TEHISÕPE I

(statistical learning, machine learning)

Loengukonspekt

Jüri Lember

kevad 2017

Kirjandus:

- "The elements of statistical learning"
  T. Hastie, R. Tibshirani, T. Friedman.
  Springer, 2001.

- "Statistical learning theory"
  V. N. Vapnik.
  Wiley, 1998.

- Ä probabilistic theory of pattern recognition"
  L. Devroye, L. Györfi, G. Lugosi.
  Springer, 1996.

- Än introduction to support vector machines and other kernel-based learning methods"
  N. Cristianini, J. Shawe-Taylor.
  Cambridge University Press, 2003.

- "Kernel methods for pattern analysis"
  J. Shawe-Taylor, N. Cristianini.
  Cambridge University Press, 2004.

- "Learning with kernels : support vector machines, regularization, optimization, and beyond"
  B. Schölkopf ; A. J. Smola.
  MIT Press, 2002.

- "Pattern recognition and machine learning"
  C. Bishop
  Springer 2006.

- "Discriminant analysis and statistical pattern recognition"
  G. McLachlan
  Wiley, 1992.

- "Introduction to machine learning"
  E. Alpaydin
  MIT, 2004.

- "Statistical pattern recognition"(2nd edition)
  A. Webb
  Wiley, 2002

- "Pattern classification"(2nd edition)
  R. Duda, P. Hart, D. Stork
  Wiley 2000

- "Pattern recognition"
  S. Theodoridis
  Academic Press, 2003

- ...

- http://www.statistical-pattern-recognition.net
  (üldine infolehekülg)

- http://www.cs.berkeley.edu/ bartlett/courses/281b-sp08/ (P. Bartletti (Berkely) SLT kursuse lehekülg. Loengukonspekt + "Readings"(muuhulgas Lugosi ülevaateartiklid))

- http://www.cs.berkeley.edu/ jordan/courses/281B-spring04/lectures.html
  (M. Jordani (Berkely) SLT kursus)

- http://www.ai.mit.edu/courses/6.867-f04/lectures.html
  (T. Jaakola (MIT) SLT slaidid)

- http://www.ece.wisc.edu/ nowak/
  (R. Nowak, (Wisconsin), SLT kursus)

- http://www.econ.upf.es/ lugosi/surveys.html
  (G. Lugosi (Pompeu Fabra) ülevaateartiklid)

- ...

Sisukord

Peatükk 1

Sissejuhatus klassifitseerimisse

1.1 Probleemi kirjeldus

Paljudes olukordades tuleb meil teha valik mitme võimaluse vahel, omamata sealjuures täielikku informatsiooni. Näiteks peab panga laenuosakond otsustama, milliseid taotlusi rahuldada või mitte. Ka eksami hindamist võib vaaldelda sama tüüpi protsessina, kus osalise informatsiooni põhjal leitakse otsitav tulemus (hinne).

Klassifitseerimine (ik *pattern recognition, classification, discrimination*) (kasutatakse ka terminit "diskrimineerimine") on teatud objekti liigitamine ühte *etteantud* klassi.

Näiteid klassifitseerimisprobleemidest:

- Pangakliendi krediidiriski hindamine. Klassid: "madal risk"ja "kõrge risk".
- Spammifilter. Klassid: "spam"ja "pole spam".
- Isendi soo määramine. Klassid: "emane"ja "isane".
- Üliõpilaste hindamine. Klassid: "A", "B", "C", "D", "E", "F".
- Kirjatähtede tuvastamine (käekirja lugemine või auto numbrimärkide tuvastamine). Klassid: "A", "B", "C",..., "Y".
- Numbrite tuvastamine. Klassid: "0", "1",..., "9".
- Meditsiooniline diagnoos. Meditsiinilise info põhjal haiguse diagnoosimine. Klassid on võimalikud haigused.
- Nägude tuvastamine: kujutise põhjal inimese identifitseerimne. Klassid: võimalikud isikud.
- ...

Kõigis neis näidetes tuleb klassifitseerijal (olgu see masin või inimene) teatava ja tihti üsna piiratud informatsiooni põhjal objekt liigitada ühte etteantud klassi. Võimalikud klassid on teada ja üldisust kitsendamata tähistame nad $\{0, \ldots, k-1\}$. Seega kokku k klassi. Klasside kodeerimisel pole arusaadavalt mingit tähtsust, kodeering $0, 1, \ldots, k-1$ on standardne, kuid kasutatakse ka $1, \ldots, k$, kahe klassi korral ka $-1, 1$ jne. Seega võib klassifitseerijat kirjeldada kui operaatorit, mille argument (sisend) on objekti kirjeldus ja väärtus (väljund) on number hulgast $\{0, \ldots, k-1\}$ – klass.

Edaspidises eeldame üldisust kitsendamata, et klassifitseeritava objekti kirjeldus esitatakse vektorina $x \in \mathbb{R}^d$. See vektor on **tunnus või tunnusvektor** (ik *feature, pattern*). Kui klassifitseerimisülesanne on isendi soo määramine, võivad tunnusteks olla näitaks isendi kaal ja pikkus. Tunnusvektor x koosneb sellisel juhul kahest komponendist. Praktikas muidugi pole tunnused alati arvulised ning kvalitatiivsed (mittearvulised) tunnused tuleb enamasti arvulisteks kodeerida. Seejuures tuleb olla ettevaatlik, sest kodeerimine mõjutab üldiselt tulemust.

Kokkuvõtteks: Klassifitseerija seab igale tunnusvektorile vastavusse klassi, s.o. arvu hulgast

$$\mathcal{Y} := \{0, \ldots, k-1\}.$$

Seega matemaatiliselt on **klassifitseerija** (ik *classifier*) funktsioon

$$g : \mathbb{R}^d \mapsto \mathcal{Y}. \tag{1.1.1}$$

Paneme tähele, et (1.1.1) avaldub kujul

$$g = \sum_{i=0}^{k-1} i I_{C_i}(x), \tag{1.1.2}$$

kus I_C on hulga C indikaatorfunktsioon, st

$$I_C(x) := \begin{cases} 1, & \text{kui } x \in C; \\ 0, & \text{kui } x \notin C. \end{cases}$$

Seega defineerib klassifitseerija g ruumi $\mathbb{R}^d$ tükelduse, tunnusega x objekt klassifitseeritakse klassi i parajasti siis, kui $x \in C_i$.

Märkus: Edaspidi on vaja (riski arvutamiseks), et klassifitseerija g oleks integreeruv ja seega mõõtuv. Funktsioon (1.1.2) on mõõtuv parajasti siis kui hulgad C_i on kõik Boreli hulgad. Seega edaspidi vaatleme ainult mõõtuvaid klassifitseerijaid.

1.2 Klassifitseerimisteooria ((Bayesian) decision theory) alused

1.2.1 Tunnuse ja klassi jaotus

Kuidas mõõta klassifitseerija g s.o. funktsiooni (1.1.1) headust? Sisestatavaid tunnuseid (klassifitseeritavaid objekte) me kindlasti ette ei tea, mistõttu neid võib käsitleda juhuslikena. Igal juhuslikul vektoril on jaotus, mida üheselt kirjeldab jaotusfunktsioon. Olgu F tunnusvektori x jaotusfunktsioon. Tuleta meelde, et suvalise (integreeruva) funktsiooni $h : \mathbb{R}^d \to \mathbb{R}$ korral, keskväärtus $h(x)$ üle tunnusvektori on Lebesgue'i integraal

$$\int h(x)dF(x).$$

Näide. Olgu ülesanne inimese soo määramine pikkuse ja kaalu põhjal. Oletame, et nii pikkus ja kui ka kaal antakse täisarvulise täpsusega, s.t. $d = 2$ ja tunnusvektor on kujul $x = (x_1, x_2)$, kus $x_1, x_2 \in \mathbb{Z}^+$. Oletame, et pikkusega x_1 ja kaaluga x_2 olevate inimeste arv populatsioonis on $N(x_1, x_2)$. Eeldades, et klassifitseeritav objekt (antud juhul küll pigem subjekt) valitakse juhuslikult ning iga inimene valitakse võrdse tõenäosusega, saame: tõenäosus, et tunnusvektori väärtus on (x_1, x_2) (loe: tõenäosus, et juhuslikult valitud inimese kaal on x_1 ja pikkus on x_2), avaldub $\frac{N(x_1,x_2)}{N}$, kus N on populatsioonis olevate inimeste arv. Tähistame selle arvu $p(x_1, x_2)$. Meie tunnusvektori võimalike väärtuste hulk on loenduv. Sellise juhusliku vektori jaotus üheselt määratud tema võimalike väärtuste ja nende tõenäosustega: $\{(x_1, x_2), p(x_1, x_2) : x_1, x_2 \in \mathbb{Z}^+\}$. Tunnusvektori jaotusfunktsioon $F : \mathbb{R}^2 \mapsto [0, 1]$ avaldub seega

$$F(x) = F(x_1, x_2) = \sum_{x_1' \le x_1, x_2' \le x_2} p(x_1', x_2').$$

Olgu $h : \mathbb{R}^d \to \mathbb{R}$ mingi funktsioon. Selle funktsiooni keskväärtus üle tunnusvektori on seega

$$\int h(x)dF(x) = \sum_{x_1,x_2} h(x_1, x_2)p(x_1, x_2).$$

Kuigi iga konkreetne objekt kuulub kindlasti vaid ühte klassi (homme kas sajab või ei, isend on kas emane või isane), pole enamikel praktikas ettetulevates klassifitseerimisülesannetes tunnusvektori x ja klasside hulga vahel ühest seost. See tähendab, et erinevatesse klassidesse kuuluvatel objektidel võib olla sama kirjeldus (sama tunnus) ehk, ümberpööratult, antud kirjeldusega (tunnusvektoriga) objekt võib kuuluda erinevatesse klassidesse. Näiteks tänase ilma põhjal ei saa kindlalt otsustada kas homme sajab või ei, sama pikkuse ja kaaluga isend võib olla nii emane kui ka isane, sarnaselt kirjutatud number võib teinekord olla "7", teinekord "1"jne. Järelikult võib tunnusele x vastava objekti klassi käsitleda juhuslikuna ning tal on oma jaotus. Et klasse on vaid lõplik arv, on see jaotus määratud paaridega

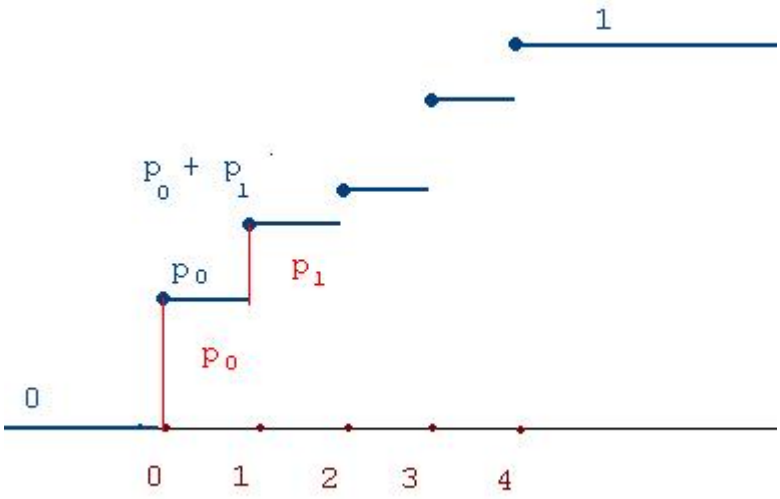
$$\{j, p(j|x) : j = 0, \ldots, k - 1\},$$

kus $p(j|x)$ on tõenäosus, et (objekti) klass on j tingimusel, et (objekti) tunnus on x. Olgu sellele jaotusele vastav jaotusfunktsioon $F(\cdot|x) : \mathbb{R} \to [0,1]$. Seega iga $y \in \mathbb{R}$ korral

$$F(y|x) := \sum_{i \le y} p(i|x).$$

Funktsiooni $h : \mathcal{Y} \to \mathbb{R}$ keskväärtus üle klasside tingliku jaoutuse $F(y|x)$ on seega

$$\int_{\mathcal{Y}} h(y) dF(y|x) = \sum_{i=0}^{k-1} h(i) p(i|x).$$



Näide. Vaatleme eelmist näidet. Vaatleme neid inimesi, kelle pikkus on 175 ja kaal 67. Olgu sellistest inimestest $\frac{2}{3}$ mehed ja $\frac{1}{3}$ naised. Siis $p(0|(175,65)) = \frac{2}{3}$ ja $p(1|(175,65)) = \frac{1}{3}$, kus klassid on kodeeritud: 1 – naine, 0 – mees.

Meie käsitluses on tunnus jaotusega F juhuslik vektor. Eelpool veendusime, et fikseeritud tunnuse x korral on ka objekti klass juhuslik, tema tinglik jaotusfunktsioon on $F(y|x)$. Seega on paar – tunnus ja tema klass – $d+1$-dimensionaalne juhuslik vektor, mille jaotus olgu määratud jaotusfunktsiooniga $F(x,y)$. Selle juhusliku vektori esimesed d-komponenti on juhuslik tunnus, viimane aga juhuslik klass, mille (tingimatu) jaotus seab igale klassile vastavusse tõenäosuse, et juhuslikult valitud objekt kuulub sinna klassi. Funktsiooni

$$h : \mathbb{R}^d \times \mathcal{Y} \mapsto \mathbb{R}$$

integreerimine mõõdu $F(x,y)$ järgi avaldub

$$\int_{\mathbb{R}^d\times\mathcal{Y}} h(x,y)dF(x,y) = \int_{\mathbb{R}^d}\int_{\mathcal{Y}} h(x,y)dF(y|x)dF(x) = \int_{\mathbb{R}^d}\sum_{i=0}^{k-1} h(x,i)p(i|x)dF(x). \tag{1.2.1}$$

Seost (1.2.1) võib vaadelda mõõdu $F(x,y)$ formaalse definitsioonina.

Näide. Jätkame ülatoodud näidet. Objekti klass on juhuslik suurus, mille võimalikud väärtused on 0 ja 1. Arv $p(0|(x_1,x_2))$ on meeste proportsioon pikkusega x_1 ja kaaluga x_2 olevate inimeste seas, $p(x_1,x_2)$ on selliste inimeste proportsioon populatsioonis. Korrutis $p(0|(x_1,x_2))p(x_1,x_2)$ on kaaluga x_1 ja pikkusega x_2 meeste proportsioon populatsioonis. Tähistame selle arvu $p(x_1,x_2,0)$. See on tõenäosus, et pikkus, kaal ja klass on $(x_1,x_2,0)$ ehk tõenäosus, et meie kolmedimensionaalne juhuslik vektor – juhuslik pikkus, juhuslik kaal ja juhuslik klass – võtab väärtuse $(x_1,x_2,0)$. Analoogiliselt olgu $p(x_1,x_2,1)$ tõenäosus, et et pikkus, kaal ja klass on $(x_1,x_2,1)$. Pikkuse, kaalu ja klassi jaotuse määravad võimalikud väärtused ja nende tõenäosused, millest saame, et jaotus $F(x,y)$ on määratud paaridega

$$\Big\{\big((x_1,x_2,i),p(x_1,x_2,i)\big) : i=0,1,\ \ x_1,x_2\in\mathbb{Z}^+\Big\}.$$

Nende arvude kaudu avaldub jaotusfunktsioon $F:\mathbb{R}^3\mapsto[0,1]$ järgmiselt:

$$F(y,x_1,x_2) = \sum_{i\le y}\sum_{x_1'\le x_1, x_2'\le x_2} p(x_1',x_2',i).$$

integraal (1.2.1) on seega (veendu selles!)

$$\int_{\mathbb{R}^d\times\mathcal{Y}} h(x,y)dF(x,y) = \sum_{x_1,x_2} h(x_1,x_2,0)p(x_1,x_2,0) + \sum_{x_1,x_2} h(x_1,x_2,1)p(x_1,x_2,1).$$

Tunnuse ja klassi ühisjaotuse kaudu avaldub kergesti juhusliku suuruse klassi jaotus. Olgu π_0 tõenäosus et (juhuslikult valitud) objekti klass on 0 (st. tõenäosus, et juhuslikult valitud inimene on mees); olgu π_1 tõenäosus, et juhuslikult valitud inimene on naine. Need tõenäosused avalduvad

$$\pi_0 = \sum_{x_1,x_2} p(x_1,x_2,0), \quad \pi_1 = \sum_{x_1,x_2} p(x_1,x_2,1).$$

Ülesanne 1.1 *Oletame, et tunnus võtab 5 võimalikku väärtust: 1,2,3,4,5. Iga väärtuse tõenäosus olgu* $\frac{1}{5}$. *Olgu klasse* $k=2$, *kusjuures* $p(1|x=i)=\frac{i}{5}$, $i=1,\dots,5$. *Leida tunnuse ja tema klassi ühisjaotus nii jaotustabeli kui ka jaotusfunktsioonina* $F(y,x)$. *Leida klasside jaotus, st tõenäosused* π_1 *ja* π_0. *Leida integraal (1.2.1), kus*

$$h(x,y) = \begin{cases} x^2, & \textit{kui } y=0; \\ -2, & \textit{kui } x\le 2, y=1; \\ -1, & \textit{kui } x=3, y=1; \\ 0, & \textit{mujal.} \end{cases}$$

Paneme tähele, et (1.1.1) on mittejuhuslik funktsioon – ühele ja samale vektorile x seatakse alati vastavusse vaid üks klass $g(x) \in \{0, \ldots, k-1\}$. Tegelikkuses see aga nii ei ole (selles just veendusime), mistõttu klassifitseerimisvead on vältimatud! Seda on lohutav teada.

Näide. Jätkame ülatoodud näidet. Olgu g mingi klassifitseerija, st funktsioon kujul (1.1.1). Oletame, et $g(175, 65) = 1$. see tähendab, et inimesed pikkusega 175 ja kaaluga 65 klassifitseeritakse alati naisteks. Et aga $\frac{2}{3}$ sellistest inimestest on hoopis mehed, teeb klassifitseerija *selliste inimeste klassifitseerimisel* kahel juhul kolmest vea. Tõsi küll, mõne teise tunnusvektori klassifitseerimisel võib see sama klassifitseerija talitada palju täpsemalt.

1.2.2 Klassifitseerija risk

Reaalses elus põhjustab klassifitseerimisviga konkreetset kahju. Näiteks kaotab pank raha, kui laenuvõtja ei suuda seda tagasi maksta jne. Teisalt toob kahju ka potensiaalse laenu tagasi lükkamine.

Definitsioon 1.1 **Kaofunktsioon** *(ik* loss function*)*

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$

seab klasside paarile (i, j) vastavusse kahju, mida toob endaga kaasa klassi i pidamine klassiks j.

Kaofunktsioon sõltub konkreetsest ülesandest, kusjuures loomulik on võtta $L(i, i) = 0$ (õigesti klassifitseerimine ei põhjusta kadu). Klassifitsserija g on seda parem, mida väiksem on kadu $L(y, g(x))$. Samas on paar (x, y) juhuslik ning samamoodi on juhuslik ka kadu $L(y, g(x))$, mistõttu pole selge, mida tähendab väike kadu. Seetõttu on loomulik vaadelda klassifitseerija keskmist kadu. Siit definitsioon

Definitsioon 1.2 *Klassifitseerija g* **risk** *(ik.* risk*) on keskmine kahju üle tunnusvektori ja klasside ühisjaotuse $F(x, y)$:*

$$R(g) := \int L(y, g(x)) dF(x, y). \tag{1.2.2}$$

Seosest (1.2.1) saame

$$R(g) = \int_{\mathbb{R}^d} \int_{\mathcal{Y}} L(y, g(x)) dF(y|x) dF(x) = \int_{\mathbb{R}^d} \sum_{j=0}^{k-1} L(j, g(x)) p(j|x) dF(x). \tag{1.2.3}$$

Seosest (1.1.2) saame

$$R(g) = \sum_{i=0}^{k-1} \int_{C_i} \sum_{j=0}^{k-1} L(j, i) p(j|x) dF(x) = \sum_{j,i=0}^{k-1} L(j, i) \int_{C_i} p(j|x) dF(x).$$

1.2.3 Bayesi klassifitseerija – parim võimalik klassifitseerija

On selge, et heale klassifitseerijale vastab väike risk, kuid milline on parim võimalik klassifitseerija? Teisisõnu: milline klassifitseerija minimiseerib riski $R(g)$ üle kõikide võimalike klassifitseerijate hulga?

Olgu x mingi tunnusvektor ning g klassifitseerija. Leiame **tingliku riski** ehk keskmise klassifitseerimiskahju tunnuse x korral. Selleks peame keskmistama võimalikud kulud $L(j, g(x))$ üle kõikvõimalike klasside j tõenäosuste (sageduste) $p(j|x)$

$$R(g(x)|x) := \sum_{j=0}^{k-1} L(j, g(x))p(j|x) = \int_{\mathcal{Y}} L(y, g(x))dF(y|x).$$

Seega tinglik risk $R(i|x)$ on keskmine risk tunnusega x objekti klassifitseerimisel klassi i. Keskmistades tingliku riski üle tunnusvektorite x jaotuse saame kogu riski (seos (1.2.3))

$$\int_{\mathbb{R}^d} R(g(x)|x)dF(x) = \int_{\mathbb{R}^d} \int_{\mathcal{Y}} L(y, g(x))dF(y|x)dF(x) = R(g).$$

Ülaltoodud võrduste ahelast lähtub, et minimiseerides igas punkis x tingliku riski oleme minimiseerinud ka kogu riski. Tõepoolest, olgu klassifitseerija g^* defineeritud järgmiselt

$$g^*(x) := \arg\min_{g(x)} R(g(x)|x) = \arg\min_{i\in\mathcal{Y}} R(i|x) = \arg\min_{i\in\mathcal{Y}} \sum_{j=0}^{k-1} L(j, i)p(j|x). \tag{1.2.4}$$

Seega

$$R(g^*(x)|x) = \min_{i\in\mathcal{Y}} R(i|x),$$

millest järeldub, et iga teise klassifitseerija g korral $R(g^*(x)|x) \leq R(g(x)|x)$. Integraali monotoonsuse tõttu

$$R(g^*) = \int R(g^*(x)|x)dF(x) \leq \int R(g(x)|x)dF(x) = R(g). \tag{1.2.5}$$

Seega klassifitseerija g^* risk on väikseim üle kõikvõimalike klassifitseerijate:

$$R(g^*) = \inf_g R(g).$$

Pane tähele, et $g^*(x)$ alati leidub, kuid ei pruugi olla ühene. Kuid ülaltoodud võrratused kehtivad iga seosega (1.2.4) defineeritud klassifitseerija g^* korral, mistõttu kõik nad on minimaalse riskiga.

Definitsioon 1.3 *Seosega (1.2.4) defineeritud klassifitseerijat g^* nimetatakse* **Bayesi klassifitseerijaks** *ning tema riski*

$$R^* := R(g^*) = \int L(y, g^*(x))dF(y, x)$$

nimetatakse **Bayesi riskiks** .

Arvestades g^* definitsiooni (1.2.4), saame Bayesi riskile kuju

$$R^* = \int \min_{i\in\mathcal{Y}} R(i|x)dF(x) = \int \min_{i\in\mathcal{Y}} \sum_{j=0}^{k-1} L(j,i)p(j|x)dF(x).$$

Kokkuvõttes: Bayesi klassifitseerija on parim võimalik klassifitseerija, Bayesi risk on väikseim võimalik keskmine klassifitseerimiskahju.

Erijuht $k = 2$**:** Seostest (1.2.4) on Bayesi klassifitseerijat lihtne leida. Kui $L(i,i) = 0$ ja $k = 2$, siis iga x korral tinglik risk avaldub (veendu selles!)

$$R(i|x) = \begin{cases} L(1,0)p(1|x) & \text{kui } i = 0, \\ L(0,1)p(0|x) & \text{kui } i = 1. \end{cases}$$

(Üks võimalik) Bayesi klassifitseerija on siis

$$g^*(x) = \begin{cases} 0 & \text{kui } L(1,0)p(1|x) \le L(0,1)p(0|x), \\ 1 & \text{kui } L(1,0)p(1|x) > L(0,1)p(0|x). \end{cases} \tag{1.2.6}$$

Näide. Jätkame ülaltoodud näidet. Olgu g mingi klassifitseerija kujul (1.1.1). Selle risk avaldub

$$\begin{aligned} R(g) &= \int L(y, g(x))dF(x,y) = \sum_{x_1,x_2,i} L(i, g(x_1,x_2))p(x_1,x_2,i) \\ &= \sum_{x_1,x_2} \Big(\sum_{i=0}^{1} L(i, g(x_1,x_2))p(i|x_1,x_2)\Big)p(x_1,x_2), \end{aligned} \tag{1.2.7}$$

kus tinglik

$$R(g(x)|x) = \sum_{i=0}^{1} L(i, g(x_1,x_2))p(i|x_1,x_2)$$

on keskmine kahju, mida g tekitab pikkusega x_1 ja kaaluga x_2 olevate inimeste klassifitseerimisel. Riski arvutamisel aga korrutatakse see arv läbi selliste inimeste proportsiooniga populatsioonis $p(x_1,x_2)$. Seega, kui selliseid inimesi on väga vähe, s.t. $p(x_1,x_2) \approx 0$, ei suurenda selliste inimeste klassifitseermisel tehtav kahju oluliselt klassifitseerija riski; kui aga $p(x_1,x_2)$ on suur, võib tunnusvektori (x_1,x_2) klassifitseerimisel olla riskile suur mõju.

Ülesanne 1.2 *Olgu objekti klass tunnusvektori x funktsioon, st $\exists f : \mathbb{R}^d \to \mathcal{Y}$ nii, et $p(f(x)|x) = 1$. Leida Bayesi klassifitseerija ja risk.*

1.2.4 Juhuslik ehk stohhastiline klassifitseerija *

Nägime, et Bayesi klassifitseerija on parim võimalik klassifitseerija kõikide kujul (1.1.1) olevate klassifitseerijate hulgast. Et aga tunnusega x objekti kuulumine mingisse kassi on juhuslik, võib tekkida kiusatus kasutada selliseid klassifitseerijaid, kus funktsiooni $g(x)$ väärtus pole üheselt määratud. Näiteks võime klassifitseerijat ette kujutada kui programmi, mis vastuse saamiseks viskab vahetevahel kulli ja kirja. Sisuliselt genereeritakse juhuslikul klassifitseerimisel üks juhuslik suurus nn. lisajuhuslikkus, mis tunnusvektori x kõrval on samuti klassifitseerija sisend. Lisajuhuslikkuse jaotus võib sõltuda tunnusvektorist x: nii näiteks võime tunnusvektori teatud väärtuste korral visata kulli ja kirja, teiste väärtuste korral täringut ning mõnikord kasutada hoopis determineeritud klassifitseerijat. Et aga suvalise jaotusega juhusliku suuruse saab genereerida ühtlasest jaotusest teatava (jaotusest sõltuva) funktsiooni abil (nn Skorohodi esitus), saame juhusliku klassifitseerija alati esitada funktsioonina

$$g : \mathbb{R}^d \times \Omega \to \mathcal{Y},$$

kus $\omega \in \Omega$ on lisajuhuslikkus, mille jaotus enam tunnusektorist ei sõltu (näiteks ühtlse jaotusega juhuslik suurus).

Juhusliku klassifitseerija korral on loomulik eeldus, et ühegi vektori x väärtuse korral ei sõltu juhuslik komponent ω ei objekti tegelikust klassist. On ju formaalselt juhuslik ka selline klassifitseerija, mis igale objektile seab vastavusse tema tegeliku klassi. Selline klassifitseerija aga eeldab olulist lisainformatsiooni, meie käsutuses on aga vaid tunnusvektor. Seega on iga tunnusvektori x korral on juhuslik komponent ja objekti klass sõltumatud, mistõttu nende tinglikud jaotusfunksioonid rahuldavad seost

$$F(y, \omega|x) = F(y|x)F(\omega|x).$$

Et lisajuhuslikkuse jaotus ei sõltu tunnusvektorist, siis $F(\omega|x) = F(\omega)$ ja

$$F(y, \omega|x) = F(y|x)F(\omega).$$

Ka juhusliku klassifitseerija korral on tinglik risk defineeritud kui tunnusega x objektide klassifitseerimisel tekkiv keskmine kahju, kuid keskmine on nüüd arusaadavalt nii üle klasside kui ka juhusliku komponendi. Seega avaldub tinglik risk

$$R(g(x)|x) = \int L(y, g(x, \omega))dF(y, \omega|x) = \int_\Omega \int_\mathcal{Y} L(y, g(x, \omega))dF(y|x)dF(\omega) =: \int_\Omega h(x, \omega)dF(\omega).$$

Iga funktsiooni $\omega \mapsto h(x, \omega)$ korral leidub väärtustus ω^* (sõltub x-st) mis on väiksem või võrdne keskmisest

$$h(x, \omega^*) \leq \int h(x, \omega)dF(\omega),$$

millest järeldub, et leidub selline determineeritud klassifitseerija, $g_o(x) = g(x, \omega^*)$ (sõltub x-st), mille tinglik risk tunnusvektori x korral pole suurem $g(x, \omega)$ tinglikust riskist:

$$R(g_o(x)|x) = \int_\mathcal{Y} L(y, g(x, \omega^*))dF(y|x) = h(x, \omega^*) \leq \int h(x, \omega)dF(\omega) = R(g(x)|x).$$

Et aga iga x ja iga determineeritud klassifitseerija korral kehtib (1.2.5), siis on Bayesi klassifitseerija tinglik risk tunnusvektori korral risk väiksem või võrdne juhusliku klassifitseerija tinglikust riskist. Kokkuvõtvalt Bayesi klassifitseerija on parim nii juhuslike kui determineeritud klassifitseerijate seas! Edaspidi me juhuslikke klassifitseerijaid ei käsitle.

1.2.5 Sümmeetriline kaofunktsioon

Kuidas g^* tegelikult välja näeb, sõltub kaofunktsioonist L. Vast kõige enam levinud kaofunktsioon on **sümmeetriline ehk 0-1 kaofunktsioon**. Sellise kaofunktsiooni korral on täpse klassifitseerimise kahju alati 0 ning kõikide vigade kahju on võrdne (näiteks 1). Seega on sümmeetriline kaofunktsioon järgmine

$$L(j,i) = \begin{cases} 0 & \text{kui } i = j, \\ 1 & \text{mujal.} \end{cases} \tag{1.2.8}$$

Sellise kaofunktsiooni korral avaldub tinglik risk $R(i|x)$ järgmiselt:

$$R(i|x) = \sum_{j=0}^{k-1} L(j,i)p(j|x) = \sum_{j\neq i} p(j|x) = 1 - p(i|x).$$

Seega tinglik risk $R(i|x)$ on tõenäosus, et tunnusega x objekt ei kuulu klassi i. Seega Bayesi klassifitseerija on (definitsioon (1.2.4))

$$g^*(x) = \arg\min_i R(i|x) = \arg\min_i (1 - p(i|x)) = \arg\max_i p(i|x).$$

Teisisõnu: sümmeetrilise kaofunktsiooni korral seab Bayesi klassifitseerija igale tunnusele vastavusse selle klassi, mille (tinglik) tõenäosus on suurim. Mis võiks veel loogilisem olla?

Olgu g mingi klassifitseerija. Sümmeetrilise kaofunktsiooni korral on selle tinglik risk

$$R(g(x)|x) = 1 - p(g(x)|x)$$

valesti klassifitseerimise (tinglik) tõenäosus tingimusel, et objekti tunnus on x (Tõepoolest, kui $g(x) = i$, siis $R(g(x)|x) = R(i|x) = 1 - p(i|x)$. See aga on tõenäosus, et tunnusega x objekt ei kuulu klassi i). Keskmistades valesti klassifitseerimise tinglikku tõenäosust üle tunnusvektori jaotuse $F(x)$ saame üldise **klassifitseerimisvea tõenäosuse**. Et aga tingliku riski keskmistamine üle tunnusvektori jaotuse $F(x)$ annab riski, siis sümmeetrilise kaofunktsiooni korral on klassifitseerija risk klassifitseerimisvea tõenäosus ja Bayesi klassifitseerija annab väikseima võimaliku klassifitseerimisvea.

Sümmeetrilise kaofunktsiooni korral avaldub Bayesi risk

$$R^* = \int \min_i R(i|x)dF(x) = \int \min_i (1 - p(i|x))dF(x) = 1 - \int \max_i p(i|x)dF(x). \tag{1.2.9}$$

Erijuht $k = 2$: Vaatleme kuidas avalduvad Bayesi klassifitseerija ja Bayesi risk (eri)juhul, kui klasse on kaks. Et $p(1|x) + p(0|x) = 1$, siis $p(1|x) \geq p(0|x)$ parajasti siis, kui $p(1|x) \geq 0.5$. Seega sümmeetrilise kaofunktsiooni korral on Bayesi klassifitseerija (1.2.6):

$$g^*(x) = \begin{cases} 1 & \text{kui } p(1|x) \geq 0.5, \\ 0 & \text{kui } p(1|x) < 0.5. \end{cases} \tag{1.2.10}$$

Bayesi klassifitseerija tinglik risk avaldub

$$R(g^*(x)|x) = \min\{p(0|x), p(1|x)\},$$

millest Bayesi risk on seega

$$R^* = \int \min\{p(0|x), p(1|x)\} dF(x). \tag{1.2.11}$$

Paneme tähele: kui x on selline, et $p(1|x) = 0.5$, siis selle klassifitseerimine ei mõjuta riski.

Ülesanne 1.3 *Olgu klasse kaks ja kaofunktsioon sümmeetriline. Olgu $g(x) = I_C$ suvaline klassifitseerija. Avaldada selle risk $R(g)$ funktsioonide $p(0|x)$ ja $p(1|x)$ kaudu st üldistada valemit (1.2.11).*

Ülesanne 1.4 *Olgu klasse kaks ja kaofunktsioon L järgmine:*

$$L(i,j) = \begin{cases} 0, & \textit{kui } i = j; \\ t_1, & \textit{kui } i = 1, j = 0; \\ t_2, & \textit{kui } i = 0, j = 1. \end{cases}$$

Leida Bayesi klassifitseerija ja avaldada see $p(1|x)$ kaudu (st üldistada valemit (1.2.10)). Leida Bayesi risk ja üldistada valemit (1.2.11). Kuidas avalduvad Bayesi klassifitseerija ja risk juhul, kui $t_1 = t_2$?

Ülesanne 1.5 *Olgu $d = 1$, $k = 2$. Iga funktsiooni $h : \mathbb{R} \to \mathbb{R}$ korral, defineerime kaofunktsiooni*

$$J_p(h) = \int_{\mathbb{R}\times\mathcal{Y}} |h(x) - y|^p dF(x,y).$$

Leida

$$h_p^* = \arg\inf_h J_p(h), \quad p = 1, 2.$$

Veendu, et $J_1(h_1^*) = R^*$.

Ülesanne 1.6 *Tõestada, et valem (1.2.11) avaldub*

$$R^* = \frac{1}{2} - \frac{1}{2} \int |2p(1|x) - 1| dF(x).$$

1.2.6 Risk ja klassifitseerija tinglike tiheduste kaudu

Olgu $F(x|i)$, $i \in \mathcal{Y}$ tunnusvektori tinglik jaotusfunktsioon klassis i. Seega (täistõenäosuse valem)

$$F(x) = \sum_{i=0}^{k-1} F(x|i)\pi_i,$$

kus

$$\pi_i := \int_{\mathbb{R}^d} p(i|x)dF(x), \quad i = 0, \ldots, k-1$$

on klassi i tõenäosus. Teisisõnu, jaotus $\pi_0, \ldots, \pi_{k-1}$ on $F(x, y)$ marginaaljaotus.

Olgu f_i jaotuste $F(x|i)$, $i = 0, \ldots, k-1$ on tihedus mingi ühe ja sama mõõdu suhtes. Kui see mõõt on Lebesgue'i mõõt, on jaotused absoluutselt pidevad ning tihedus f_i on sisuliselt tõenäosusteooria algkursusest tuntud pideva jaotuse tihedus. Juhul, kui $F(x|i)$ on diskreetsed jaotused, on f_i tõenäosusteooria algkursusest tuntud tõenäosusfunktsioon, sest tõenäosusfunktsioon on tihedus loendava mõõdu suhtes. Tiheduse olemasolu ei ole kitsendav eeldus jaotustele $F(x|i)$, sest igal mõõdul leidub tihedus mingi teise mõõdu suhtes. Käesoleval juhul eeldame, et jaotustel $F(x|i)$ on tihedus ühe ja sama mõõdu suhtes. Ka see ei ole kitsendav, sest alati leidub mõõt, mille suhtes kõik jaotused $F(x|i)$ on absoluutselt pidevad (seega tihedust omavad).

Edaspidi tähistame mõõtu, mille suhtes f_i on tihedus dx. Seega iga (integreeruva) funktsiooni h korral

$$\int_{\mathbb{R}^d} h(x)dF(x|i) = \int_{\mathbb{R}^d} h(x)f_i(x)dx.$$

Jaotusel $F(x)$ on seega tihedus f, mis avaldub $f(x) = \sum_i f_i(x)\pi_i$.

Tinglike tiheduste kaudu avaldub $p(i|x)$ järgmiselt (Bayesi valem):

$$p(i|x) = \frac{f_i(x)\pi_i}{f(x)}. \tag{1.2.12}$$

Seega Bayesi klassifitseerija (1.2.4) avaldub

$$g^*(x) = \arg\min_{i\in\mathcal{Y}} \sum_{j=0}^{k-1} L(j, i)f_j(x)\pi_j. \tag{1.2.13}$$

Sümmeetriline kaofunktsioon

Juhul, kui kaofunktsioon on sümmetriline, on Bayesi klassifitseerija kujul

$$g^*(x) = \arg\max_{i\in\mathcal{Y}} p(i|x) = \arg\max_{i\in\mathcal{Y}} \pi_i f_i(x). \tag{1.2.14}$$

Et $R(i|x) = 1 - p(i|x)$, siis funktsioon $R(i|x)f(x)$ avaldub sellisel juhul

$$R(i|x)f(x) = f(x) - \pi_i f_i(x) = \sum_{j,j\neq i} \pi_j f_j(x).$$

Seega Bayesi risk on (tuleta meelde (1.2.9))

$$R^* = \int \min_i R(i|x)dF(x) = \int \min_i R(i|x)f(x)dx = \int \min_i \big(\sum_{j,j\neq i} \pi_j f_j(x)\big)dx$$
$$= \int \big(f(x) - \max_i \pi_i f_i(x)\big)dx = 1 - \int \max_i \pi_i f_i(x)dx.$$

Erijuht $k = 2$**:** Kahe klassi korral on eeskiri (1.2.14) järgmine

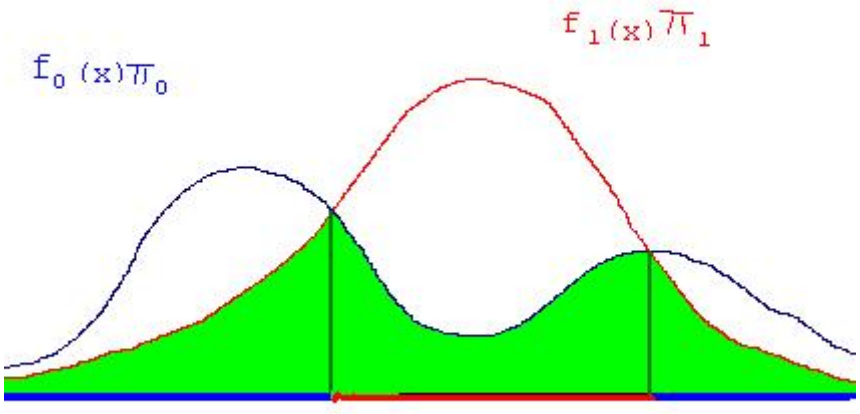
$$g^*(x) = \begin{cases} 1, & \text{kui } \pi_1 f_1(x) \geq \pi_0 f_0(x); \\ 0, & \text{mujal.} \end{cases}$$

ja selle võib esitada **tõepärasuhte** (ik *likelihood ratio*) kaudu järgmiselt

$$g^*(x) = \begin{cases} 1, & \text{kui } \frac{f_1(x)}{f_0(x)} \geq \frac{\pi_0}{\pi_1}; \\ 0, & \text{mujal.} \end{cases} \tag{1.2.15}$$

Bayesi risk kahe klassi (ja sümmeetrilise kaofunktsiooni) korral on

$$R^* = \int \min\{\pi_1 f_1(x), \pi_0 f_0(x)\}\, dx. \tag{1.2.16}$$



Näide. Jätkame näidet. Arvud π_0 ja π_1 on vastavalt meeste ja naiste proportsioon populatsioonis. Olgu $p(x|0)$ ja $p(x|1)$ tunnuse jaotus meesta ja naiste seas:

$$p(x|0) = p(x_1, x_2|0) = \frac{\text{pikkusega } x_1 \text{ ja kaaluga } x_2 \text{ meeste arv}}{\text{meeste arv}}.$$

Vastavalt Bayesi valemile

$$p(i|x) = \frac{p(x,i)}{p(x)} = \frac{p(x|i)\pi_i}{p(x)}, \quad i = 0, 1.$$

Järelikult $p(0|x) < p(1|x)$ parajasti siis, kui $p(x|0)\pi_0 < p(x|1)\pi_1$ ja sümmeetrilise kaofunktsiooni korral on Bayesi klassifitseerija seega

$$g^*(x) = \begin{cases} 0 & \text{kui } \pi_0 p(x|0) < \pi_1 p(x|1), \\ 1 & \text{vastasel korral.} \end{cases}$$

See on seos (1.2.14). Bayesi risk (1.2.11) on seega

$$R^* = \int \min\{p(0|x), p(1|x)\} dF(x) = \sum_x \min\{p(0|x), p(1|x)\} p(x)$$

$$= \sum_x \min\{\frac{p(x|0)\pi_0}{p(x)}, \frac{p(x|1)\pi_1}{p(x)}\} p(x) = \sum_x \min\{p(x|0)\pi_0, p(x|1)\pi_1\}.$$

Toodud võrduste parem pool on (1.2.16).

Ülesanne 1.7 *Olgu klasse kaks ja kaofunktsioon sümmeetriline. Olgu $g = I_C$ suvaline klassifitseerija. Avalda risk $R(g)$ tiheduste f_0 ja f_1 kaudu, st üldista seost (1.2.16).*

Ülesanne 1.8 *Olgu klasse kaks ja kaofunktsioon L nagu ülesandes 1.4. Näita, et Bayesi klassifitseerija avaldub tõepärasuhte statistiku kaudu, st üldista seost (1.2.15) ebasümmeetrilisele kaofunktsioonile.*

Ülesanne 1.9 *Olgu tunnus x tundide arv, mida üliõpilane kulutab aine "Tehisõpe"õppimiseks. Olgu tõenäosus, et x tundi õppiv tudeng sooritab arvestuse $p(1|x)$. Tõenäosus $p(0|x) = 1 - p(1|x)$ olgu tõenäosus, et tudeng kukub läbi. Oletame, et $p(1|x)$ on järgmine (kasvav) funktsioon*

$$p(1|x) = \frac{x}{c + x}, \quad c > 0$$

Leida Bayesi klassifikaator klassifitseerimaks tudengit läbikukkujaks või mitte (sümmeetrilise kaofunktsiooni korral) õppimiseks kulutatud aja põhjal. Oletame, et õppimise aeg tudengite seas on ühtlase jaotusega 0-st $4c$, st

$$f(x) = \frac{1}{4c} I_{[0,4c]}.$$

Avaldada π_i, f_i (tinglikud tihedused) ja Bayesi risk ($R^* = \frac{1}{4}\ln\frac{5e}{4}$, $\pi_1 = 1 - \frac{1}{4}\ln 5$).

Ülesanne 1.10 *Tõestada, et kui $\pi_0 = \pi_1$, siis valem (1.2.16) avaldub*

$$R^* = \frac{1}{2} - \frac{1}{4}\int |f_1(x) - f_0(x)| dx.$$

1.3 Võimalus otsustamata jätta

Teinekord võib valesti otsustamise riski vähendada sel moel, et tunnust ei klassifitseerita. Näiteks lükatakse klassifitseerimine täiendava informatsiooni saabumiseni edasi. Muidugi on loomulik, et seda ei tehta kuitahes tihti ikka vaid nn "kriitiliste" tunnuste korral. Ingliskeelses kirjanduses nimetatakse klassifitseerimist võimalusega otsustamata jätta *reject option*. Tagasilükkamise korral on klassifitseerijal lisaväärtus – olgu see "r" – mis tähendab, et tunnus jäetakse klassifitseerimata. Selline klassifitseerija g, on sellisel juhul funktsioon

$$g : \mathbb{R}^d \to \mathcal{Y} \cup \{r\}, \tag{1.3.1}$$

Klassifitseerija g jagab ruumi $\mathbb{R}^d$ $k+1$ lõikumatuks piirkonnaks $C_0, \ldots, C_{k-1}$ ja R (k klassi ja r) nii, et

$$g = \sum_{i=0}^{k-1} iI_{C_i} + rI_R.$$

Kui $x \in R$, siis tunnust ei klassifitseerita, piirkond R on **tagasilükkamispiirkond** (*reject region*), piirkond $A := \cup_{i=0}^{k-1} C_i$ on **otsustamispiirkond** (*acceptance region*). On loomulik nõuda, et piirkonna R tõenäosus (st tõenäosus, et juhuslik tunnusvektor kuulub sinna piirkonda) on suhteliselt väike.

Tuletame meelde, et Bayesi klassifitseerija $g^*(x)$ minimiseerib tingliku riski, st

$$R(g^*(x)|x) = \min_i R(i|x).$$

Seega Bayesi klassifitseerija on tunnuse x klassifitseerimisel suhteliselt riskivaba, kui $\min_i R(i|x)$ on väike. Riskantse tunnuse x korral aga on $\min_i R(i|x)$ suhteliselt suur. Sellest loogikast lähtudes vaatleme piirkonda R kujul

$$R(t) = \{x : \min_i R(i|x) > t\} \subset \mathbb{R}^d \tag{1.3.2}$$

kus $t \geq 0$ on mingi fikseeritud lävi. Otsustamispiirkond on sellisel juhul

$$A(t) = \{x : \min_i R(i|x) \leq t\}$$

ja sel piirkonnal on loomulik kasutada Bayesi klassifitseerijat. Nii saame klassifitseerija

$$g_t(x) := \begin{cases} \arg\min_i R(i|x), & \text{kui } \min_i R(i|x) \leq t \ ; \\ r, & \text{kui } \min_i R(i|x) > t. \end{cases}$$

Olgu g suvaline tagasilükkamise võimalusega klassifikaator otsustamispiirkonnaga A ja tagasilükkamispiirkonnaga R. Et $r \notin \mathcal{Y}$, siis pole defineeritud kadu $L(i, r)$ ning seega juhul kui $x \in R$, pole defineeritud ka tinglik risk $R(g(x)|x)$. Seega pole defineeritud ka tagasilükkamisvõimalusega klassifikaatori g risk, küll on aga defineeritud integraal

$$\int_A R(g(x)|x) dF(x). \tag{1.3.3}$$

(tuleta meelde, et riski korral on integraal üle kogu $\mathbb{R}^d$, antud juhul vaid üle alamhulga A). Kuidas sellisel juhul võrrelda erinevaid tagasilükkamise võimalusega klassifikaatoreid? Integraalide (1.3.3) vahetu võrdlemine pole õigustatud, sest selle saab teha kuitahes väikeseks võttes A piisavalt väikese (õigemini piisavalt väikese tõenäosusega). Küll võib aga võrrelda integraale (1.3.3) juhul, kui hulga A tõenäosused on võrdsed. Hulga A tõenäaosus siin on tõenäosus, et tunnusvektor võtab väärtuse sellest hulgast ja see tõenäosus on

$$P(A) := \int_A dF.$$

Järgnev lihtne lemma näitab, et kõigi nende tagasilükkamisvõimalustega klassifikaatorite hulgas, mille otsustuspiirkonna tõenäosus on vähemalt sama suur kui $P(A(t))$ (mingi t korral), on justnimelt g_t see, mille korral integraal (1.3.3) väikseim. Oletame, et otsustuspiirkonna tõenäosus peab olema vähemalt 0.95. Lemmast järeldub, et kui leidub t nii, et $P(A(t)) = 0.95$ (see eeldus ei pruugi alati täidetud olla, miks?), siis kõikide lubatud tagasilükkamisvõimalustega klassifikaatorite hulgast on just g_t korral (1.3.3) väikseim.

Lemma 1.1 *Olgu g mingi tagasilükkamisvõimalusega klassifitseerija, selle otsustuspiirkond ja tagasilükkamispiirkonnad olgu vastavalt A ja R. Kui $P(R) \leq P(R(t))$, siis*

$$\int_A R(g(x)|x)dF(x) \geq \int_{A(t)} R(g_t(x))dF(x) = \int_{A(t)} \min_i R(i|x)dF(x). \tag{1.3.4}$$

Tõestus. Et

$$\int_A R(g(x)|x)dF \geq \int_A \min_i R(i|x)dF,$$

siis piisab kui näitame, et

$$\int_A \min_i R(i|x)dF(x) \geq \int_{A(t)} \min_i R(i|x)dF(x). \tag{1.3.5}$$

Võrratuse (1.3.5) tõestus on ülesanne 1.11. ■

Ülesanne 1.11 *tõesta järgnev võrratus*

$$\int_A \min_i R(i|x)dF(x) - \int_{A(t)} \min_i R(i|x)dF(x) \geq t(P(A) - P(A(t)) \geq 0.$$

Märkus: Tähistades

$$R(g|A) := \frac{\int_A R(g(x)|x)dF}{P(A)}$$

klassifitseerija g riski tingimusel, et tunnus klassifitseeritakse (kuulub piirkonda A), võib lemma 1.1 esitada kujul: kui $P(R) \leq P(R(t))$, siis $R(g|A) \geq R(g_t|A(t))$.

Sümmeetriline kaofunktsioon ja tagaslükkamisvõimalus. Sümmeetrilise kaofunktsiooni korral $R(i|x) = 1 - p(i|x)$ ja seega klassifitseerija g_t on kujul

$$g_t(x) := \begin{cases} \arg\max_i p(i|x), & \text{kui } \max_i p(i|x) \geq 1 - t \ ; \\ r, & \text{kui } \max_i p(i|x) < 1 - t. \end{cases} \tag{1.3.6}$$

Piirkond $R(t)$ kahaneb kui t kasvab; kui $t \geq \frac{k-1}{k}$, siis $R(t) = \emptyset$.

Kui klasse on kaks, siis klassifitseerijal on tagasilükkamisvõimalus, kui $t < 0.5$ Olgu $0.5 - t = c$ ehk $t = 0.5 - c$. Seega g_t võib esitada kujul

$$g_t(x) := \begin{cases} 1, & \text{kui } p(1|x) \geq \frac{1}{2} + c \ ; \\ 0, & \text{kui } p(1|x) \leq \frac{1}{2} - c \ ; \\ r, & \text{kui } |\frac{1}{2} - p(1|x)| < c. \end{cases} \tag{1.3.7}$$

Kui $c = 0$, on (1.3.7) (sisuliselt) Bayesi klassifitseerija (1.2.10). Klassifitseerija (1.3.7) on intuitiivselt väga mõistetav: tunnust x on suhteliselt kindel klassifitseerida, kui $p(1|x)$ või $p(0|x)$ on piisavalt suur (suurem $0.5 + c$), siis klassifitseerimine toimub Bayesi reegli alusel. Kui aga $p(1|x)$ on aga ligikaudu $\frac{1}{2}$, siis on tunnuse x klassifitseerimine suhteliselt riskantne ja nii lükatakse selle klassifitseerimine edasi. Läve c (või t) muutmisega saab muuta piirkonna $R(t)$ suurust ja mõõtu.

1.4 Klassifitseerimine kui hüpoteeside kontroll, Neyman-Pearsoni lemma*

Vaatleme olukorda, kus klasse on kaks. Reegel (1.1.2) sellisel juhul on kujul $g = I_C$, C on piirkond, kus tunnus klassifitseeritakse klassi 1. Klassifitseerimisülesannet kahe klassi korral võib vaadelda statistikast tuntud lihtüpoteeside kontrollina. Hüpoteesid:

$$H_0 : \text{objekti klass on } 0$$
$$H_1 : \text{objekti klass on } 1$$

ja valim koosneb tunnusest x. Funktsioon I_C on statistiline test ja klassifitseerimisel tehtud viga võib olla kaht tüüpi: tegelikult klassi 0 kuuluv element klassifitseeritakse klassi 1 ja vastupidi. Tihti nimetatakse neid esimest ja teist tüüpi vigadeks. Olgu nende vigade tõenäosused vastavalt α ja β, st

$$\alpha(g) := \int_C f_0(x)dx, \quad \beta(g) := \int_{C^c} f_1(x)dx.$$

Siin f_0 ja f_1 on tinglikud tihedused; α on tõenäosus, et klassi 0 kuuluv element klassifitseeritakse (klassifikaatori I_C abil) klassi 1 ja β on tõenäosus, et klassi 1 kuuluv element klassifitseeritakse klassi 0. Vigade kaalutud keskmine

$$\pi_0\alpha + \pi_1\beta = R(g)$$

on klassifitseerimisvea tõenäosus ja sümmeetrilise kaofunktsiooni korral klassifitseerija risk (ülesanne 1.7). Esimest tüüpi vea vähendamine üldiselt suurendab teist tüüpi viga ja vastupidi.

Tihti on oluline, et esimest tüüpi viga ei ületaks teatud olulisusenivood α_0 (näiteks $\alpha_0 = 0.05$). Seega on eesmärk küll üldise klassifitseerimisvea – riski $R(g)$ – vähendamine, kuid lisatingimus on, et esimest tüüpi viga püsiks teatud piirides. Seega optimeerimisülesanne on järgmine:

$$\begin{aligned} &\min_g \big(\pi_0\alpha(g) + \pi_1\beta(g)\big). \\ &\alpha(g) \le \alpha_0. \end{aligned} \tag{1.4.1}$$

Teinekord on aga tähtis vähendada just teist tüüpi viga, hoides esimest tüüpi viga kontrolli all. Optimeerimisülesanne on siis

$$\begin{aligned} &\min_g \beta(g) \\ &\alpha(g) \le \alpha_0. \end{aligned} \tag{1.4.2}$$

Statistikast tuntud Neyman-Pearsoni lemma annab üldkuju ülesannete (1.4.1) ja (1.4.2) lahendeile.

Lemma 1.2 (Neyman-Pearsoni lemma) *Iga klassifitseerija g ja iga $t > 0$ korral kehtib võrratus*

$$\big(\beta(g) - \beta(g_t)\big) + t\big(\alpha(g) - \alpha(g_t)\big) \ge 0, \tag{1.4.3}$$

kus g_t on järgmine klassifitseerija (tõepärasuhte statistik):

$$g_t = I_{C(t)}, \quad C(t) := \Big\{x : \frac{f_1(x)}{f_0(x)} > t\Big\}.$$

Tõestus. Olgu $g = I_B$. Olgu $C := C(t)$, st $g_t = I_C$. Veendu, et iga x korral kehtib võrratus

$$\big(I_{C^c}(x) - I_{B^c}(x)\big)\big(tf_0(x) - f_1(x)\big) \ge 0. \tag{1.4.4}$$

[Võta $x \in C^c$ ja näita, et (1.4.4) kehtib. Siis võta $x \in C$ ja näita sama.] Seos (1.4.4) on

$$tf_0(x)I_{C^c}(x) - tf_0(x)I_{B^c}(x) - f_1(x)I_{C^c}(x) + f_1(x)I_{B^c}(x) \ge 0.$$

Integreerides saame

$$\begin{aligned} 0 &\le \int_{C^c} f_0(x)dx - t\int_{B^c} f_0(x)dx - \int_{C^c} f_1(x)dx + \int_{B^c} f_1(x)dx \\ &= t(1-\alpha(g_t)) - t(1-\alpha(g)) - \beta(g_t) + \beta(g) = t(\alpha(g) - \alpha(g_t)) + \beta(g) - \beta(g_t). \end{aligned}$$

■

Võrratusest (1.4.3) järeldub, et klassifitseerija g_t on optimaalne järgmises mõttes: kui leidub klassifitseerija g nii, et $\beta(g) < \beta(g_t)$, siis $\alpha(g) > \alpha(g_t)$. Seega, kui leidub t nii, et $\alpha(g_t) = \alpha_0$, siis g_t on ülesande (1.4.2) lahend. Veendume, et sellisel juhul on g_t ka ülesande (1.4.1) lahend. Olgu

$$t^* := \frac{\pi_0}{\pi_1}, \quad \alpha^* := \alpha(g_{t^*}).$$

Teame, et klassifitseerija g_{t^*} on Bayesi klassifitseerija. Järelikult kui $\alpha_0 \geq \alpha^*$, siis lisatingimus pole kitsendav ja ülesande (1.4.1) lahend on Bayesi klassifitseerija $g^* = g_{t^*}$. Seega edaspidi vaatleme olukorda, kus $\alpha_0 < \alpha^*$. Kui leidub selline t et $\alpha(g_t) = \alpha_0$, peab kehtima võrratus $t > t^*$. [Tõepoolest, kui $t \leq t^*$, siis $C(t^*) \subset C(t)$ ning $\alpha(g_t) \geq \alpha^*$.] Oletame vastuväiteliselt, et g_t pole (1.4.1) lahend. See tähendab, et leidub g nii, et $\alpha(g) \leq \alpha_0$ ja

$$R(g) = \pi_0\alpha(g) + \pi_1\beta(g) < \pi_0\alpha(g_t) + \pi_1\beta(g_t) = R(g_t).$$

Viimane võrratus tähendab aga, et

$$\pi_1\big(\beta(g_t) - \beta(g)\big) + \pi_0\big(\alpha(g_t) - \alpha(g)\big) > 0$$

ehk arvestades, et $\alpha(g_t) = \alpha_0$

$$\big(\beta(g) - \beta(g_t)\big) < \frac{\pi_0}{\pi_1}\big(\alpha_0 - \alpha(g)\big) = t^*\big(\alpha_0 - \alpha(g)\big).$$

Neyman-Pearsoni lemmast saame aga, et

$$\big(\beta(g) - \beta(g_t)\big) \geq t\big(\alpha_0 - \alpha(g)\big) \geq 0.$$

Mõlemad võrratused saavad kehtida vaid siis kui $t < t^*$. See on aga vastuolu eeldusega.

Seega, kui leidub t nii, et $\alpha(g_t) = \alpha_0$, siis ülesannete (1.4.1) ja (1.4.2) lahendi (optimaalseima testi) annab tõepärasuhete satistik. Ülesandest 1.8 järeldub aga, et tõepärasuhete statistiku põhjal saadud klassifitseerija g_t on Bayesi klassifitseerija sobiva ebasümmeetrilise kaofunktsiooni korral. Seega erinevaid vigu saab sama hästi kontrollida kaofunktsiooni valikuga.

ROC-graafik: suuruse $1 - \beta(g_t)$ [tõenäosus, et klass 1 (signaal) klassifitseeritakse õigesti] sõltuvus suurusest $\alpha(g_t)$ [tõenäosus, et klass 0 (müra) loetakse klassiks 1 (valehäire)] Neyman-Pearsoni klassifikaatori g_t korral. Mida väiksem on $\alpha(g_t)$ (valehäire tõenäosus), seda väiksem on ka $1-\beta(g_t)$. Seega graafik kasvav. Mida kiiremini, seda parem (kasvamise kiirus sõltub mudelist).

Peatükk 2

Sissejuhatus Vapnik-Cervonenkise teooriasse

2.1 Treeningandmetest õppimine

Ülaltoodust nägime, et kui tinglikud tõenäosused $p(i|x)$ või tinglikud jaotused $F(y|x)$ on teada, on parimat võimalikku klassifitseerijat võimalik teoreetiliselt leida (seos (1.2.4)). Samuti teame, et tõenäosusi $p(i|x)$ on lihtne leida, kui teame tinglikke jaotusi F_i või tinglikke tihedusi f_i ja tõenäosusi π_i iga $i = 0, \dots k-1$ korral (valem (1.2.13)). Tinglike jaotuste F_i ja tõenäosuste π_i teadmine on samaväärne tunnusvektori ja tema klassi ühisjaotuse $F(x, y)$ teadmisega, sest (täistõenäosuse valem)

$$F(x, y) = \sum_{i=0}^{k-1} F_i(x)\pi_i.$$

Eeldus, et $F(x, y)$ (ja seetõttu ka $p(i|x)$, samuti $F(x|i)$ ja kõik muu) on teada pole kuigi realistlik. Praktikas on jaotus $F(x, y)$ on enamasti tundmatu, selle asemel on meil antud **treeningandmed** ehk **valim** (ik. *sample*):

$$\mathcal{D}_n := \{(x_1, y_1), \dots, (x_n, y_n)\}. \tag{2.1.1}$$

Valimis on n objekti, mille tunnusvektorid $x_1, \dots, x_n$ ning klassifikatsioon $y_1, \dots, y_n$ on teada. Enamasti eeldatakse, et kõik valimi paarid (x_i, y_i) on saadud sõltumatult jaotusest $F(x, y)$. Teisisõnu, valim (2.1.1) on n sõltumatu ning jaotusega $F(x, y)$ juhusliku vektori realisatsioon. Sellist valimit nimetatakse **iid** (*independent and identically distributed*) juhuslikuks valimiks. Sellised eeldused pole alati põhjendatud ning tihti on polegi see nii. Näiteks võivad paarid (x_i, y_i) olla ajalises sõltuvuses (aegrida) vms. Käesolevas kursuses vaatleme vaid iid valimit.

Kokkuvõttes: jaotus $F(x, y)$ pole meil teada, selle asemel on meil seda jaotust lähendav iid valim (2.1.1). Seetõttu ei saa me ka kasutada Bayesi klassifitseerijat, sest selleks vajame

jaotust $F(x, y)$. Tehisõppe valdkonna – klassifitseerimisteooria (*statistical pattern recognition*) – eesmärk on treeningandmete põhjal võimalikult hea klassifitseerija leidmine. Seda protsessi nimetatakse tehisõppes tihti klassifitseerija **treenimiseks** (ik *training, learning*). Statistikas nimetatakse sama asja klassifitseerija **hindamiseks** (ik *estimation*). Asjaolu, et valimis on teada ka vektorite õiged klassid nimetatakse **õpetajaga treenimiseks** (*supervised learning*). Sel moel saadud andmetest sõltuv klassifitseerija on sisuliselt funktsioon

$$g_n : (\mathbb{R}^d \times \mathcal{Y})^n \times \mathbb{R}^d \to \mathcal{Y} \tag{2.1.2}$$

ning seega sõltub selle risk

$$R(g_n) = \int L(y, g(\mathcal{D}_n, x))dF(x, y) \tag{2.1.3}$$

samuti valimist. Et $F(x, y)$ jaotus pole teada, siis on samuti teadmata risk $R(g_n)$, selle hindamine on teoorias kesksel kohal.

Klassifitseerimisreegel. Vaadeldes klassifitseerijat kui algoritmi, siis on mõistlik nõuda, et klassifitseerija aksepteeriks kuitahes suurt hulka treeningandmeid. Selle idee matemaatiliseks formalisatsiooniks on **klassifitseemimisreegel** (ik *classification rule*).

Definitsioon 2.1 *Klassifitseerijate (2.1.2) jada $g_1, g_2, \ldots, g_k, \ldots$ nimetatakse klassifitseerimisreegliks.*

Sisuliselt on klassifitseerimisreegel klassifitseerimispritntsiip (algoritm) mis (üldiselt) ei sõltu treeningandmete hulgast ja kujust. Küll aga sõltub treeningandemtest reegli rakendamisel saadud klassifitseerija g_n. Näiteks kahe klassi korral võib klassifitseerimisreegel olla järgmine: klassifitseeri tundmatu objekt klassi 1, kui treeningandmetes $y_1, \ldots, y_n$ on ühtesid vähemalt sama palju kui nulle. Sellise reegli abil saadud klassifitseerija sõltub küll andmetest, kuid reegel on rakendatav suvalise n ja suvaliste treeningandmete korral. Loomulikult pole selline klassifitseerimisreegel eriti mõttekas, sest sellisel teel saadud klassifitseerija ei arvesta tunnusvektorit x.

2.2 Formulatsioon juhuslike vektorite abil

Formuleerime klassifitseerimisprobleemi tõenäosusteooria keeles. Olgu $(\Omega, \mathcal{F}, \mathbf{P})$ tõenäosusruum ning (X, Y) juhuslik vektor, kus

$$X : \Omega \to \mathbb{R}^d, \quad Y : \Omega \to \mathcal{Y}.$$

Seega X on $\mathbb{R}^d$-dimensionaalne juhuslik vektor ja Y võtab väärtusi hulgas $\mathcal{Y}$. Vektor (X, Y) modelleerib tunnust ja tema klassi. Olgu vektori (X, Y) jaotusfunktsioon $F(x, y)$. Seega iga $x \in \mathbb{R}^d$ ja $y \in \mathbb{R}$ korral (vektorite $x, x' \in \mathbb{R}^d$ korral võrratus $x \leq x'$ on defineeritud komponentide kaupa)

$$F(y, x) = \mathbf{P}(X \leq x, Y \leq y).$$

Klassi i tinglik tõenäosus $p(i|x)$ on seega $\mathbf{P}(Y = i|X = x)$ ning tingimatu tõenäosus $\pi_i = \mathbf{P}(Y = i)$. Tunnuse X jaotusfunktsioon ja tinglik jaotusfunktsioon on vastavalt

$$F(x) = \mathbf{P}(X \leq x) \text{ ja } F(x|i) = \mathbf{P}(X \leq x|Y = i).$$

Saab näidata, et $p(i|x)$ on mõõtuv iga $i = 0, \ldots, k-1$ korral.

Risk. Iga (mõõtuva) funktsiooni $h : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ korral

$$Eh(X,Y) = \int h(X,Y)d\mathbf{P} = \int_{\mathbb{R}^d \times \mathcal{Y}} h(x,y)dF(x,y).$$

Seega suvalise klassifitseerija g korral (tuleta meelde, et klassifitseerija g on alati mõõtuv ja nii on ka funktsioon $L(y, g(x))$) on risk

$$R(g) = EL(Y, g(X)) = \int L(Y, g(X))d\mathbf{P}.$$

Juhul, kui L on sümmeetriline kaofunktsioon (1.2.8), siis

$$L\big(Y(\omega), g(X(\omega))\big) = \begin{cases} 1, & \text{kui } Y(\omega) \neq g(X(\omega)); \\ 0, & \text{kui } Y(\omega) = g(X(\omega)). \end{cases}$$

Seega sümmeetrilise kaofunktsiooni korral klassifitseerija g risk on klassifitseerimisvea tõenäosus:

$$R(g) = EL(Y, g(X)) = \int I_{\{Y \neq g(X)\}}d\mathbf{P} = \mathbf{P}(Y \neq g(X)).$$

Kui klasse on kaks (ja kaofunktsioon sümmeetriline), siis esimest ja teist tüüpi vead on vastavalt

$$\alpha(g) := \mathbf{P}(g(X) = 1|Y = 0), \quad \beta(g) := \mathbf{P}(g(X) = 0|Y = 1),$$

millest (täistõenäosuse valem)

$$R(g) = \mathbf{P}(g(X) \neq Y) = \mathbf{P}(g(X) = 1, Y = 0) + \mathbf{P}(g(X) = 0, Y = 1) =$$
$$\mathbf{P}(g(X) = 1|Y = 0)\mathbf{P}(Y = 0) + \mathbf{P}(g(X) = 0|Y = 1)\mathbf{P}(Y = 1) = \alpha(g)\pi_0 + \beta(g)\pi_1.$$

Sümmeetrilise kaofunktsiooni korral Bayesi risk on (vt (1.2.9))

$$R^* = \min_g \mathbf{P}(g(X) \neq Y) = 1 - E(\max_i p(i|X)),$$

kui klasse on kaks siis viimane valem on (vt (1.2.11))

$$R^* = E(\min\{p(1|X), 1 - p(1|X)\}).$$

Treeningandmed. Olgu

$$D_n := ((X_1, Y_1), \ldots, (X_n, Y_n)) \tag{2.2.1}$$

sõltumatud sama jaotusega juhuslikud vektorid, paari (X_i, Y_i) jaotus on sama, mis (X, Y) jaotus; D_n modelleerib valimit. Klassifitseerija g_n sõltub vektoritest D_n. Seega

$$g_n(\cdot; D_n) : \mathbb{R}^d \mapsto \mathcal{Y} \tag{2.2.2}$$

on valimist sõltuv juhuslik funktsioon. Tema risk $R(g_n)$ on samuti valimi D_n funktsioon, sest

$$R(g_n) = E[L(Y, g_n(X; D_n))|D_n],$$

kus (X, Y) on valimist D_n sõltumatu juhuslik vektor. Tinglik keskväärtus ülaltoodud avaldises sisuliselt tähendab, et keskväärtus võetud üle X ja Y. Kui valim on fikseeritud, st juhuslike vektorite $(X_1, Y_1), \ldots, (X_n, Y_n)$ väärtused on $(x_1, y_1), \ldots, (x_n, y_n)$, on juhusliku suuruse $R(g_n)$ väärtus

$$EL(Y, g_n(X; x_1, y_1, \ldots, x_n, y_n)).$$

Et $R(g_n)$ on juhuslik suurus, siis edaspidi vaatleme tihti tema keskväärtust $ER(g_n)$, kus keskväärtus on arusaadavalt võetud üle juhusliku valimi D_n. Seega

$$ER(g_n) = E\big(E[L(Y, g_n(X; D_n))|D_n]\big) = E[L(Y, g_n(X; D_n)],$$

kus viimane võrdus järeldub tingliku keskväärtuse omadusest. Pane tähele: et (X, Y) ja D_n on omavahel sõltumatud ja D_n on iid valim, siis keskväärtus $ER(g_n)$ sõltub ainult klassifitseerijast g_n (ja seeläbi arvust n, sest g_n on ikka kujul (2.1.2)) ning vektori (X, Y) jaotusest $F(x, y)$. Seetõttu jäetakse D_n tihti tähistustest välja ning kirjutatakse

$$ER(g_n) =: EL(Y, g_n(X)).$$

Sümmeetrilise kaofunktsiooni korral seega risk $R(g_n)$ on tinglik tõenäosus

$$R(g_n) = \mathbf{P}[Y \neq g_n(X, D_n)|D_n]$$

ning

$$ER(g_n) = E\big(\mathbf{P}[Y \neq g_n(X, D_n)|D_n]\big) = \mathbf{P}[Y \neq g_n(X, D_n)] =: \mathbf{P}(Y \neq g_n(X)).$$

Ülesanne 2.1 *Olgu $T : \mathbb{R}^d \to \mathbb{R}^d$ suvaline mõõtuv funktsioon. Olgu R^* juhusliku vektori (X, Y) Bayesi risk ning R_T^* olgu juhusliku vektori $(T(X), Y)$ Bayesi risk. Tõestada, et $R_T^* \geq R^*$, st tunnuse X teisendus (andmete töötlemine) ei saa vähendada riski.*

Ülesanne 2.2 *Olgu X' sõltumatu vektorist (X, Y). Olgu R^* vektori (X, Y) Bayesi risk ja R' olgu $((X, X'), Y)$ Bayesi risk. Tõestada, et $R^* = R'$.*

Ülesanne 2.3 *Olgu L sümmeetriline kaofunktsioon ning g, g' olgu kaks klassifikaatorit. Tõestada, et*

$$|R(g) - R(g')| \leq \mathbf{P}\big(g'(X) \neq g(X)\big).$$

Tõestada, et kui klasse on kaks, siis

$$|R(g) - R(g')| \leq E[|2p(1|X) - 1|I_{\{g'(X) \neq g(X)\}}].$$

2.3 Mõjusus

Et valimi põhjal konstrueeritud klassifitseerija g_n risk $R(g_n)$ on juhuslik suurus (sõltub valimist), tekib jällegi küsimus klassifitseerija (nüüd juba kui funktsiooni (2.2.2)) headuse mõõtmisest. Võimalikult väikese riski nõue siin ei tööta: risk $R(g_n)$ võib olla mõne valimi korral väike, kuid mõne teise valimi korral suur. Valim on aga juhuslik. Küll on aga mõistlik kasutada klassifikaatoreid, mille keskmine risk üle kõigi valimite on väike. See annab lootust, kuid ei garanteeri, et ka konkreetse valimi $\mathcal{D}_n$ korral on risk väike. Seega vaatleme keskmist riski $ER(g_n)$. Paneme tähele, et iga valimi korral $R(g_n) \geq R^*$. Siit järeldub, et võrdus $ER(g_n) = R^*$ kehtib vaid siis, kui $R(g_n) = R^*$ p.k. Sellisel juhul peaks aga klassifitseerija g_n olema (peaaegu kindlasti) võrdne Bayesi klassifitseerijaga ning viimase konstrueerimine vaid valimi põhjal on ebareaalne – (üldiselt) pole võimalik konstrueerida klassifitseerijat (2.2.2) mille keskmine risk oleks Bayesi risk. Ometi on aga selge, et kui meie käsutuses on väga palju andmeid (n on väga suur), siis on meil nii palju informatsiooni jaotuse $F(x, y)$ kohta, et peaks olema võimalik defineerida klassifitseerijat g_n, mille keskmine risk oleks peaaegu R^*.

Definitsioon 2.2 *Ütleme, et klassifitseerimisreegel* $\{g_n\}$ *on* **mõjus** *(ik* consistent, asymptotically consistent*), kui*

$$ER(g_n) \to R^*. \tag{2.3.1}$$

Me ütleme, et reegel on **tugevalt mõjus** *(ik* strongly consistent*), kui*

$$R(g_n) \to R^* \quad p.k.$$

Kui reegel on (tugevalt) mõjus iga jaotuse $F(x, y)$ korral, nimetame seda **universaalselt (tugevalt) mõjusaks** .

Paneme tähele, et mõjusus on reegli, mitte klassifitseerija omadus. Et $R(g_n)$ on tõkestatud juhuslik suurus, on koondumine (2.3.1) ekvivalentne koondumisega tõenäosuse järgi, s.t. $\forall \epsilon > 0$,

$$\lim_n \mathbf{P}(R(g_n) - R^* > \epsilon) = 0.$$

Mõjusus on hea omadus. Kui reegel on mõjus, siis tõenäosus, et keskmine risk $R(g_n)$ erineb Bayesi riskist vähem kui ϵ, läheneb ühele. Seega suure n korral on g_n "peaaegu niisama hea kui g^*". Oluline on, et g_n ei pruugi olla (mingis mõttes) lähedal Bayesi reeglile g^*. Seega, kui $\{g_n\}$ on mõjus ja n suur, siis ei pruugi kehtida $g_n \approx g^*$, küll aga on lähedased g_n ja g^* klassifitseerimisomadused. Vapniku terminoloogias mõjus reegel imiteerib kuid ei identifitseeri parimat võimalikku klassifitseerijat. Kuid mida me ühelt masinalt ootame: kas seda et ta töötaks hästi või et ta oleks mingis mõttes sarnane parima võimaliku masinaga? (Lugeja võiks siinkohal meelde tuletada vanake Hottabõtši võlutud mustast marmorist telefoni).
Tugev mõjusus tähendab, et risk $R(g_n)$ läheneb Bayesi riskile (peaaegu) iga võimaliku treeningvalimi korral. Sellest järeldub mõjusus.

Üldiselt sõltub reegli $\{g_n\}$ mõjusus jaotusest $F(x,y)$. Iga jaotuse $F(x,y)$ korral leidub vähemalt üks mõjus reegel: $g_n \equiv g^*$ iga n korral. Mõne teise jaotuse jaoks ei pruugi see reegel aga mõjus olla. Seetõttu on hea, kui reegel on mõjus võimalikult laia klassi jaotuste korral. Universaalselt mõjus reegel on aga mõjus iga võimaliku jaotuse korral. Universaalselt mõjusad reeglid õnneks eksisteerivad (see on üllatav ning üldsegi mitte triviaalne); nende olemasolu on väga hea, sest tihti (tegelikult peaaegu mitte kunagi, kuigi statistikud sellest kõva häälega ei räägi) pole meil mingit aimu tegeliku jaotuse kohta. Ning mitteuniveraalse reegli korral pole meil siis ka mingit aimu selle reegli mõjususe kohta.

2.3.1 Aeglasest koondumisest

Olgu $\{g_n\}$ universaalselt mõjus reegel ja $\epsilon > 0$. Universaalne mõjusus tähendab, et iga jaotuse $F(x,y)$ korral leidub n_o (sõltub jaotusest ja ka ϵ-ist) nii, et

$$ER(g_n) - R^* < \epsilon \text{ kui } n > n_o \tag{2.3.2}$$

Kas aga leidub n_o (sõltub muidugi ϵ-ist) nii, et (2.3.2) kehtib iga jaotuse $F(x,y)$ korral? Kui see nii oleks, siis piisavalt suure valimi mahu $n > n_o$ korral oleks keskmine risk Bayesi riskile lähemal kui ϵ *mistahes jaotuse korral*. Järgmine teoreem ([1], Thm 7.1) näitab, et sellist universaalset n_o ei saa olemas olla ehk keskmise riski koondumine üle jaotuste pole ühtlane.

Teoreem 2.3 *Olgu klasse kaks ja kaofunktsioon sümmeetriline. Iga $\epsilon > 0$, n ja reegli $\{g_n\}$ korral leidub jaotus $F(x,y)$ nii, et $R^* = 0$ ja $ER(g_n) > \frac{1}{2} - \epsilon$.*

Esmapilgul võib tunduda, et teoreem 2.3 on vastuolus universalselt mõjusa reegli olemasoluga. Tõepoolest, kui $\{g_n\}$ on universaalselt mõjus reegel, siis koondumine $ER(g_n) \to R^*$ kehtib iga jaotuse korral, kaasaarvatud kõik sellised jaotused, kus $R^* = 0$. Vastuolu siiski pole, sest teoreemis figureeriv halb jaotus sõltub n-st. See tähendab, et näiteks $n = 1000000$ korral leidub jaotus $F'(x,y)$ nii, et $R^* = 0$ ja

$$ER(g_{1000000}) > \frac{1}{2} - \epsilon.$$

Et aga $\{g_n\}$ on universaalselt mõjus reegel ja jaotuse F' korral $R^* = 0$, siis ka jaotuse F' korral kehtib $ER(g_n) \to 0$. Ent see koondumine on nii aeglane, et $ER(g_{1000000}) > 0.5 - \epsilon$.

Seega: iga reegel võib teatud jaotuste korral käituda väga halvasti, isegi siis, kui reegel on universaalselt mõjus ja valimi maht n kuitahes suur!

Statistikas pakub tihti huvi koonduva jada kiirus. Ülaltoodud teoreem ei ütle tegelikult midagi jada $ER(g_n)$ koondumise kiiruse kohta. Nii ei välista teoreem 2.3, et iga n korral $ER(g_n) - R^* \leq \frac{C}{n}$, kus C on konstant mis sõltub jaotusest $F(x,y)$. Sellisel juhul oleks $ER(g_n)$ koondumiskiirus vähemalt $\frac{1}{n}$ ja kui leiduks konstant $c > 0$ nii, et mingi jaotuse korral $ER(g_n) - R^* \geq \frac{c}{n}$, siis ütleme, et parim võimalik koondumiskiirus (minimax mõttes) on $\frac{1}{n}$. Ja kui mõne teise universaalselt mõjusa reegli $\{g'_n\}$ korral jada $ER(g_n) - R^*$

parim võimalik kiirus on näiteks $\frac{1}{\sqrt{n}}$, siis tuleks eelistada kiiremat kuigi kiirus on siin puhtalt asümptootilises tähenduses ja sellest ei saa teha järeldusi ühegi konkreetse n korral, sest teoreem 2.3 ju kehtib. Järgmine negatiivne tulemus aga näitab et ühelgi reeglil pole universaalset (üle jaotuste) koondumiskiirust, sest iga reegli ja kuitahes aeglaselt nulliks koonduva jada $\{a_n\}$ korral leidub jaotus $F(x, y)$ (sõltub sellest jadast ja ka reeglist) nii, et

$$ER(g_n) - R^* \geq a_n, \quad \forall n. \tag{2.3.3}$$

Teoreem 2.4 *Olgu $\{a_n\}$ nulliks koonduvate reaalarvude jada ning $\frac{1}{16} \geq a_1 \geq a_2 \geq \cdots$. Olgu $\{g_n\}$ klassifitseerimisreegel. Siis leidub jaotus $F(x, y)$ nii, et $R^* = 0$ ja kehtib (2.3.3).*

Teoreemi tõestus on raamatus [1] (Thm 7.2). Teoreem 2.4 näitab, et iga (ka universaalselt mõjusa) reegli keskmise riski koondumine võib olla halva jaotuse korral kuitahes aeglane. See aga ei tähenda, et ei võiks leiduda universaalselt mõjusat "superreeglit"$\{g_n\}$ nii, et iga teise reegli $\{g'_n\}$ korral $ER(g_n) \leq ER(g'_n)$ $\forall n$ korral. Tõsi, halva jaotuse korral koondub ka "superreegel" väga aeglaselt, kuid teised on kõik veel halvemad. Selgub aga, et sellist ülireeglit ei saa olla, sest iga klassifitseerimisreegli $\{g_n\}$ korral leidub jaotus $F(x, y)$ ning mingi teine klassifitseerimisreegel $\{g'_n\}$ nii, et

$$ER(g'_n) < ER(g_n), \quad \forall n. \tag{2.3.4}$$

Seega pole sellist reeglit, mis iga jaotuse korral minimiseerib riski üle kõikvõimalike klassifitseerijate (1.1.1). Kui selline superreegel leiduks, poleks klassifitseerimisteoorial sügavat mõtet – kogu teooria tegeleks vaid selle reegliga.

2.4 Empiirilise riski minimiseerimise printsiip

Reeglina otsitakse klassifitseerijat teatud funktsioonide klassist $\mathcal{G}$. Loomulikult võib klassina $\mathcal{G}$ vaadelda ka kõiki võimalikke klassifitseerijaid, kuid enamasti on $\mathcal{G}$ väiksem ja mingis mõttes lihtsam alamklass. Alljärgnevas tutvume lihtsa kuid fundamentaalse meetodiga klassist $\mathcal{G}$ andmete põhjal parima funktsiooni valimiseks. Seda meetodi nimetatakse empiirilise riski minimiseerimise printsiibiks (ERM-printsiip) ning teatavas mõttes on ERM-printsiip statistikas (eriti tehisõppes) kesksel kohal.

2.4.1 Empiirilise riski minimiseerimine

Kuidas valida valimi põhjal klassifitseerijat hulgast $\mathcal{G}$? Tuletame meelde, et parim klassifitseerija klassist $\mathcal{G}$ on see, mis minimiseerib riski $R(g)$ üle $\mathcal{G}$. Seda valikuprintsiipi ei saa me rakendada, sest pole teada jaotus $F(y, x)$. Selle asemel on meil aga sellest jaotusest genereeritud iid valim $\mathcal{D}_n$. Valimit $\mathcal{D}_n$ võib vaadelda empiirilise jaotusfunktsioonina F_n. Tuleta meelde, et iga $(x, y) \in \mathbb{R}^{d+1}$ korral

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq x, y_i \leq y\}},$$

ja iga funktsiooni $h : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ korral

$$\int h(x,y)dF_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} h(x_i,y_i). \tag{2.4.1}$$

Seega on F_n jaotuse F hinnang, kusjuures hea hinnang – kehtib **Glivenko-Cantelli teoreem:**

$$\sup_{x,y} |F_n(x,y) - F(x,y)| \to 0, \quad p.k..$$

Siit idee: riski arvutamisel kasutame tundmatu jaotuse F asemel valimi põhjal leitud empiirilist jaotust F_n. Teisisõnu, riski arvutamise valemis (1.2.2) asendame tundmatu jaotuse F selle valimi põhjal saadud hinnanguga F_n. Nii saame **empiirilise riski** (ik *empirical risk*)

$$R_n(g) := \int L(y,g(x))dF_n(y,x) = \frac{1}{n}\sum_{i=1}^{n} L(y_i,g(x_i)). \tag{2.4.2}$$

Klassifikaatoriks g_n võtame loomulikult selle, mis minimiseerib empiirilise riski üle $\mathcal{G}$ ehk

$$g_n = \arg\inf_{g\in\mathcal{G}} R_n(g).$$

Seega oleme g_n leidnud **empiirilise riski minimiseerimisel.**

Kui L on sümmeetriline, siis empiiriline risk on

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{y_i \neq g(x_i)\}}.$$

Seega on $R_n(g)$ proportsionaalne klassifitseerimisvigade arvuga, mida g teeb kui ta on rakendatud treeningvalimile. Teisisõnu: sümmeetrilise kaofunktsiooni korral saadakse empiirilise riski minimiseerimisel selline klassifitseerija, mis minimiseerib klassifitseerimisvigade arvu treeningvalimis. Üsna loogiline, kas pole?

2.4.2 Näiteid empiirilise riski minimiseerimisest statistikaprobleemidel

Vaatleme ERM printsiipi mõnevõrra laiemas kontekstis kui vaid klassifitseerimisteooria. Formuleerime järgmise üldise tehisõppe probleemi. Olgu meil iid valim jaotusest $F(x,y)$

$$\mathcal{D}_n := (x_1,y_1),\ldots,(x_n,y_n), \tag{2.4.3}$$

kus $(x_i,y_i) \in \mathcal{X}\times\mathcal{Y}$ ja $\mathcal{X}$, $\mathcal{Y}$ on suvalised hulgad. Tunnuseid $x_1,\ldots,x_n$ interpreteerime kui sisendeid, millele on vastavusse seatud väljundid $y_1,\ldots,y_n$. Väljundi tinglik jaotus sisendi x korral on $F(y|x)$. Tehisõppe eesmärk on antud (väljundita) sisendile x sellise väljundi

y vastavusse seadmine, mis (mingis mõttes) kõige paremini ennustab või jäljendab tegelikkust.

Olgu $\mathcal{G}$ funktsioonide klass hulgast $\mathcal{X}$ hulka $\mathcal{Y}$, iga funktsioon $g \in \mathcal{G}$ korral defineerime riski

$$R(g) := \int L(y, g(x))dF(y, x), \tag{2.4.4}$$

kus kaofunktsioon L sõltub ülesande püstitusest. Risk $R(g)$ mõõdab keskmist kadu.

Näited statistikaülesannetest

Klassifitseerimisülesanne. Siin $\mathcal{Y} = \{0, \ldots, k-1\}$ (klassid) ja g on $\mathcal{Y}$-väärtuseline funktsioon. Enamasti L on sümmeetriline kaofunktsioon ja sellisel juhul risk on klassifitseerimisvea tõenäosus ning parim võimalik klassifitseerija üle kõikide funktsioonide on Bayesi klassifitseerija, mis igale tunnnusvektorile x seab vastavussse klassi, mille tinglik tõenäosus on suurim:

$$g^*(x) = \arg\max_i p(i|x).$$

Regressiooniülesanne. Siin $\mathcal{Y} = \mathbb{R}$ ja kaofunktsioon L on enamasti $L(y, g) = (y-g)^2$. Risk on sellisel juhul

$$R(g) = \int (y - g(x))^2 dF(x, y)$$

ja parim võimalik funktsioon üle kõikide funktsioonide on tinglik keskväärtus:

$$g^*(x) = \int y dF(y|x).$$

Tiheduse hindamise ülesanne. Siin valim koosneb vaid sisenditest $x_1, \ldots, x_n$ st väljundid puuduvad (formaalselt $\mathcal{Y} = \emptyset$) ja klass $\mathcal{G}$ olgu tihedusfunktsioonide klass. Kaofunktsioon olgu järgmine: $L(g) = -\ln g$. Seega risk on **tõepärakontrast** (*likelihood contrast*):

$$R(g) = -\int \ln g(x) dF(x).$$

Juhul, kui jaotusel F on tihedus $p(x)$, siis

$$R(g) = -\int \ln g(x) p(x) dx$$

ning selle funktsiooni minimiseerib (üle kõikide tiheduste) just $p(x)$, st $g^*(x) = p(x)$.

ERM-printsiip: funktsiooni (2.4.4) asemel minimiseeri empiirilist riskifunktsiooni

$$R_n(g) := \int L(y, g(x))dF_n(y, x) = \frac{1}{n}\sum_i L(y_i, g(x_i)). \tag{2.4.5}$$

ERM-printsiip statistikaülesannetel

Klassifitseerimisülesanne. Sümmmeetrilise kao korral ERM-printsiip on treeningvea minimiseerimine:

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i)) = \frac{1}{n}\sum_{i=1}^{n} I_{\{(x,y):g(x)\neq y\}}(x_i, y_i).$$

Regressiooniülesanne. ERM-printsiip on vähimruutude printsiip:

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i)) = \frac{1}{n}\sum_{i=1}^{n} (y_i - g(x_i))^2.$$

Tiheduse hindamine. ERM-printsiip on suurima tõepära printsiip:

$$R_n(g) = -\frac{1}{n}\sum_{i=1}^{n} \ln g(x_i).$$

Seega on ERM printsiip juba ligi sada aastat edukalt kasutusel erinevates statistikavaldkondades. Erinevates statistikaharudes on sellel küll erinevad nimetused ja ka rakendused, põhimõtteline olemus on aga sama. Seetõttu on loomulik eeldada ühtse teooriat, mis käsitleks ERM-printsiibi olemust ja hõlmaks korraga nii klassifitseerimisteooriat, regressioonanalüüsi, suurima tõepära hinnanguid ja palju muud. Selline teooria on nn Vapnik-Chervonenkise teooria, mis on teatavas mõttes tehisõppe teoreetiline baas.

2.4.3 Lähendamisviga ja hindamisviga

Olgu $\mathcal{G}$ funktsioonide klass ja g_n valimi põhjal (mitte ilmtingimata ERM meetodil) valitud klassifitseerija hulgast $\mathcal{G}$. Klassifitseerija g_n risk $R(g_n)$ on muidugi suurem Bayesi riskist, vahe võib aga jagada kahte ossa

$$R(g_n) - R^* = \Big(R(g_n) - \inf_{g\in\mathcal{G}} R(g)\Big) + \Big(\inf_{g\in\mathcal{G}} R(g) - R^*\Big). \tag{2.4.6}$$

Neist esimene – **hindamisviga** (ik *estimation error*) – sõltub valimist ja klassifikaatori g_n valimise meetoodist. See viga on juhuslik suurus. Intuitiivselt on selge, et kui klass $\mathcal{G}$ on suur, on hindamisviga ka üldiselt suur. Näitena liiga suurest klassist $\mathcal{G}$ vaatleme sümmeetrilise kaofunktsiooniga klassifitseerimisülesannet. Olgu $\mathcal{G}$ kõikide funktsioonide klass ja g_n ERM-printsiibil valitud klassifitseerija. Seega g_n on selline hulga $\mathcal{G}$ element, mis minimiseerib treeningviga. Kui tunnusvektori jaotus on absoluutselt pidev, siis tõenäosusega üks on kõik valimi punktid erinevad. Siis leidub aga klassifitseerijaid, mis ei tee treeningvalimi klassifitseerimisel ühtegi viga. Üks selline klassifitseerija on näiteks selline

$$g_n(x) = \begin{cases} y_i & \text{kui } x = x_i, \\ 0 & \text{otherwise.} \end{cases}$$

Selline klassifitseerija on aga peaaegu kindlasti 0, mistõttu

$$R(g_n) = \mathbf{P}(Y \neq g(X)) = \mathbf{P}(Y \neq 0).$$

Seega $R(g_n)$ on valimist sõltumatu konstant, mis on üldiselt palju suurem Bayesi riskist. Kahtlemata pole klassifitseerija, mis (peaaegu) kõik punktid klassifitserib klassi 0, just see resulataat, mida me masinalt ootame. Põhjus sellise ebaloomuliku klassifitseerija saamiseks empiirilise riski minimiseerimisel on see, et $\mathcal{G}$ on liiga suur. See fenomen on **ülesobitumus** (ik *overfitting*).

Teine viga – **lähendamisviga** (ik *approximation error*) – on mittejuhuslik ja sõltub klassist $\mathcal{G}$. Mida suurem on klass, seda väiksem on lähendamisviga. Kui $\mathcal{G}$ on kõikide mõõtuvate funktsioonide hulk, on lähendamisviga 0. Ent nagu nägime, on aga sellisel juhul enamasti hindamisviga lootusetult suur. Näitena liiga väikesest klassist vaatame klassi, mis koosneb vaid ühest funktsioonist $\mathcal{G} = \{g\}$. Sellisel juhul $g_n = g$ (sest valida midagi suurt pole) ning $R(g_n) = R(g) = \inf_{\mathcal{G}} R(g)$ ehk hindamisviga on 0. Samas lähendamisviga on $R(g) - R^*$, mis, sõltuvalt jaotusest, võib olla väga suur.

Ülaltoodust järeldub, et hea klassifitseerija pärineb klassist $\mathcal{G}$, mille suurus on teatavas mõttes optimaalne. Selgu, et klassi $\mathcal{G}$ füüsilisest suurusest (võimsusest) olulisem on tema nn **kompleksus**. Küsimus $\mathcal{G}$ optimaalsest kompleksusest on teoorias kesksel kohal.

2.4.4 Lähendamisvea ja riski hinnangud

Olgu $\mathcal{G}$ klassifikaatorite hulk. Olgu $g_n \in \mathcal{G}$ valimi põhjal valitud (hinnatud, treenitud) klassifitseerija. Et (X, Y) jaotus $F(x, y)$ meile teadmata, ei saa me ka arvutada riski $R(g_n)$. Samas on just risk klassifitseerija g_n headust iseloomustav suurus, mistõttu praktikas pakub huvi riski $R(g_n)$ ülemine hinnang (usalduspiir), mille leidmiseks kasutatakse tavaliselt empiirilist riski $R_n(g_n)$ (tuleta meelde (2.4.2)), mida saab valimi põhjal alati arvutada. Kui ülemine hinnang riskile $R(g_n)$ on väike, siis suure tõenäosusega on väike ka $R(g_n)$ ja see on hea. Teisest küljest aga suur risk $R(g_n)$ ei tähenda alati, et g_n on klassist $\mathcal{G}$ halvasti valitud. Tõepoolest, et $R(g_n) \geq \inf_{g \in \mathcal{G}} R(g)$, võib suur risk tähendada hoopis seda, et klass $\mathcal{G}$ on (antud jaotuse jaoks) ebasobiv või on üldkogumi jaotus klassifitseerimise mõttes halb. Seda, kas g_n on hästi või halvasti hinnatud mõõdab hindamisviga $R(g_n) - \inf_{g \in \mathcal{G}} R(g)$. Seetõttu ongi Vapnik-Cervonenkise teooria kesksed objektid:

- hindamisviga (*estimation error, excess risk*):

$$R(g_n) - \inf_{g \in \mathcal{G}} R(g)$$

- g_n empiirilise riski erinevus riskist (*generalization error*):

$$|R(g_n) - R_n(g_n)|.$$

Tuleta meelde, et $R_n(g_n)$ ja $R(g_n)$ on juhuslikud suurused, sest sõltuvad juhuslikust valimist D_n, kusjuures iga g korral

$$ER_n(g) = R(g)$$

(miks?). Seetõttu saame neile anda vaid tõenäosuslikke hinnanguid. Tüüpilised hinnangud on kujul

$$\mathbf{P}\big(R(g_n) - \inf_{g\in\mathcal{G}} R(g) > \epsilon\big) \le \delta_1(\epsilon, n, \mathcal{G}) \tag{2.4.7}$$

$$\mathbf{P}\big(|R(g_n) - R_n(g_n)| > \epsilon\big) \le \delta_2(\epsilon, n, \mathcal{G}), \tag{2.4.8}$$

kus $\delta_i(\epsilon, n, \mathcal{G})$ on (soovitavalt nulliks koonduvad) funktsioonid. Hinnangud (2.4.7) on (muuhulgas) olulised ka mõjususe tõestamisel. Kombineeridas hinnangud (2.4.8) ja (2.4.7) saame empiirilise riski abil hinnata ka suurust $\inf_{g\in\mathcal{G}} R(g)$:

$$\mathbf{P}(\inf_{g\in\mathcal{G}} R(g) \le R_n(g_n) + \epsilon\big) \ge \mathbf{P}\big(R(g_n) \le R_n(g_n) + \epsilon\big) \ge 1 - \delta_2(\epsilon, n, \mathcal{G}). \tag{2.4.9}$$

Tehisõppes avaldatakse ϵ tavaliselt δ kaudu ning sellisel juhul on (2.4.7) ning (2.4.9) kujul:

$$\text{tõenäosusega } 1-\delta \text{ (üle valimite): } R(g_n) - \inf_{g\in\mathcal{G}} R(g) \le \epsilon_1(\delta, n, \mathcal{G}) \tag{2.4.10}$$

$$\text{tõenäosusega } 1-\delta \text{ (üle valimite): } \inf_{g\in\mathcal{G}} R(g) \le R(g_n) \le R_n(g_n) + \epsilon_2(\delta, n, \mathcal{G}). \tag{2.4.11}$$

Neid nimetatakse teinekord ka **PAC** (*probably almost correct*) hinnanguteks. Seega olulised hinnangud on (2.4.7) ja (2.4.8), ülejäänud siintoodud hinnangud järelduvad nendest. Järgnev väga lihtne lemma näitab, et ERM-printsiibil saadud klassifikaatori korral järelduvad mõlemad hinnangud hinnangust ülemisest hinnangust juhuslikule suurusele

$$\sup_{g\in\mathcal{G}} |R_n(g) - R(g)|.$$

Lemma 2.1 (Vapnik, Chervonenkis, 1974) *Olgu $g_n \in \mathcal{G}$ suvaline klassifikaator, $\hat{g}_n \in \mathcal{G}$ olgu ERM-printsiibil leitud klassifikaator, st*

$$\hat{g}_n = \arg\inf_{g\in\mathcal{G}} R_n(g).$$

Siis

$$|R_n(g_n) - R(g_n)| \le \sup_{g\in\mathcal{G}} |R_n(g) - R(g)| \tag{2.4.12}$$

$$R(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g) \le 2\sup_{g\in\mathcal{G}} |R_n(g) - R(g)|. \tag{2.4.13}$$

Tõestus. Esimene võrratus on ilmne.
Teine võrratus: olgu $g^* \in \mathcal{G}$ selline, et (eeldame korraks, et leidub):

$$R(g^*) = \inf_{g\in\mathcal{G}} R(g).$$

Siis

$$\begin{aligned} R(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g) = R(\hat{g}_n) - R(g^*) &= R(\hat{g}_n) - R_n(\hat{g}_n) + R_n(\hat{g}_n) - R_n(g^*) + R_n(g^*) - R(g^*) \\ &\le R(\hat{g}_n) - R_n(\hat{g}_n) + R_n(g^*) - R(g^*) \\ &\le 2\sup_{g\in\mathcal{G}} |R_n(g) - R(g)|. \end{aligned}$$

Juhul, kui optimaalset g^* ei leidu, siis (inf definitsiooni kohaselt) iga $\epsilon > 0$ korral leidub $g_\epsilon \in \mathcal{G}$ nii, et $R(g_\epsilon) \le \inf_{g\in\mathcal{G}} R(g) + \epsilon$. Siis

$$\begin{aligned} R(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g) &\le R(\hat{g}_n) - R(g_\epsilon) + \epsilon \\ &\le 2\sup_{g\in\mathcal{G}} |R_n(g) - R(g)| + R_n(\hat{g}_n) - R_n(g_\epsilon) + \epsilon \\ &\le 2\sup_{g\in\mathcal{G}} |R_n(g) - R(g)| + \epsilon. \end{aligned}$$

Et ϵ oli suvaline, saame siit (2.4.13). ■

Järelikult on teoorias kesksel kohal hinnangud kujul

$$\mathbf{P}\Big(\sup_{g\in\mathcal{G}} |R_n(g) - R(g)| > \epsilon\Big) = \mathbf{P}\Big(\sup_{g\in\mathcal{G}} |R_n(g) - ER_n(g)| > \epsilon\Big) \le \delta(\epsilon, n, \mathcal{G}). \tag{2.4.14}$$

Hinnangud kujul (2.4.14) põhinevad tõenäosusteooriast tuntud **kontsentratsiooni-võrratustel**. Tõenäosus

$$\mathbf{P}\Big(\sup_{g\in\mathcal{G}} |R_n(g) - R(g)| > \epsilon\Big) \tag{2.4.15}$$

sõltub klassi $\mathcal{G}$ kompleksusest. Eespooltoodud lihtsast näitest nägime, et liiga suure klassi korral leidub alati selline g, et $R_n(g) = 0$ ning sellisel juhul iga $\epsilon < R(g)$ ja (mis kõige hullem) iga n korral tõenäosus (2.4.15) võrdub ühega. Kuidas klassifikaatorite hulga $\mathcal{G}$ kompleksust mõõta? Klassi võimsus – näiteks lõplike klasside korral selle elementide arv – pole ilmtingimata õige mõõdupuu. Nii näiteks võib klass, milles on väga palju üksteisega mingis mõttes sarnaseid funktsioone olla hoopis väiksema kompleksusega kui klass, mis koosneb vaid 10-st kuid teatud mõttes vaid väga erinevast funktsioonist. Vapnik-Cervonenkise teoorias mõõdetakse klassi $\mathcal{G}$ komplektsust nn. Vapnik-Cervonenkise dimensiooniga, millega nüüd tutvume.

2.5 VC dimensioon

Olgu $\mathcal{A}$ ruumi $\mathbb{R}^d$ alamhulkade klass..

Definitsioon 2.5 *Olgu $n \in \mathbb{N}$. Arvu*

$$\mathbb{S}_{\mathcal{A}}(n) := \max_{x_1,\ldots,x_n} |\{x_1,\ldots,x_n\} \cap A : A \in \mathcal{A}\}|$$

nimetame klassi $\mathcal{A}$ **tükelduskoefitsiendiks** (ik shatter coefficient).

Olgu $x_1,\ldots,x_n$ n-elemendiline hulk. Arv $|\{x_1,\ldots,x_n\} \cap A : A \in \mathcal{A}\}|$ näitab, mitu erinevat alamhulka saab klassi $\mathcal{A}$ elementide abil sellest hulgast välja eraldada (alamhulk on välja eraldatud, kui ta on mingi $A \in \mathcal{A}$ korral kujul $\{x_1,\ldots,x_n\} \cap A$). Tükelduskoefitsient on klassi $\mathcal{A}$ iseloomustav arv, mis näitab maksimaalset väljaeraldavate alamhulkade arvu (üle kõikide n-elemendiliste alamhulkade). On selge, et $\mathbb{S}_{\mathcal{A}}(n) \leq 2^n$ ning võrdus kehtib parajasti siis, kui leidub n hulk $\{x_1,\ldots,x_n\}$ nii, et selle kõikvõimalikud alamhulgad on võimalik klassi $\mathcal{A}$ elementide abil välja eraldeda. Kui see on nii, siis ütleme, et $\mathcal{A}$ **tükeldab** (ik *shatters*) hulga $\{x_1,\ldots,x_n\}$.

Näited:

1. Olgu $d = 1$ ja $\mathcal{A} := \{(-\infty, a] : a \in \mathbb{R}\}$. Siis $\mathbb{S}_{\mathcal{A}}(n) = n + 1$.
2. Olgu $d = 2$ ja $\mathcal{A} = \{(-\infty, a^1] \times (-\infty, a^2] : a = (a^1, a^2) \in \mathbb{R}^2\}$. Siis (ülesanne 2.4)
$$\mathbb{S}_{\mathcal{A}}(n) = 1 + \sum_{k=1}^{n}(n-k+1) = 1 + \sum_{k=1}^{n} k = 1 + \frac{n(n+1)}{2}.$$
3. Olgu $d = 1$ ja $\mathcal{A} := \{[a, b] : a, b \in \mathbb{R}\}$. Siis (ülesanne 2.4)
$$\mathbb{S}_{\mathcal{A}}(n) = 1 + \sum_{k=1}^{n}(n-k+1) = 1 + \frac{n(n+1)}{2}.$$
4. Olgu $d = 2$ ja $\mathcal{A} = \{x : w'x \geq b : w \in \mathbb{R}^2, b \in \mathbb{R}\}$. Seega $\mathcal{A}$ elemendid on pooltasandid. Siis (ülesanne 2.4)
$$\mathbb{S}_{\mathcal{A}}(2) = 4, \quad \mathbb{S}_{\mathcal{A}}(3) = 8, \quad \mathbb{S}_{\mathcal{A}}(4) = 14.$$
5. Eelmise näite üldistus. Olgu $\mathcal{A}$ kõikide ruumi $\mathbb{R}^d$ pooltasandite hulk. Siis iga $n \geq d+1$ korral
$$\mathbb{S}_{\mathcal{A}}(n) = 2\sum_{i=0}^{d}\binom{n-1}{i} \leq 2(n-1)^d + 2$$
([1], Cor. 13.1)

Ülesanne 2.4 *Tõesta näited 2, 3 ja 4.*

Tükelduskoefitsiendi omadused.

Teoreem 2.6 *Olgu $\mathcal{A}$ ja $\mathcal{B}$ hulga $\mathbb{R}^d$ alamhulkade klassid, olgu $n, m \geq 1$ täisarvud. Siis*

1. $\mathbb{S}_{\mathcal{A}}(n) \leq |\mathcal{A}|$;
2. $\mathbb{S}_{\mathcal{A}}(n+m) \leq \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{A}}(m)$;
3. *Kui* $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, *siis* $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n) + \mathbb{S}_{\mathcal{B}}(n)$;
4. *Kui* $\mathcal{C} = \{A^c : A \in \mathcal{A}\}$, *siis* $\mathbb{S}_{\mathcal{A}}(n) = \mathbb{S}_{\mathcal{C}}(n)$;
5. *Kui* $\mathcal{C} = \{A \cap B : A, \in \mathcal{A}$ *ja* $B \in \mathcal{B}\}$, *siis* $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$;
6. *Kui* $\mathcal{C} = \{A \cup B : A, \in \mathcal{A}$ *ja* $B \in \mathcal{B}\}$, *siis* $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$;
7. *Kui* $\mathcal{C} = \{A \times B : A, \in \mathcal{A}$ *ja* $B \in \mathcal{B}\}$, *siis* $\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$.

Ülesanne 2.5 *Tõesta teoreem.*

Definitsioon 2.7 *Klassi $\mathcal{A}$* **Vapnik-Cervonenkise (VC) dimensioon V** *on suurim n nii, et $\mathbb{S}_{\mathcal{A}}(n) = 2^n$. Kui $\mathbb{S}_{\mathcal{A}}(n) = 2^n$ iga n korral, siis $V = \infty$.*

VC dimensioon on korrektselt defineeritud, sest kui $\mathbb{S}_{\mathcal{A}}(n) < 2^n$, siis iga $m > n$ korral $\mathbb{S}_{\mathcal{A}}(m) < 2^m$.

Seega klassi $\mathcal{A}$ VC dimension on suurima hulga võimsus, mida see klass tükeldab. Teisisõnu, kui VC dimensioon on V, siis leidub vähemalt üks V-elemendiline hulk, mida klass $\mathcal{A}$ tükeldab, kuid iga $n > V$ korral pole ühtegi hulka $\{x_1, \ldots, x_n\}$ mida $\mathcal{A}$ tükeldaks.

Näited:

1. Olgu $d = 1$ ja $\mathcal{A} := \{(-\infty, a] : a \in \mathbb{R}\}$. Siis $V = 1$. Ühtegi kaheelemendilist hulka see klass ei tükeldada.
2. Olgu $d = 2$ ja $\mathcal{A} = \{(-\infty, a^1] \times (-\infty, a^2] : a = (a^1, a^2) \in \mathbb{R}^2\}$. Siis $\mathbb{S}_{\mathcal{A}}(2) = 4 = 2^2$ ja $\mathbb{S}_{\mathcal{A}}(3) = 1 + 6 = 7 < 2^3$. järelikult $V = 2$.
3. Eelnevat näidet saab üldistada järgnevalt: kui $\mathcal{A} = \{(-\infty, a^1] \times \cdots \times (-\infty, a^d] : (a^1, \ldots, a^d) \in \mathbb{R}^d\}$, siis $V = d$ (ülesanne 2.6).
4. Olgu $d = 1$ ja $\mathcal{A} := \{[a, b] : a, b \in \mathbb{R}\}$. Siis $\mathbb{S}_{\mathcal{A}}(2) = 4 = 2^2$ ja $\mathbb{S}_{\mathcal{A}}(3) = 1+6 = 7 < 2^3$. Järelikult $V = 2$.
5. Eelnevat näidet saab üldistada järgnevalt: kui $\mathcal{A} = \{[a^1, b^1] \times \cdots \times [a^d, b^d] : a, b \in \mathbb{R}^d\}$, so $\mathcal{A}$ on ristkülikute klass, siis $V = 2d$ ([1], Thm. 13.8).

6. Olgu $\mathcal{A}$ ruumi $\mathbb{R}^d$ pooltasandid. Et

$$\sum_{i=0}^{k} \binom{k}{i} = 2^k,$$

saame

$$S_{\mathcal{A}}(d+1) = 2\sum_{i=0}^{d} \binom{d}{i} = 2 \cdot 2^d = 2^{d+1}$$

$$S_{\mathcal{A}}(d+2) = 2\sum_{i=0}^{d} \binom{d+1}{i} = 2(2^{d+1} - 1) < 2^{d+1}.$$

Järelikult $V = d + 1$.

7. Olgu $\mathcal{A} = \{x \in \mathbb{R}^d : \|x - a\| \leq r, a \in \mathbb{R}^d, r \geq 0\}$ so $\mathcal{A}$ koosneb kõikidest keradest ruumis $\mathbb{R}^d$. Siis $V \leq d + 2$ ([1], Cor. 13.2).

8. Olgu $\mathcal{A}$ kõik kumerad hulknurgad ruumis $\mathbb{R}^2$. Selle klassi VC dimensioon on lõpmatu, st $V = \infty$. Tõepoolest, iga n korral see klass tükeldab hulga $x_1, \ldots, x_n$, kuis see hulk asetseb ringjoonel.

Ülesanne 2.6 *Tõesta näide 3.*

Kui klassi $\mathcal{A}$ VC-dimensioon V on lõplik, siis iga $n > V$ korral $\mathbb{S}_{\mathcal{A}}(n) < 2^n$. Järgmine tähtis lemma näitab aga, et sellisel tükelduskoefitsient kasvab ülimalt polünomiaalselt astmega V.

Lemma 2.2 (Saueri lemma) *Olgu $\mathcal{A}$ klass, mille VC dimensioon V on lõplik. Siis iga n korral*

$$\mathbb{S}_{\mathcal{A}} \leq \sum_{i=0}^{V} \binom{n}{i} \leq (n+1)^V$$

ja iga $n \geq V$ korral

$$\mathbb{S}_{\mathcal{A}} \leq \left(\frac{ne}{V}\right)^V$$

Seega kui $V \geq 3$, siis $\mathbb{S}_{\mathcal{A}} \leq n^V$.

Lemma tõestuse võib leida [1], Thm 13.2 ja Thm 13.3, samuti [2], Corollary 1.3.

2.6 Vapnik-Cervonenkise võrratus ja riski hinnangud

Empiiriline mõõt. Olgu $Z_1, \dots, Z_n$ iid d-dimensionaalsed juhuslikud vektorid. Iga hulga $A \subset \mathbb{R}^d$ korral selle **empiiriline mõõt** $P_n(A)$ on hulka A kuuluvate vektorite proportsioon:

$$P_n(A) := \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i \in A\}}.$$

Fikseeritud (mõõtuva) A korral on empiiriline mõõt $P_n(A)$ juhuslik suurus keskväärtusega

$$EP_n(A) = \mathbf{P}(Z_1 \in A) =: P(A),$$

kusjuures hulka A kuuluvate vektorite arv on binoomjaotusega, st $nP_n(A) \sim B(n, P(A))$. (Tugevast) suurte arvude seadusest järeldub, et $P_n(A) \to P(A)$, p.k.. Veel enam, tuntud **Höffdingi võrratus** väidab, et iga $\epsilon > 0$

$$\mathbf{P}(|P_n(A) - P(A)| > \epsilon) \le 2 \exp[-2n\epsilon^2]. \tag{2.6.1}$$

Koondumine $P_n(A) \to P(A)$ p.k. (mis muidugi järeldub suurte arvude seadusest) järeldub Borel-Cantelli lemma kaudu vahetult võrratusest (2.6.1).

Ülesanne 2.7 *Olgu $\mathcal{A}$ lõplik mõõtuvate hulkade klass. Tõesta*

$$\mathbf{P}(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon) \le 2|\mathcal{A}| \exp[-2n\epsilon^2]. \tag{2.6.2}$$

VC võrratus. Võrratus (2.6.2) annab eksponentsiaalse tõkke lõpliku klassi korral, aga kui $|\mathcal{A}| = \infty$, on see võrratus kasutu. Fundamentalne Vapnik-Cervonenkise võrratus üldistab võrratust (2.6.2) lõpmatutele klassidele. Võrratus on kasulik, kui klassi VC-dimensioon on lõplik.

Teoreem 2.8 (Vapnik ja Cervonenkis' 71) *Iga klassi $\mathcal{A}$, iga täisarvu n ja $\epsilon > 0$ korral*

$$\mathbf{P}\Big(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon \Big) \le 8 \mathbb{S}_{\mathcal{A}}(n) \exp[-\frac{\epsilon^2}{32} n] \tag{2.6.3}$$

$$E\Big(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \Big) \le 2 \sqrt{\frac{\ln \mathbb{S}_{\mathcal{A}}(n) + \ln 2}{n}}. \tag{2.6.4}$$

Võrratuse (2.6.3) tõestust vaata [1], Thm. 12.5; võrratuse (2.6.4) tõestus on [2], Thm 1.9.

Neist kahest võrratusest on teedrajav ja olulisem (ja üldse mitte triviaalne) just esimene võrratus (2.6.3). Teine võrratus järeldub sisuliselt esimesest. Veendumaks, et võrratusega (2.6.4) sarnaseid võrratusi (ehk teiste konstantidega) saab suhteliselt kergelt järeldada esimesest võrratusest (2.6.3) vaatleme järgmist lauset.

Lause 2.1 *Olgu Z mittenegatiivne juhuslik suurus nii, et iga $\epsilon > 0$ korral kehtib*

$$\mathbf{P}(Z > \epsilon) \leq C(n) \exp[-A(n)\epsilon^2], \tag{2.6.5}$$

kus $A(n) > 0$ ja $C(n)$ ei sõltu ϵ-st. Siis

$$EZ \leq \sqrt{\frac{\ln C(n) + 1}{A(n)}}. \tag{2.6.6}$$

Ülesanne 2.8 *Tõesta lause.*
Näpunäide: kasuta

$$(EZ)^2 \leq EZ^2 = \int_0^u \mathbf{P}(Z^2 > t)dt + \int_u^\infty \mathbf{P}(Z^2 > t)dt \leq u + \int_u^\infty \mathbf{P}(Z^2 > t)dt.$$

Leia u, mis minimiseerib ülemist tõket.

Järeldus 2.6.1 *Kui klassi $\mathcal{A}$ VC-dimensioon V on lõplik, siis*

$$\mathbf{P}\Big(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon \Big) \leq 8(n+1)^V \exp[-\frac{\epsilon^2}{32}n] \tag{2.6.7}$$

$$E\Big(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \Big) \leq 2\sqrt{\frac{V \ln(n+1) + \ln 2}{n}}. \tag{2.6.8}$$

Tõestus. Kasuta võrratust $\mathbb{S}_\mathcal{A} \leq (n+1)^V$ võrratustes (2.6.3) ja (2.6.4). ■

Ülesanne 2.9 *Järelda seosest (2.6.7) Clivenko-Cantelli teoreem:*

$$\sup_x |F_n(x) - F(x)| \to 0, \ \ a.s..$$

Riski hinnangud sümmeetrilise kaofunktsooni korral. Vaatleme sümmeetrilist kaofunktsiooni

$$L(g(x), y) = I_{A_g}, \quad A_g := \{(x, y) : g(x) \neq y\} \subset \mathbb{R}^d \times \{0, \ldots, k-1\}.$$

Seega, võttes $Z_i = (X_i, Y_i)$ saame riski $R(g)$ ja empiirilise riski $R_n(g)$ avaldada kui hulga A_g tõenäosus ning empiiriline mõõt:

$$R(g) = P(A_g), \quad R_n(g) = P_n(A_g).$$

Nendest võrdustest paremaks arusaamiseks tuleta meelde, et

$$R(g) = \mathbf{P}(g(X) \neq Y) = \mathbf{P}((X_1, Y_1) \in A_g) = P(A_g),$$
$$R_n(g) = \frac{1}{n}\sum_{i=1}^n I_{\{g(X_i)\neq Y_i\}} = \frac{1}{n}\sum_{i=1}^n I_{\{(X_i,Y_i)\in A_g\}} = P_n(A_g).$$

Nüüd (2.6.3) ja (2.6.4) on

$$\mathbf{P}\Big(\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|>\epsilon\Big)\le 8\mathbb{S}_{\mathcal{A}_\mathcal{G}}(n)\exp[-\frac{n\epsilon^2}{32}] \tag{2.6.9}$$

$$E\Big(\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|\Big)\le 2\sqrt{\frac{\ln\mathbb{S}_{\mathcal{A}_\mathcal{G}}(n)+\ln 2}{n}}, \tag{2.6.10}$$

kus

$$\mathcal{A}_\mathcal{G}=\{A_g : g\in\mathcal{G}\}.$$

Kui klassi $\mathcal{A}_\mathcal{G}$ VC-dimensioon on lõplik, siis, just nagu võrratustes (2.6.7) ja (2.6.8), saame selle kaudu anda ülemise hinnangu koefitsiondile $\mathbb{S}_{\mathcal{A}_\mathcal{G}}(n)$. Klassi $\mathcal{A}_\mathcal{G}$ VC-dimensiooni nimetatakse teinekord $\mathcal{G}$ **graafikudimensiooniks** (*graph-dimension of* $\mathcal{G}$).

Riski hinnangud sümmeetrilise kao ja kahe klassi korral. Kui klasse on kaks, st $k=2$, siis iga klassifitseerija g on kujul I_A (binaarne klassifitseerija) nii et klassifitseerijate klass $\mathcal{G}$ on üksüheses vastavuses hulkade klassiga $\mathcal{A}$. Seetõttu klassi $\mathcal{A}$ tükelduskoefitsienti ja VC dimensiooni nimetame ka (vastavalt) **klassi $\mathcal{G}$ tükelduskoefitsiendiks** ja klassi **klassi $\mathcal{G}$ VC dimensiooniks**. Iga binaarse klassifikaatori $g=I_A$, korral hulk A_g on

$$A_g=\{A\times\{0\}\cup A^c\times\{1\}\}.$$

Tuleta meelde, et $A_g\subset\mathbb{R}^d\times\{0,1\}$, kuid $A\subset\mathbb{R}^d$. Seega oleme binaarsete klassifitseerijate klassi korral defineerinud nii graafikudimensiooni kui ka VC dimensiooni. Samas pole raske näha ([1], Thm. 13.1) et klassi $\mathcal{A}_\mathcal{G}$ tükelduskoefitsient on sama mis $\mathcal{A}$ tükelduskoefitsient nii, et binaarsete klassifitseerijate klassi $\mathcal{G}$ graafikudimensioon on sama, mis klassi $\mathcal{G}$ VC dimensioon ja see on (definitsiooni kohaselt) klassi $\mathcal{A}$ VC dimensioon. Seega sümmeetrilise kaofunktsiooni ja kahe klassi korral riski hinnangud (2.6.9) ja (2.6.10) on

$$\mathbf{P}\Big(\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|>\epsilon\Big)\le 8\mathbb{S}_{\mathcal{A}}(n)\exp[-\frac{n\epsilon^2}{32}]\le 8(n+1)^V\exp[-\frac{n\epsilon^2}{32}] \tag{2.6.11}$$

$$E\Big(\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|\Big)\le 2\sqrt{\frac{\ln\mathbb{S}_{\mathcal{A}}(n)+\ln 2}{n}}\le 2\sqrt{\frac{V\ln(n+1)+\ln 2}{n}}, \tag{2.6.12}$$

kus V on $\mathcal{G}$ (klassi $\mathcal{A}$) VC dimensioon. Kui $V<\infty$, siis nende võrratuste parem pool koondub nulliks eksponentsiaalse kiirusega.

Ülesanne 2.10 *Tõesta võrratused (2.6.13), (2.6.14) ja (2.6.15):*

$$ER(\hat{g}_n)-\inf_{g\in\mathcal{G}}R(g)\le 4\sqrt{\frac{V\ln(n+1)+\ln 2}{n}} \tag{2.6.13}$$

$$\mathbf{P}\Big(R(\hat{g}_n)-\inf_{g\in\mathcal{G}}R(g)>\epsilon\Big)\le 8(n+1)^V\exp[-\frac{n\epsilon^2}{128}], \tag{2.6.14}$$

ja tõenäosusega $1-\delta$

$$R(g_n) \le R_n(g_n) + 4\sqrt{\frac{2(V\ln(n+1) - \ln\delta + \ln 8)}{n}}. \tag{2.6.15}$$

kus $\hat{g} \in \mathcal{G}$ *on saadud ERM-printsiibil ja* $g_n \in \mathcal{G}$ *on suvaline klassifitseerija.*

Ülesanne 2.11 *Tõesta, et kui* $V < \infty$*, siis*

$$R(\hat{g}_n) \to \inf_{g\in\mathcal{G}} R(g) \quad a.s., \quad ER(\hat{g}_n) \to \inf_{g\in\mathcal{G}} R(g). \tag{2.6.16}$$

Märkused: 1. Võrratused (2.6.3) ja (2.6.4) pole kõige täpsemad. Vapniku ja Cervonenkise originaalartiklis oli eksponendi $\frac{-n\epsilon^2}{32}$ asemel mõnevõrra parem $\frac{-n\epsilon^2}{8}$. Kirjandusest võib leida veelgi paremaid eksponente, vaata näiteks [1], 12.8, samuti [2]. Vapniku raamatus ([5], Thm 4.1) on VC võrratus (2.6.3) kujul

$$\mathbf{P}\Big(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\Big) \le 4\mathbb{S}_{\mathcal{A}}(2n)\exp[-(\epsilon - \frac{1}{n})^2 n]. \tag{2.6.17}$$

On selge, et täpsem VC võrratus annab ka täpsemad hinnangud (2.6.14) ja (2.6.15). Samas kõik toodud hinnangud on kujul $A\exp[-c\epsilon^2 n]$, kus A ja c on mingid konstandid ja seda kuju ei saa üldiselt parandada. Kuigi konkreetse n korral on konstantidel tähtsus ja parem eksponent võib anda oluliselt täpsema hinnangu, siis asümptootiliselt on oluline vaid kuju $A\exp[-c\epsilon^2 n]$. Selgub, et erijuhul $\inf_{g\in\mathcal{G}} R(g) = 0$ (väga kitsendav tingimus mis samaaegselt eeldab et Bayesi risk on 0 ja klass $\mathcal{G}$ sisaldab Bayesi klassifitseerijat) saab ka seda kuju muuta ja siis hinnang (2.6.14) on kujul $A\exp[-c\epsilon n]$ st ϵ^2 on asendatud palju suurema arvuga ϵ ([1], 12.7). Samas selle võrratuse kasutamiseks PAC võrratuses (2.6.15) peame teadma, et lisatingumus $\inf_{g\in\mathcal{G}} R(g) = 0$ kindlasti kehtib.
Selgub, et ka võrratust (2.6.13) saab parandada nii, et $\ln(n+1)$ tegur kaob. Nimelt kehtib võrratus ([2], 1.4.6)

$$ER(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g) \le c\sqrt{\frac{V}{n}}, \tag{2.6.18}$$

kus c on konstant. Kui n on väga suur, siis (2.6.18) on parem kui (2.6.13).

2. Kui kaofunktsioon L on sümmeetriline kuid klasse on enam kui kaks, siis võrratused (2.6.13), (2.6.14) ja (2.6.15) kehtivad kui V on klassi $\mathcal{G}$ graafikudimensioon.

2.7 Regulariseerimise alused

Tuleta meelde, et sümmeetrilise kaofunktsiooni ja lõpliku graafikudimensiooniga klassi $\mathcal{G}$ korral kehtivad koondumised (2.6.16):

$$R(\hat{g}_n) \to \inf_{g\in\mathcal{G}} R(g) \quad \text{a.s.}, \quad ER(\hat{g}_n) \to \inf_{g\in\mathcal{G}} R(g).$$

Kui klass $\mathcal{G}$ oleks selline, et

$$\inf_{g\in\mathcal{G}} R(g) = R^* \tag{2.7.1}$$

(st lähendamisviga on 0), siis nimetatud koondumised tähendaksid ERM-printsiibil saadud reegli $\hat{g}_n$ tugevat mõjusust. Paraku tähendab seos (2.7.1) üldiselt, et valitud klass $\mathcal{G}$ on liiga kompleksne ning tema graafikudimensioon lõpmatu. Sellisel juhul ei kehti Vapnik-Chervonenkise võrratus ega ükski eelmises osas toodud hinnang ning loomulikult mitte ka koondumine (2.6.16). Saamaks siiski mõjusat hinnangut, vaadeldakse klasside jada $\mathcal{G}_n$, kus klasside kompleksus suureneb valimimahu n kasvades nii, et lähenemisviga väheneks. Samas peab vähenema ka hindamisviga ehk klassi $\mathcal{G}_n$ kompleksuse kasv olema piisavalt aeglane. Siit kasvab välja **regulariseerimisteooria** (*complexity regularization*).

Sõelad (*sieves*). Olgu

$$\mathcal{G}_1 \subset \mathcal{G}_2 \subset \cdots \tag{2.7.2}$$

sellised klassifikaatorite klassid, et k suurenedes hindamisviga läheneb nullile, st

$$\lim_k \inf_{g\in\mathcal{G}_k} R(g) = R^*. \tag{2.7.3}$$

Et klassid on üksteisesse sisestatud, siis nende kompleksus (graafikudimensioon) suureneb. Olgu

$$n \mapsto k(n)$$

teatav kasvav funktsioon. Iga n korral valime treeningandmete põhjal hulgast $\mathcal{G}_{k(n)}$ klassifikaatori g_n (näiteks ERM abil). Seega

$$R(g_n) - R^* = \big(R(g_n) - \inf_{g\in\mathcal{G}_{k(n)}} R(g)\big) + \big(\inf_{g\in\mathcal{G}_{k(n)}} R(g) - R^*\big).$$

Et k on kasvav, siis n kasvades kasvab klassi $\mathcal{G}_{k(n)}$ kompleksus ja seega ka hindamisviga. Samal ajal väheneb lähenemisviga. Kui indeksid $k(n)$ õnnestub valida nii, et $k(n) \to \infty$ kuid klasside $\mathcal{G}_{k(n)}$ kompleksuse kasv on kontrolli all, on võimalik saavutada reegli mõjusus. Näitena vaatleme järgmist teoreemi, mis kahe klassi ja sümmeetrilise kaofunktsiooni korral garanteerib selliselt saadud ERM-reegli mõjususe ([1], Thm. 18.1).

Teoreem 2.9 *Olgu klass kaks. Olgu klassid $\mathcal{G}_k$ sellised, et iga (X, Y) jaotuse korral kehtib (2.7.3). Olgu $k(n)$ sellised täisarvud, et*

$$\lim_n k(n) = \infty, \quad \lim_n \frac{V_{k(n)} \ln n}{n} = 0, \tag{2.7.4}$$

kus V_k on klassi $\mathcal{G}_k$ VC dimensioon. Siis reegel $\{g_n\}$, kus $g_n = \hat{g}_{k(n)}$ on universaalselt tugevalt mõjus.

Ülaltoodud teoreemis ei pruugi klassid olla ilmtingimata üksteistesse sisestatud, kuid enamasti nad seda on.

Ülesanne 2.12 *Tõestada, et teoreemi eeldustel on $\{g_n\}$ universaalselt mõjus.*

Ülesanne 2.13 *Tõestada teoreem.*

SRM (*structural risk minimization*). SRM on ERM-printsiibi samaaegne rakendamine mitmele klassile. Olgu antud klassid $\mathcal{G}_k$ nagu (2.7.2), kusjuures V_k olgu klassi $\mathcal{G}_k$ kompleksust mõõtev suurus, sümmeetrilise kaofunktsiooni korral enamasti graafikudimensioon. Siis defineeritakse iga k ja n korral klassi $\mathcal{G}_k$ kompleksust mõõtev **karistusliige** (*penalty term*) $C(V_k, n)$. Funktsioon $C(V, n)$ on reeglina V järgi kasvav ja n järgi kahanev. Näiteks

$$C(V, n) = a\sqrt{\frac{V \ln n + b}{n}},$$

kus a, b on konstandid. Ülaltoodud karistusliige tuleneb võrratustest (2.6.13) või ka (2.6.15).

SRM käib nii: iga n ja k korral valitakse klassist $\mathcal{G}_k$ empiirilist riski minimiseeriv funktsioon $\hat{g}_{k,n}$ ning seejärel leitakse iga n korral hulgast $\hat{g}_{1,n}, \hat{g}_{2,n}, \ldots$ selline, mis minimiseerib

$$R_n(\hat{g}_{k,n}) + C(V_k, n).$$

Teisisõnu

$$g_n = \hat{g}_{k(n),n}, \tag{2.7.5}$$

kus

$$k(n) = \arg\inf_k \Big(R_n(\hat{g}_{k,n}) + C(V_k, n) \Big).$$

Erinevus sõelade meetoodist seisneb sellest, et funktsioon $n \mapsto k(n)$ pole ette antud vaid $k(n)$ valitakse andmetest lähtuvalt automaatselt. Pane tähele: kui $C(V_k, n)$ on väga väike võrreldes riskiga $R_n(\hat{g}_{k,n})$, siis summa $R_n(\hat{g}_{k,n}) + C(V_k, n)$ minimiseerimine on sisuliselt sama, mis riski $R_n(\hat{g}_{k,n})$ minimiseerimine ja see tähendab, et $k(n)$ tuleb väga suur – ülesobituvus. Teisalt, kui $C(V_k, n)$ on väga suur võrreldes riskiga $R_n(\hat{g}_{k,n})$, siis summa $R_n(\hat{g}_{k,n}) + C(V_k, n)$ minimiseerimine on sisuliselt karistusliikme $C(V_k, n)$ minimiseerimine ja et viimane suureneb k kasvades, tuleb $k(n)$ sellisel juhul liiga väike – alasobituvus. Seega küsimus korrektselt valitud karistusliikmest on SRM korral kesksel kohal.

Ülesanne 2.14 *Olgu $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \cdots$. Tõestada, et klassifitseerija (2.7.5) võib leida ka nii*

$$g_n = \arg\inf_{\cup_k \mathcal{G}_k} \big(R_n(g) + C(g, n) \big), \tag{2.7.6}$$

kus $C(g, n) = C(V_k, n)$, kui $g \in \mathcal{G}_k / \mathcal{G}_{k-1}$.

Selgitame SRM-reegli olemust ja häid omadusi. Olgu kaofunktsioon L sümmeetriline ja V_k graafikudimensioon. Tuletame meelda, et $\hat{g}_{k,n} \in \mathcal{G}_k$ keskmise (üle valimite) riski erinevus Bayesi riskist avaldub teadupoolest hindamisvea ja lähendamisvea summana, kusjuures hindamisviga saab ülalt hinnata võrratusega (2.6.13):

$$\begin{aligned} ER(\hat{g}_{k,n}) - R^* &= \Big(ER(\hat{g}_{k,n}) - \inf_{g \in \mathcal{G}_k} R(g) \Big) + \Big(\inf_{g \in \mathcal{G}_k} R(g) - R^* \Big) \\ &\le 4\sqrt{\frac{V_k \ln(n+1) + \ln 2}{n}} + \Big(\inf_{g \in \mathcal{G}_k} R(g) - R^* \Big). \end{aligned}$$

Saamaks head hinnangut oleks loomulik valida iga n korral selline $k(n)$ mis minimiseerib ülaltoodud võrratuse paremat poolt üle kõigi k-de. Kahjuks seda me teha ei saa, sest suurus $\inf_{g\in\mathcal{G}_k} R(g)$ on meile tundmatu. Regulariseerimisalases kirjanduses nimetatakse tundmatut $k(n)$ teinekord ka *oraakliks*. Selgub aga, et SRM teeb minimiseerimise töö meie eest ära ja SRM printsiibil saadud klassifitseerija g_n keskmine risk erineb Bayesi riskist ligikaudu (konstantide täpsusega) samapalju kui oraakli abil saadud keskmine risk $ER(\hat{g}_{k(n),n})$. Näitena nn *oraaklivõrratusest* vaatleme järgmist teoreemi ([2] Thm. 1.20 ja 1.6.3), mis on tõestatud juhul kui kaofunktsioon on sümmeetriline ja karistusliige on

$$C(V_k,n)=2\sqrt{\frac{V_k\ln(n+1)+\ln 2}{n}}+\sqrt{\frac{\ln k}{n}}. \tag{2.7.7}$$

Teoreem 2.10 *Olgu klasside $\mathcal{G}_k$ graafikudimensioonid V_k lõplikud ning g_n olgu karistusliikme (2.7.7) abil saadud SRM-klassifitseerija. Siis*

$$ER(g_n)-R^*\le\min_k\left[\sqrt{\frac{V_k\ln(n+1)+\ln 2}{n}}+\Big(\inf_{g\in\mathcal{G}_k}R(g)-R^*\Big)+\sqrt{\frac{\ln k}{n}}\right]+\sqrt{\frac{1}{2n}}.$$

Järeldus 2.7.1 *Olgu klassid $\mathcal{G}_k$ sellised, et iga (X,Y) jaotuse korral kehtib (2.7.3). Siis teoreemis 2.10 defineeritud SRM-reegel on universaalselt mõjus.*

Ülesanne 2.15 *Tõestada järeldus.*

Järgnev teoreem ([1], Thm. 18.2) väidab, et SRM-printsiibil saadud klassifiteseerija on ka (universaalselt) tugevalt mõjus. Teoreem on tõestatud juhul kui klasse on kaks (seega V_k on klassi $\mathcal{G}_k$ VC dimensioon), kaofunktsioon sümmeetriline ja karistusliige on kujul

$$C(V_k,n)=\sqrt{\frac{32V_k\ln n+1}{n}}. \tag{2.7.8}$$

Teoreem 2.11 *Olgu klassid $\mathcal{G}_k$ sellised, et iga (X,Y) jaotuse korral kehtib (2.7.3). Olgu klasside $\mathcal{G}_k$ VC dimensioonid V_k lõplikud, kusjuures kehtib*

$$\sum_{k=1}^{\infty}e^{-V_k}<\infty. \tag{2.7.9}$$

Olgu g_n karistusliikmega (2.7.8) SRM-reegel. Siis $\{g_n\}$ on universaalselt mõjus.

Pane tähele: et V_k on täisarv, on eeldus (2.7.9) täideteud, kui iga k korral $V_k<V_{k+1}$. Kui see nii pole, saab valida sellise alamjada.

Märkus koondumiskiirusest: Oletame, et Bayesi klassifitseerija g^* kuulub ühte klassidest $\mathcal{G}_{k_o}$, st $g^*\in\cup_k\mathcal{G}_k$,. Siis ERM-printsiibil saadud klassifitseerija üle selle klassi, nimelt $g_{k_o,n}$ rahuldab võrratust (2.6.18)

$$ER(\hat{g}_{k_o,n})-R^*\le C(k_o)\sqrt{\frac{1}{n}},$$

kus $C(k_o)$ on arvust k_o sõltuv konstant. Seega, kui me teaksime "õiget klassi"$\mathcal{G}_{k_o}$, me kasutaksime vaid seda klassi ja valiksime sealt ERM printsiibil klassifitseerija. Sellisel juhul meie reegel $\{g_n\}$, kus $g_n = \hat{g}_{k_o,n}$ oleks selline, et kehtib

$$ER(g_n) - R^* = O\Big(\sqrt{\frac{1}{n}}\Big).$$

Saab näidata, et see on parim võimalik koondumiskiirus (minimax mõttes üle kõikide jaotuste). Teisest küljest aga, teoreemist 2.10 järeldub, et SRM printsiibil saadud reegli $\{g_n\}$ keskmine risk koondub kiirusega:

$$ER(g_n) - R^* \le \sqrt{\frac{V_{k_o}\ln(n+1)+\ln 2}{n}} + \sqrt{\frac{\ln k_o}{n}} + \sqrt{\frac{1}{2n}} \le B(k_o)\sqrt{\frac{\ln(n+1)+\ln 2}{n}},$$

kus $B(k_o)$, on k_o-st sõltuv konstant. Seega teadmata "õiget klassi"$\mathcal{G}_{k_o}$, SRM-printsiip annab meile klassifitseerija g_n nii, et $ER(g_n) - R^*$ koondub nulliks kiirusega

$$ER(g_n) - R^* = O\Big(\sqrt{\frac{\ln n}{n}}\Big).$$

See on lähedane (vaid $\sqrt{\ln n}$-teguriga korrutatud) optimaalsele kiirusele, mille saavutaksime kui teaksime k_o. Selline peaaegu parima koondumiskiiruse saavutamine näitab, et SRM printsiip on põhimõtteliselt õigem kui sõelade meetod, sest seal kiirus sõltub etteantud jadast $k(n)$.

Üldine regulariseerimine. Olgu $\mathcal{G}$ on lai funktsioonide klass (näiteks lõpmatu graafikudimensiooniga), kuid igale funktsioonile defineerime klassifitseerija kompleksusest sõltuva mõõtva karistusliikme $C(g,n)$. Klassifitseerija g_n leiame nüüd nii

$$g_n = \arg\inf_{g\in\mathcal{G}}\Big(R_n(g) + C(n,g)\Big). \tag{2.7.10}$$

Karistusliige $C(g,n)$ sõltub funktsioonist g enamasti läbi mingi g kompleksust mõõtva funktsiooni $v(g)$. Nagu ikka, on $C(v,n)$ v järgi kasvav ja n järgi kahanev. Funktsioon v sõltub ülesande püstitusest, see võib olla g norm (mingis ruumis), tema kirjeldamiseks kasutatava koodi pikkus, eelmõõdust sõltuv funktsioon vms. Paneme tähele, et (2.7.6) on erijuht üldisemast seosest (2.7.10), kus karistusfunktsioon $g \mapsto C(g,n)$ on konstantne klassidel $\mathcal{G}_1, \mathcal{G}_k/\mathcal{G}_{k-1}$, $k = 2,3,\ldots$. Seega, kui $g \in \mathcal{G}_k/\mathcal{G}_{k-1}$, siis $v(g) = V_k$.

Peatükk 3

Lineaarsed klassifitseerijad

3.1 Meeldetuletus

3.1.1 Hüpertasand ja punkti kaugus sellest

Normeeritud ruum. Olgu f normeeritud ruumil $\mathbb{X}$ antud lineaarne funktsionaal ning H olgu funktsionaali f abil defineeritud affiine hulk ehk **hüpertasand** (ik *hyperplane*):

$$H = \{x \in \mathbb{X} : f(x) = c\} = \{x \in \mathbb{X} : f(x) - c = 0\} = \{x \in \mathbb{X} : \frac{1}{\|f\|}\big(f(x) - c\big) = 0\},$$

kus $c \in \mathbb{R}$. Suvalise punkti $y \in \mathbb{X}$ kaugus hüpertasandist H on

$$d(y, H) := \inf_{x \in H} \|y - x\| = \frac{1}{\|f\|}\big|f(y) - c\big|. \tag{3.1.1}$$

Hilberti ruum. Juhul, kui $\mathbb{X}$ on Hilberti ruum (või lihtsalt sakalaarkorrutisega ruum), on lineaarne funktsionaal kujul $f(x) = \langle w, x\rangle$, kus $w \in \mathbb{X}$. Hüpertasand Hilberti ruumis on

$$H := \{x : \langle w, x\rangle + w_0 = 0\}$$

ja punkti $y \in \mathbb{X}$ kaugus sellest on järelikult

$$\frac{1}{\|w\|}\big|\langle w, y\rangle + w_0\big|.$$

Võttes (üldisust kitsendamata) w sellise, et $\|w\| = 1$ ning defineerides

$$g(x) := \langle w, x\rangle + w_0,$$

saame, et punkti y kaugus hüperasandist $\{x \in \mathbb{X} : g(x) = 0\}$ on $|g(y)|$.

Eukelidiline ruum. Erijuhul kui $\mathbb{X} = \mathbb{R}^d$ on $\langle w, x\rangle = w'x$. Seega punkti y kaugus hüpertasandist

$$H := \{x \in \mathbb{X} : w'x + w_0 = 0\}$$

on

$$d(y, H) = \frac{|w'y + w_0|}{\|w\|} = \frac{|g(y)|}{\|w\|}, \text{ kus } g(x) = w'x + w_0.$$

3.1.2 Kovariatsioonimaatriks ja selle lahutus

Olgu X, Y juhuslikud suurused. Tuleta meelde, et

$$D(X) = E(D[X|Y]) + D(E[X|Y]).$$

Olgu X juhuslik vektor, keskväärtusega $m = EX$. Vektori X *kovariatsioonimaatriks* on

$$\mathrm{Cov}(X) = E(X-m)(X-m)' = E(XX') - mm'. \tag{3.1.2}$$

Lause 3.1 *Olgu Y juhuslik suurus. Siis*

$$\mathrm{Cov}(X) = E\big(E\big[(X-E(X|Y))(X-E(X|Y))'|Y\big]\big) + E\big((E(X|Y)-m)(E(X|Y)-m)'\big)$$

ehk

$$\mathrm{Cov}(X) = E\big(\mathrm{Cov}[X|Y]\big) + \mathrm{Cov}\big(E[X|Y]\big). \tag{3.1.3}$$

Tõestus.

$$E(X-m)(X-m)' = E\Big(E[(X-m)(X-m)'|Y]\Big) =$$
$$E\Big(E\big[\big(X-E(X|Y)+E(X|Y)-m\big)\big(X-E(X|Y)+E(X|Y)-m\big)'|Y\big]\Big).$$

Tinglik keskväärtus on lineaarne, millest

$$E[(X-E(Y|X))(E(X|Y)-m)'|Y] = E[(E(X|Y)-m)(X-E(X|Y))'|Y] = 0.$$

Seega

$$E\big[\big(X-E(X|Y)+E(X|Y)-m\big)\big(X-E(X|Y)+E(X|Y)-m\big)'|Y\big] =$$
$$E\big[\big(X-E(X|Y)\big)\big(X-E(X|Y)\big)'|Y\big] + E[(E(X|Y)-m)(E(X|Y)-m)'|Y].$$

■

Olgu Y võimalike väärtuste hulk arvud $0, 1, \ldots, k-1$ ja olgu

$$m_i := E[X|Y=i], \quad \pi_i = \mathbf{P}(Y=i), \quad \Sigma_i := \mathrm{Cov}[X|Y=i], \quad \Sigma_X := \mathrm{Cov(X)}.$$

Siis (3.1.3) on järgmine:

$$\Sigma_X = \sum_{i=0}^{k-1} \pi_i \Sigma_i + \sum_{i=0}^{k-1} \pi_i (m_i - m)(m_i - m)' \tag{3.1.4}$$

Tihti tähistatakse

$$\Sigma_W := \sum_{i=0}^{k-1} \pi_i \Sigma_i, \quad \Sigma_B := \sum_{i=0}^{k-1} \pi_i (m_i - m)(m_i - m)'$$

ja siis (3.1.4) on $\Sigma_X = \Sigma_W + \Sigma_B$. Viimast võib interpreteerida kui tunnuse X hajuvuse lahutumist klassisises hajuvuse Σ_W (*within-class*) ja klassidevahelise hajuvuse Σ_B (*between-class*) summaks.

Ülesanne 3.1 *Veendu, et kui klasse on kaks, siis*

$$\Sigma_B = \pi_1 \pi_0 (m_1 - m_0)(m_1 - m_0)'. \tag{3.1.5}$$

Valimi kovariatsioonimaatriks. Kui juhusliku vektori (X, Y) jaotuse asemel on meil vaid treeningvalim (2.1.1), hinnatakse keskväärtusi, tõenäosusi ja kovariatsioonimaatrikseid järgmiselt:

$$\hat{\Sigma}_X := \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{m})(x_j - \hat{m})', \quad \hat{m} := \frac{1}{n}\sum_{j=1}^{n} x_j,$$
$$\hat{\Sigma}_i := \frac{1}{n_i}\sum_{j:y_j=i}(x_j - \hat{m}_i)(x_j - \hat{m}_i)', \quad \hat{\pi}_i := \frac{n_i}{n}, \quad \hat{m}_i = \frac{1}{n_i}\sum_{j:y_j=i} x_j, \quad i = 0, \ldots, k-1$$

kus n_i on klassi i kuuluvate valimielementide arv. Seos (3.1.4) on siis

$$\begin{aligned}\hat{\Sigma}_X &= \sum_{i=0}^{k-1}\hat{\pi}_i\hat{\Sigma}_i + \sum_{i=0}^{k-1}\hat{\pi}_i(\hat{m}_i - m)(\hat{m}_i - m)' \\ &= \frac{1}{n}\sum_{i=0}^{k-1}\sum_{j:y_j=i}(x_j - \hat{m}_i)(x_j - \hat{m}_i)' + \frac{1}{n}\sum_{i=0}^{k-1} n_i(\hat{m}_i - m)(\hat{m}_i - m)'.\end{aligned}$$

Seega korrutades mõlemad pooled n-ga, saame

$$S := \sum_{j=1}^{n}(x_j - \hat{m})(x_j - \hat{m})' = \sum_{i=0}^{k-1}\sum_{j:y_j=i}(x_j - \hat{m}_i)(x_j - \hat{m}_i)' + \sum_{i=0}^{k-1} n_i(\hat{m}_i - m)(\hat{m}_i - m)' \tag{3.1.6}$$

$$= \sum_{i=0}^{k-1} S_i + \sum_{i=0}^{k-1} n_i(\hat{m}_i - m)(\hat{m}_i - m)' = S_W + S_B, \tag{3.1.7}$$

kus

$$S_i := \sum_{j:y_j=i}(x_j - \hat{m}_i)(x_j - \hat{m}_i)', \quad S_W := \sum_{i=0}^{k-1} S_i, \quad S_B := \sum_{i=0}^{k-1} n_i(\hat{m}_i - m)(\hat{m}_i - m)'.$$

Kui klasse on kaks, siis seosest (3.1.5) saame

$$S_B = \frac{n_1 n_0}{n}(\hat{m}_1 - \hat{m}_0)(\hat{m}_1 - \hat{m}_0)'. \tag{3.1.8}$$

3.2 Lineaarne klassifitseerija

3.2.1 Eeldused ja definitsioonid

Eeldus: Alljärgnevas vaatleme olukorda kus $k = 2$, s.t. klasse on kaks ning L on sümmeetriline. Tuletame meelde, et sellisel juhul avaldub Bayesi klassifitseerija kujul (1.2.10), s.o.

$$g^*(x) = \begin{cases} 1 & \text{kui } p(1|x) \geq 0.5, \\ 0 & \text{kui } p(1|x) < 0.5. \end{cases} \tag{3.2.1}$$

Bayesi risk R^* on aga

$$\begin{aligned} R^* = R(g^*) = \mathbf{P}(g^*(X) \neq Y) &= \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbf{P}(g(X) \neq Y) \\ &= E \min\{p(1|X), 1 - p(1|X)\} = \int \min\{p(1|x), 1 - p(1|x)\} dF(x). \end{aligned}$$

kus $p(1|x) = \mathbf{P}(Y = 1|X = x)$ ja $F(x)$ on vektori X jaotusfunktsioon.

Lineaarne klassifitseerija klassifitseerib (tunnusvektorite ruumis) $\mathbb{R}^d$ antud hüpertasandi abil: ühele poole hüpertasandit jäävad punktid klassifitseeritakse klassi 0, ülejäänud klassifitseeritakse klassi 1. Formaalselt on reegel järgmine:

$$g(x) = \begin{cases} 1 & \text{kui } w'x + w_0 = \sum_{i=1}^d w^i x^i + w_0 \geq 0, \\ 0 & \text{mujal}, \end{cases} \tag{3.2.2}$$

kus $w = (w^1, \ldots, w^d)$ ja $x = (x^1, \ldots, x^d)$ kuuluvad hulka $\mathbb{R}^d$ ning $w_o \in \mathbb{R}$. Klassifitseeriv hüpertasand on seega $H = \{x \in \mathbb{R}^d : w'x + w_0 = 0\}$.

Märkused:

- Üldisust kitsendamata võib alati võtta $\|w\| = 1$.
- Tihti on klassid kodeeritud kui +1 ja -1 (saame $\{0, 1\}$ kodeeringust transformatsiooni $2Y - 1$ kaudu) ja sellisel juhul on lineaarne klassifitseerija antud kui

$$g(x) = \text{sgn}(w'x + w_0).$$

- tehisõppe-alases kirjanduses (eriti närvivõrkudega seoses) nimetatakse lineaarset klassifitseeijat tihti **pertseptroniks** (ik *perceptron*)

3.3 Riski hinnangud kovariatsioonimaatriksi kaudu

3.3.1 Ühedimensionaalne ruum

Ühemõõtmelises ruumis on hüpertasand punkt (*split*). Seega, kui $d = 1$, on lineaarne klassifitseerija järgmine

$$g(x) = \begin{cases} y', & \text{kui } x \le x'; \\ 1 - y', & \text{kui } x > x'. \end{cases}, \tag{3.3.1}$$

kus $x' \in \mathbb{R}$ ja $y' \in \{0, 1\}$. Seega iga lineaarne klassifitseerija on antud punkti x' ja "poole"y' abil.

Olgu $\mathcal{G}$ kõikide lineaarsete klassifitseerijate hulk. Uurime, kui hästi töötab parim võimalik lineaarne klassifitseerija, st uurime riski

$$R := \inf_{\mathcal{G}} R(g).$$

Mida väiksem R, seda paremini on klassid lineaarselt eralduvad.

Ülesanne 3.2 *Veendu, et ühemõõtmelisel juhul*

$$R = \inf_{x',y'} \Big[I_{\{y'=0\}} \big(\pi_1 F_1(x') + \pi_0 (1 - F_0(x')\big) + I_{\{y'=1\}} \big(\pi_0 F_0(x') + \pi_1 (1 - F_1(x')\big) \Big] \le \pi_1 \wedge \pi_0 \le \frac{1}{2}, \tag{3.3.2}$$

kus

$$F_i(x) = \mathbf{P}(X \le x | Y = i), \quad \pi_i = \mathbf{P}(Y = i), \quad i = 0, 1.$$

Seega kehtib

$$R^* \le R \le \frac{1}{2},$$

kusjuures on lihtne näidata (vaata [1], lemma 4.1), et $R = \frac{1}{2}$ parajasti siis, kui $R^* = \frac{1}{2}$ ja sellisel juhul on igasugune klassifitseerimine mõttetu.

Ülesanne 3.3 *Tõestada, et kui* $\pi_1 = 0.5$*, siis*

$$R = \frac{1}{2} - \frac{1}{2} \sup_x |F_1(x) - F_0(x)|.$$

($\min\{a, b\} = \frac{a+b-|a-b|}{2}$*)*

Järgnev lemma annab hinnangu riskile R kasutades klasside keskmisi ja dispersioone. Viimaseid on valimi põhjal tihti võimalik hinnata. Olgu

$$m_i = E(X|Y = i), \quad \sigma_i^2 = D(X|Y = i), \quad i = 1, 0.$$

Tuletame meelde Tŝebõŝev-Cantelli võrratuse (TNT2, Trm 8.20)

$$\mathbf{P}(X - EX \ge c) \le \frac{DX}{c^2 + DX}.$$

Lemma 3.1

$$R \le \left(1 + \frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2}\right)^{-1}. \tag{3.3.3}$$

Tõestus. Üldisust kitsendamata eeldame, et $m_0 < m_1$. Olgu $\Delta_1 > 0$, $\Delta_0 > 0$ sellised, et $m_1 - m_0 = \Delta_1 + \Delta_0$. Vaatleme reeglit

$$g(x) = \begin{cases} 0, & \text{kui } x \le m_0 + \Delta_0; \\ 1, & \text{kui } x > m_1 + \Delta_1. \end{cases}$$

On selge, et

$$R \le R(g) = \mathbf{P}(Y = 1)\mathbf{P}(X \le m_1 - \Delta_1 | Y = 1) + \mathbf{P}(Y = 0)\mathbf{P}(X > m_0 + \Delta_0 | Y = 0). \tag{3.3.4}$$

Kasutades Tšebõšev-Cantelli võrratust, saame

$$\mathbf{P}(X > m_0 + \Delta_0 | Y = 0) \le \frac{\sigma_0^2}{\sigma_0^2 + \Delta_0^2}$$

ning (kuidas?)

$$\mathbf{P}(X \le m_1 - \Delta_1 | Y = 1) \le \frac{\sigma_1^2}{\sigma_1^2 + \Delta_1^2}.$$

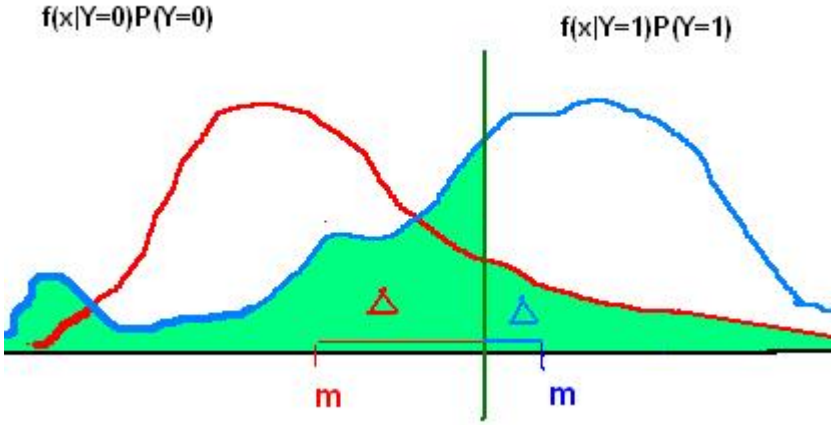
Seega

$$R \le \frac{\pi_1 \sigma_1^2}{\sigma_1^2 + \Delta_1^2} + \frac{(1 - \pi_1)\sigma_0^2}{\sigma_0^2 + \Delta_0^2},$$

kus $\pi_1 = \mathbf{P}(Y = 1)$. Võttes

$$\Delta_0 = \frac{(m_1 - m_0)\sigma_0}{\sigma_0 + \sigma_1}, \quad \Delta_1 = \frac{\sigma_1}{\sigma_0}\Delta_0,$$

saame (3.3.3). ■



Kui klasside keskmised on teineteisest kaugel võrreledes nende dispersioonidega, on suurus

$$\frac{|m_1 - m_0|}{\sigma_0 + \sigma_1}$$

suur ning R ja R^* väikesed. Sellisel juhul on mõttekas otsida (parimat) lineaarset klassifitseerijat.

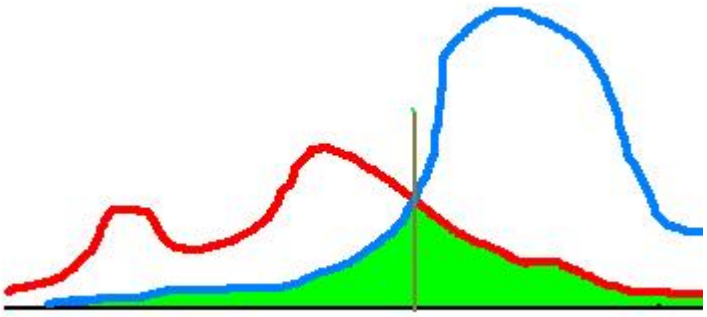
Paraku on lihtne näha, et tihti ei anna (3.3.3) mõistlikku hinnangut.

Näide. Olgu $X \sim U[0,1]$ ning olgu

$$Y = \begin{cases} 0, & \text{kui } X \in [\frac{1}{3}, \frac{2}{3}]; \\ 1, & \text{mujal.} \end{cases}$$

Ülesanne 3.4 *Veendu, et $R^* = 0$, $R = \frac{1}{3}$. Leida võrratus (3.3.3) parem pool ning parim lineaarne klassifitseerija.*

Lõpetuseks märgime, et ühedimensionaalsel juhul on tihti ka Bayesi klassifitseerija kujul (3.3.1) ehk Bayesi klassifitseerija on lineaarne. Selline olukord on näiteks siis, kui klasside tinglikud tihedused lõikuvad vaid ühes punktis. Sellisel juhul, loomulikult, $R^* = R$.



3.3.2 Ruum $\mathbb{R}^d$

Käsitleme praegu olukorda, kus $x' = (x^1, \ldots, x^d)$. Seega on iga lineaarne klassifitseerija määratud d-dimensionnalse vektoriga w ja skalaariiga w_0. Üldisust kitsendamata võime eeldada, et $\|w\| = 1$.

Olgu jällegi $\mathcal{G}$ kõikide lineaarsete klassifitseerijate hulk, ning vaatluse all olgu parima lineaarse klassifitseerija risk $R := \inf_{\mathcal{G}} R(g)$.

Analoogiliselt ühemõõtmelise juhuga saab näidata, et $R \leq \frac{1}{2}$, kusjuures $R = \frac{1}{2}$ parajasti siis, kui $R^* = \frac{1}{2}$. Tuletame meelde, et tinglikud keskmised ja kovariatsioonimaatriksid avalduvad

$$m_i = E[X|Y = i], \quad \Sigma_i = E[(X - m_i)(X - m_i)'|Y = i], \quad i = 0, 1.$$

Lemmast 3.1 saame tõkke suurusele R klasside keskmiste ja kovariatsioonimaatriksite kaudu ka d-dimensionaalsel juhul.

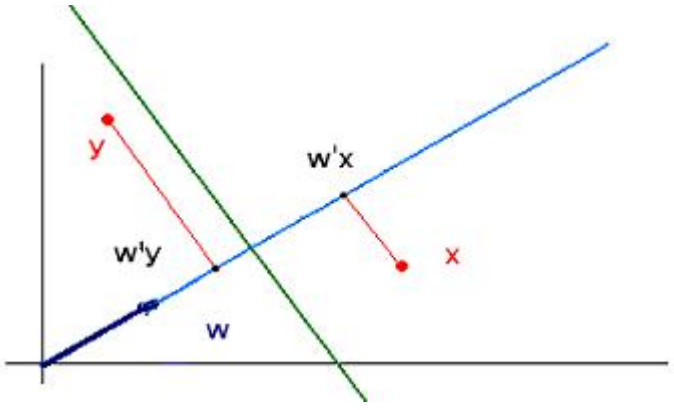
Teoreem 3.1

$$R^* \leq R \leq \inf_{w \in \mathbb{R}^d} \Big(1 + \frac{\big(w'(m_1 - m_0)\big)^2}{\big((w'\Sigma_1 w)^{\frac{1}{2}} + (w'\Sigma_0 w)^{\frac{1}{2}}\big)^2}\Big)^{-1}. \tag{3.3.5}$$

Tõestus. Olgu g lineaarne klassifitseerija (3.2.2), mis on esitatav vektoriga $w \in \mathbb{R}^d$ ja vabaliikmega $w_0 \in \mathbb{R}$. Üldisust kitsendamata (miks?) eeldame, et $\|w\| = 1$. Tunnusvektori x klassifitseerimiseks projekteeritakse x vektori w sihis olevale ühemõõtmelisele alamruumile ning saadud projektsiooni $w'x$ võrreldakse konstandiga w_0. Konstandiga võrdlemine on aga ühemõõtmeline lineaarne klassifitseerimine. Viimase riski saame aga hinnata lemma 3.1 abil. Selleks on aga vaja vektori w sihis olevale alamruumile projekteeritud tunnusvektori tinglikke keskmisi ja dispersioone. Viimased avalduvad $E[w'X|Y = i] = w'm_i$, $D[(w'X)|Y = i] = w'\Sigma_i w$. Ühemõõtmelisel ruumil on lineaarne klassifitseerija defineeritud punkti w_0 ja "poole " abil. Fikseerides ühemõõtmelise ruumi, st vektori w, saame lemmast 3.1 et parima lineaarse klassifitseerija risk selles ruumis, olgu see R_w, rahuldab võrratust

$$R_w \leq \Big(1 + \frac{\big(w'(m_1 - m_0)\big)^2}{\big((w'\Sigma_1 w)^{\frac{1}{2}} + (w'\Sigma_0 w)^{\frac{1}{2}}\big)^2}\Big)^{-1}.$$

■



Funktsioonid J ja J'. Seost (3.3.5) võib interpreteerida järgmiselt: R on väike kui leidub $w \in \mathbb{R}^d$, et

$$J(w) := \frac{w'(m_1 - m_0)}{(w'\Sigma_0 w)^{\frac{1}{2}} + (w'\Sigma_1 w)^{\frac{1}{2}}} \tag{3.3.6}$$

on suur. Kuid $w'(m_1 - m_2)$ on aga vektorit w läbivale alamruumile projekteeritud tinglike tunnusvektorite keskväärtuste vahe ruut ning $w'\Sigma_0 w$ ja $w'\Sigma_1 w$ on samale alamruumile projekteeritud tinglike tunnusvektorite kovariatsioonimaatriksid. Seega $J(w)$ mõõdab kui hästi on tinglikud jaotused eralduvad, kui nad on projekteeritud vektorit w läbivale alamruumile (on selge, et tunnused on hästi eralduvad, kui $J(w)$ on suur – keskväärtused on üksteisest kaugel ja hajuvused keskväärtuste ümber väikesed).

Märkus: Võrratus (3.3.5) ei ole täpne: võib konstrueerida hulganisti näiteid, kus $m_1 = m_2$, ning seega $J(w) = 0$ iga w korral ja võrratuse parem pool on 1. Selliseid näiteid võib konstrueerida ka nii, et $R^* = 0$.

Funktsiooni J maksimiseerimisest on lihtsam optimeerida funktsiooni

$$\begin{aligned} J'(w) &:= \frac{(w'(m_1 - m_0))^2}{\pi_0 w'\Sigma_0 w + \pi_1 w'\Sigma_1 w} = \frac{(w'(m_1 - m_0))^2}{w'(\pi_0\Sigma_0 + \pi_1\Sigma_1)w} \\ &= \frac{w'(m_1 - m_0)(m_1 - m_0)'w}{w'\Sigma_W w} = (\frac{1}{\pi_1\pi_0})\frac{w'\Sigma_B w}{w'\Sigma_W w}. \end{aligned}$$

Viimases võrduses kasutasime seost (3.1.5): $\Sigma_B = \pi_1\pi_0(m_1 - m_0)(m_1 - m_0)'$.

Pane tähele: erijuhul kui $\Sigma_0 = \Sigma_1 = \Sigma$, siis $\Sigma_W = \Sigma$ ja $4J^2(w) = \pi_1\pi_0 J'(w)$.

Funktsiooni $J'(w)$ on palju kergem maksimiseerida kui $J(w)$. Tõepoolest, funktsiooni $J'(w)$ maksimiseerimine on ekvivalentne järgmise maksimiseerimisülesandega

$$\max_{w:\|w\|=1} \frac{w'\Sigma_B w}{w'\Sigma_W w}. \tag{3.3.7}$$

ning pole raske veenduda, et lahend on proportsionaalne vektoriga

$$w \propto \Sigma_W^{-1}(m_1 - m_0). \tag{3.3.8}$$

Lahend pole ühene: iga vektor kujul $\alpha\Sigma_W^{-1}(m_1 - m_0)$, kus $\alpha \neq 0$ maksimiseerib $J'(w)$. Seetõttu on lahendi puhul oluline vaid tema siht.

Eelpool märkisime, et kui $\Sigma_0 = \Sigma_1 =: \Sigma$, siis funktsiooni J' maksimeerimine on ekvivalentne funktsiooni J maksimeerimisega ja sellisel juhul mõlema optimiseerimisülesande lahend on

$$w \propto \Sigma^{-1}(m_1 - m_0).$$

3.4 Millal Bayesi klassifitseerija on lineaarne?

Eespool nägime, et ühemõõtmelisel juhul $d = 1$ võib Bayesi klassifitseerija olla tihti lineaarne (ühepunktiline), kuid $d > 1$ korral on lineaarne klassifitseerija pigem õnnelik juhus kui seaduspära. Bayesi klassifitseerija on lineaarne näiteks siis, kui klasside tinglikud jaotused on ühe ja sama kovariatsioonimaatriksiga mitmemõõtmelised normaaljaotused. Tõepoolest, tähistades tinglikud tihedused vastavalt f_0 ja f_1, saame Bayesi reegliks

$$g^*(x) = \begin{cases} 1 & \text{kui } \pi f_1(x) > (1-\pi) f_0(x), \\ 0 & \text{mujal.} \end{cases} \tag{3.4.1}$$

Siin $\pi = \mathbf{P}(Y = 1)$. Kui

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp[-\frac{1}{2}(x - m_i)' \Sigma_i^{-1} (x - m_i)],$$

saame, et $g^*(x) = 1$ parajasti siis, kui

$$(x-m_1)'\Sigma_1^{-1}(x-m_1) - 2\ln\pi + \ln|\Sigma_1| < (x-m_0)'\Sigma_0^{-1}(x-m_0) - 2\ln(1-\pi) + \ln|\Sigma_0|. \tag{3.4.2}$$

Üldiselt on määrab sellise reegli mingi teist järku pind.

Erijuhul, kui $\Sigma_1 = \Sigma_0 = \Sigma$, siis võrratus (3.4.2) lihtsustub

$$(x - m_1)'\Sigma^{-1}(x - m_1) < (x - m_0)'\Sigma^{-1}(x - m_0) - 2\ln\frac{\pi}{1-\pi}. \tag{3.4.3}$$

Viimane on ekvivalentne

$$-2x'\Sigma^{-1}(m_1 - m_0) < 2\ln\frac{\pi}{1-\pi} + m_0'\Sigma^{-1}m_0 - m_1'\Sigma^{-1}m_1$$

ehk Bayesi klassifitseerija on lineaarne:

$$g^*(x) = \begin{cases} 1, & \text{kui } \underline{w}'x + w_0 > 0; \\ 0, & \text{mujal,} \end{cases}$$

kus

$$\underline{w} := \Sigma^{-1}(m_1 - m_0) \tag{3.4.4}$$

ning

$$\begin{aligned} w_0 &= \ln\frac{\pi}{1-\pi} + \frac{1}{2}\Big(m_0'\Sigma^{-1}m_0 - m_1'\Sigma^{-1}m_1\Big) = \ln\frac{\pi}{1-\pi} - \frac{(m_0+m_1)'}{2}\Sigma^{-1}(m_1 - m_0) \\ &= \ln\frac{\pi}{1-\pi} - \frac{(m_0+m_1)'}{2}\underline{w}. \end{aligned}$$

Teisisõnu $g^*(x) = 1$ parajasti siis, kui

$$\underline{w}'x > \frac{(m_0+m_1)'}{2}\underline{w} - \ln\frac{\pi}{1-\pi} = \frac{\underline{w}'m_0 + \underline{w}'m_1}{2} - \ln\frac{\pi}{1-\pi}$$

ehk,

$$|\underline{w}'x - \underline{w}'m_1| < |\underline{w}'x - \underline{w}'m_0| - 2\ln\frac{\pi}{1-\pi}, \tag{3.4.5}$$

sest

$$\underline{w}'(m_1 - m_0) = (m_1 - m_0)'\Sigma^{-1}(m_1 - m_0) \geq 0.$$

Märgime, et kui klasse on rohkem kui kaks ja $\Sigma_i = \Sigma$ iga i korral, on piir iga kahe klassi vahel lineaarne.

Väga spetsiifilisel erijuhul, kui $\pi = \frac{1}{2}$ (ning $\Sigma_1 = \Sigma_0 = \Sigma$) on

$$\frac{(m_0+m_1)}{2} = EX =: m$$

ning sellisel juhul on reegel seega

$$g^*(x) = 1 \Leftrightarrow \underline{w}'x > \underline{w}'m.$$

Seega, kui tunnusvektor X on mõlemas klassis normaalselt jaotatud, kusjuures jaotuste kovariatsioonimaatriks on sama, on parim võimalik klassifitseerija lineaarne. Kui on alust arvata, et andmed on saadud just sellisest mudelist, tasub otsida parimat klassifitseerijat lineaarsete klassifitseerijate seast.

3.5 Riski hinnangud ja empiirilise riski minimiseerimine (ERM)

Tuletame veelkord meelde, et $\mathcal{G}$ on kõikide lineaarsete klassifitseerijate hulk, $R = \inf_{g\in\mathcal{G}} R(g)$ on parima lineaarse klassfitseerija risk ning iga valimi põhjal leitud klassifitseerija g_n korral $R(g_n) = \mathbf{P}(g_n(X) \neq Y|D_n)$. Seda riski $R(g_n)$ me ei tea, kuid peatükis 2.6 tutvustasime nn PAC-tüüpi ülemisi hinnanguid. Need põhinesid klassifikaatorite hulga $\mathcal{G}$ VC-dimensioonil. Lineaarsete klassifikaatorite hulk $\mathcal{G}$ ei ole kompleksne, selle VC-dimensioon on $d+1$. Seega võrratus (2.6.11) annab järgmise hinnangu:

$$\mathbf{P}\Big(\sup_{g\in\mathcal{G}}|R_n(g) - R(g)| > \epsilon\Big) \leq 8(n+1)^{d+1}\exp[-\frac{n\epsilon^2}{32}], \tag{3.5.1}$$

millest saame PAC-tüüpi võrratuse suvalise lineaarse klassifitseerija riskile (võrratus (2.6.15)): tõenäosusega $1-\delta$

$$R(g_n) \leq R_n(g_n) + 2\sqrt{\frac{8((d+1)\ln n - \ln\delta + \ln 8)}{n}} = R_n(g_n) + 8\sqrt{\frac{(d+1)\ln n + \ln\frac{8}{\delta}}{2n}}. \tag{3.5.2}$$

Tuletame meelde, et (3.5.2) iga $g_n \in \mathcal{G}$ korral.

ERM-printsiibil saadud klassifitseerija. ERM-printsiibil saadud lineaarne klassifikaator on selline, mis minimiseerib klassifitseerimisvigade arvu treeningvalimis üle kõikvõimalike lineaarsete klassifikaatorite. Seega antud valimi $(x_1, y_1), \ldots, (x_n, y_n)$ korral

$$g_n = \arg\min_{g\in\mathcal{G}} R_n(g) = \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i)) = \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{n} I_{\{y_i\neq g(x_i)\}},$$

sest L on sümmeetriline. Hinnang (3.5.2) kehtib ka $\hat{g}_n$ korral, kuid võrratustest (2.6.13) ja (2.6.14) saame antud juhul veel võrratused

$$ER(\hat{g}_n) - R \le 4\sqrt{\frac{(d+1)\ln(n+1) + \ln 2}{n}}$$

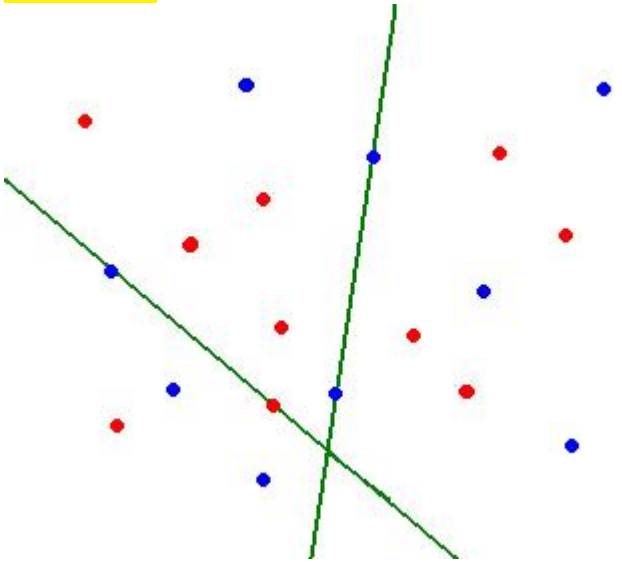
$$\mathbf{P}\Big(R(\hat{g}_n) - R > \epsilon\Big) \le 8(n+1)^{(d+1)} \exp[-\frac{n\epsilon^2}{128}],$$

millest muidugi

$$R(\hat{g}_n) \to R, \quad \text{a.s.}, \quad R_n(\hat{g}_n) \to R.$$

Ülaltoodust tuleneb, et ERM printsiibil saadud klassifikaatoril on head omadused ehk ERM printsiibi kasutamine klassifitseerimises on teoreetiliselt hästi põhjendatud.

Fingering. ERM meetodi puudus on aga see, et klassifikaatorit $\hat{g}_n$ on raske leida. Empiirilise riskifunktsiooni $R_n(g)$ gradiendid on peaaegu kõikjal võrdsed nulliga, mistõttu gradient-meetodid ei tööta. Veel enam, saab näidata, et empiirilist riski minimiseeriva lineaarse klassifikaatori leidmine on NP-raske. Üks võimalus probleemi ületamiseks on nn. **fingering**, kus vaadatakse vaid neid hüpertasandeid, mis läbivad d valimi punkti.



Iga selline hüperatasand defineerib 2 klassifikaatorit. Kui tunnuse X jaotusel on tihedus, on igal hüpertasandil maksimaalselt d valimi punkti (p.k.). Seega vaadeldakse $\binom{n}{d}$

hüpertasandit ehk $2\binom{n}{d}$ klassifikaatorit. *Fingering*-meetod korral minimiseeritakse empiirilist riski üle nende konkreetsete klassifikaatorite ja see on palju lihtsam, sest selliste klassifikaatorite hulk on lõplik. Kuigi sel teel saadud klassifitseerija g_n ei minimiseeri empiirilist riski üle kõigi võimalike lineaarsete klassifikaatorite, on empiirilise riski erinevus suhteliselt väike:

$$R_n(g_n) - \inf_{g \in \mathcal{G}} R_n(g) \leq \frac{d}{n},$$

sest erinevus võib tulla vaid hüpertasandil asuvate valimipunktide klassifitseerimisest. Seetõttu pole midagi imestada, et *fingering*-meetodil saadud klassifikaator töötab (vähemalt piisavalt suure n korral) peaaegu sama hästi kui $\hat{g}_n$: kehtivad teoreemi 3.3.2 võrratustega analoogsed võrratused. Selle postuleerib järgmine teoreem ([1], thm. 4.5), kus g_n on fingering-meetodil saadud klassifitseerija.

Teoreem 3.2 *Olgu X absoluutselt pidev (tihedusega). Olgu $d < n$ ja $\frac{2d}{n} \leq \epsilon < 1$. Siis iga (X, Y) jaotuse korral kehtivad võrratused*

$$\mathbf{P}\big(R(g_n) - R > \epsilon\big) \leq e^{2d\epsilon}\left(2\binom{n}{d} + 1\right)e^{-\frac{n\epsilon^2}{2}}$$

$$ER(g_n) - R \leq \sqrt{\frac{2}{n}(d+1)\ln n + (2d+2)}.$$

Alternatiivid. Et empiirilist riski minimiseerivat lineaarset klassifikaatorit on raske leida, siis praktikas kasutatakse mitmesuguseid alternatiive. Paneme tähele, et empiirilise riski võib defineerida kui

$$\frac{1}{n}\sum_{i=1}^{n}\big|y_i - I_{(0,\infty)}(w^T x_i + w_o)\big|^p, \tag{3.5.3}$$

kus $p \geq 1$. Üks võmalus selle funktsiooni siledaks tegemiseks (ja püüdes samas säilitada tema põhilised omadused) on indikaatori $I_{(0,\infty)}$ asendamine mingi sileda (sõlm)funktsiooniga σ. Sellisel juhul minimiseeritakse riskifunktsiooni

$$\frac{1}{n}\sum_{i=1}^{n}\big|y_i - \sigma(w^T x_i + w_o)\big|^p, \tag{3.5.4}$$

mis nüüd on diferentseeruv. Siin p on harilikult 1 või 2. Tihti kasutatakse funktsiooni σ rollis logistilist funitsiooni

$$\sigma(t) = \frac{\exp[t]}{1 + \exp[t]}$$

ja (3.5.3) on siis

$$\frac{1}{n}\sum_{i=1}^{n}\big|y_i - \frac{\exp[w^T x_i + w_o]}{1 + \exp[w^T x_i + w_o]}\big|^p.$$

Kui $p = 2$, saame sellisel juhul logistilise regressiooni vähimruutude meetodil, millest räägime edaspidi.

3.6 Klassikalised meetodid parima lineaarse klassifitseerija leidmiseks

Vaatamata sellele, et enamasti pole parim võimalik (Bayesi) klassifitseerija lineaarne, kasutatakse eelkõige nende lihtsuse tõttu lineaarseid klassifitseerijaid praktikas tihti. Alljärgnevas vaatleme põgusalt enamlevinuimaid meetode parima lineaarse klassifitseerija leidmiseks treeningandmete põhjal.

Seega käesolevas alajaotuses eeldame treeningandmete $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$ olemasolu; klass $\mathcal{G}$ on kõigi lineaarsete klassifitseerijate hulk.

3.6.1 Lineaarne regressioon

Lineaarse regressiooni põhiidee on funktsiooni $x \mapsto p(1|x)$ lähendamine (teatavas mõttes) parima lineaarse funktsiooniga $f(x) := w'x + a$, kus $w \in \mathbb{R}^d$, $a_0 \in \mathbb{R}$. Seosest (3.2.1) saame siis klassifitseerija

$$g(x) = \begin{cases} 1 & \text{kui } w'x + a \geq 0.5, \\ 0 & \text{kui } w'x + a < 0.5. \end{cases} \tag{3.6.1}$$

Võttes $w_0 := a - 0.5$, saame, et (3.6.1) on lineaarne klassifitseerija kujul (3.2.2):

$$g(x) = 1 \quad \Leftrightarrow \quad w'x + w_o \geq 0 \quad \text{mujal} \quad \text{g(x)} = 0.$$

Klassikaline meetod koefitsentide w ja a leidmiseks on **vähimruutude meetod** (ik *ordinary least squares*): leia $\hat{w}$ ja $\hat{a}$, mis minimiseerivad vähimruutude summat

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - w'x_i - a)^2. \tag{3.6.2}$$

ja nende abil konstrueeri klassifitseerija (3.6.1) g_n.

Vähimruutude meetodil saadud klassifitseerija omadused. Alljärgnevas vaatleme suvalise jaotusega juhuslikku vektorit (X, Y) ja optimiseerimisülesannet

$$\min_{w,a} E\big(Y - (w'X + a)\big)^2. \tag{3.6.3}$$

Kui vektor (X, Y) on empiirilise jaotusega F_n, on (3.6.3) sama mis summa (3.6.2).

Lause 3.2 *Olgu $f : \mathbb{R}^d \to \mathbb{R}$ suvaline funktsioon. Siis*

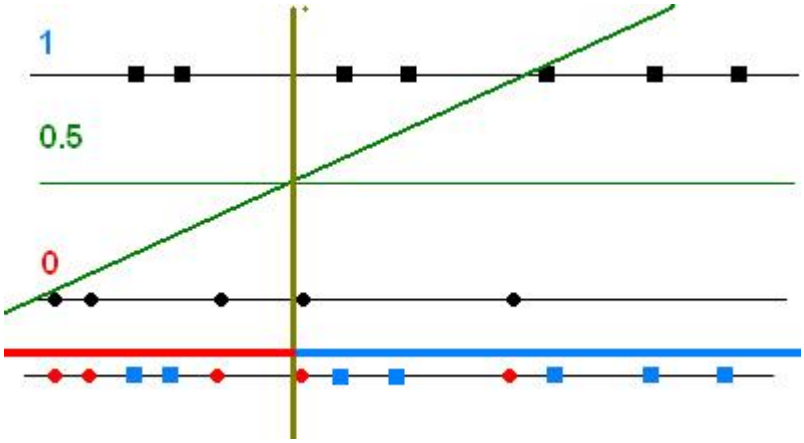
$$E\big(Y - f(X)\big)^2 = E\big(Y - p(1|X)\big)^2 + E\big(p(1|X) - f(X)\big)^2.$$

Ülesanne 3.5 *Tõesta lause.*

Ülaltoodud lausest järeldub, et iga funktsioonide klassi $\mathcal{F}$ (mitte ilmtingimata lineaarsete funktsioonide hulga) korral keskmise ruutkao $E\big(Y - f(X)\big)^2$ minimiseerimine üle $\mathcal{F}$ on ekvivalentne $E\big(p(1|X) - f(X)\big)^2$ minimiseerimisega üle klassi $\mathcal{F}$. Erijuhul kui $\mathcal{F}$ on lineaarsete funktsioonide klass, saame

$$\min_{w,a} E\big(p(1|X) - (w'X - a)\big)^2 \quad \Leftrightarrow \quad \min_{w,a} E\big(Y - (w'X - a)\big)^2. \tag{3.6.4}$$

Seega $f(x) = w'x + a$ on vähimruutude mõttes parim lineaarne funktsioon, mis lähendab funktsiooni $p(1|x)$.



Lause 3.3 *Olgu $\gamma, \beta \in \mathbb{R}$. Olgu w^* ja a^* optimiseerimisülesande (3.6.3) lahendid. Siis γw^* ja $\gamma a^* + \beta$ on järgmise optimeerimisülesande lahendid*

$$\min_{w,a} \big((\gamma Y + \beta) - (w'X + a)\big)^2 = \min_{w,a} \big(\tilde{Y} - (w'X + a)\big)^2, \tag{3.6.5}$$

kus $\tilde{Y} = \gamma Y + \beta$.

Ülesanne 3.6 *Tõesta lause.*

Lausest 3.3 järeldub muuhulgas, et klassifitseerija (3.6.1) sisuliselt ei muutu, kui klassid ümber kodeerida: $0 \leftrightarrow 1$. Olgu $f(x) = (w^*)'x + a^*$ esialgse ülesande lahend (regressioonifunktsioon). Selle abil saadud klassifitseerija (3.6.1) on järgmine

$$g(x) = 1 \quad \Leftrightarrow \quad f(x) \geq 0.5 \quad \Leftrightarrow \quad f(x) \geq 1 - f(x). \tag{3.6.6}$$

Pärast tunnuste lineaarset teisendust saame regressioonifunktsiooni

$$\tilde{f}(x) = \gamma\big((w^*)'x + a^*\big) + \beta = \gamma f(x) + \beta.$$

Vaadeldaval juhul on kodeeritud klassid $\tilde{Y} = 1 - Y$, st $\gamma = -1$ ja $\beta = 1$, millest saame, et iga x korral $\tilde{f}(x) = 1 - f(x)$ ehk $f(x) \geq 0.5$ parajasti siis, kui $\tilde{f}(x) \leq 0.5$. Seega kui klassid on kvalitatiivsed, näiteks *punased* ja *sinised*, ei sõltu lineaarsel regressioonil saadud klassifitseerija sellest, kas klassiks 1 kodeerida punased või sinised. Veel enam saadud klassifitseerija võib saada ka järgmiselt: kodeeri klassid ühtviisi, näiteks punased kui 0 ja sinised kui 1 ja leia regressioonifunktsioon $f_s(x)$ (s nagu sinised, sest need on kodeeritud kui 1). Siis kodeeri klassid vastupidi. Sellisel juhul on regressioonifunktsioon f_p, kusjuures iga x korral $f_p(x) = 1 - f_s(x)$ ning klassifitseerija g on

$$g(x) = s, \quad \Leftrightarrow \quad f_s(x) \geq 0.5 \quad \Leftrightarrow \quad f_s(x) \geq f_p(x). \tag{3.6.7}$$

Lausest on ka hästi näha, et sisuliselt ei muutu klassifitseerimine lineaarse regressiooni kaudu ka siis, kui klasse kodeerida mitte kui 0 ja 1 vaid näiteks a ja b. Tõepoolest, kodeeringust 0 ja 1 saame kodeeringu a ja b (nii, et $0 \leftrightarrow a$) lineaarse teisenduse $(b-a)y+a$ kaudu. Seega, kui $f_{0,1}$ on esialgsele kodeeringule vastav regressioonifunktsioon, siis (a, b)-kodeeringule vastav funktsioon on $f_{a,b}(x) = (b-a)f_{0,1}(x)+a$. Vahetades a ja b (st esialgsele klassile 0 vastab nüüd b) saame regressioonifunktsiooni $f_{b,a}(x) = (a-b)f_{0,1}(x)+b$. Nüüd, juhul $a < b$ saame (veendu!), et

$$f_{a,b}(x) \geq f_{b,a}(x) \quad \Leftrightarrow \quad f_{0,1}(x) \geq \frac{1}{2} \quad \Leftrightarrow \quad f_{a,b}(x) \geq \frac{b+a}{2}$$

ning juhul kui $a > b$ saame, et

$$f_{a,b}(x) \geq f_{b,a}(x) \quad \Leftrightarrow \quad f_{0,1}(x) \leq \frac{1}{2} \quad \Leftrightarrow \quad f_{a,b}(x) \leq \frac{b+a}{2}.$$

Seega suvalise klasside kodeeringu a ja b korral järgmine klassifiseerija on ekvivalentne klassifiseerijaga (3.6.6):

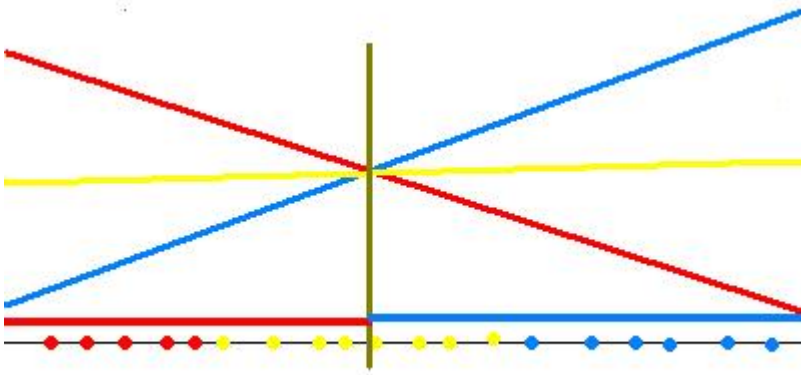
$$g(x) = \max\{a, b\} \quad \Leftrightarrow \quad f_{a,b}(x) \geq \frac{b+a}{2}$$

ehk lineaarsel regressioonifunktsiooni abil saadud klassifitseerija ei sõltu klasside kodeeringust.

Rohkem kui kaks klassi. Tuleta meelde eeskirja (3.6.7): tunnus x klassifitseeritakse klassi *punased*, kui $f_p(x) \geq f_s(x)$, kus regressioonifunktsioon $f_p(x)$ on saadud juhul kui punased on kodeeritud kui 1 ja sinised kui 0 ning f_s on saadud vastupidise kodeeringuga. Nüüd on lihtne lineaarse regressiooni abil klassifitseerimist üldistada enam kui kahele klassile. Sellisel juhul kodeeritakse iga klassi korral treeningandmed järgmiselt: vastavasse klassi kuuluvad treenigandmed kodeeritakse kui 1, ülejäänud kui 0. Nii leitakse iga klassi korral regressioonisirge $f_i(x)$, kus $i = 0, \ldots k-1$ on klassid. Saab näidata, et iga x korral need funktsioonid rahuldavad tingimust $f_0(x) + \cdots + f_{k-1}(x) = 1$ (vt (3.6.11)). Otsustusreegel on siis järgimine

$$g(x) = \arg \max_{i=0,\ldots,k-1} f_i(x). \tag{3.6.8}$$

Enam kui kahe klassi korral on selle reegli puudus see, et mõni sisuliselt hästi eristatud klass nö "kaetakse"teiste poolt ära (*masking*). Ühemõõtmelisel juhul illustreerib seda olukorda pilt.



Otimiseerimisülesande (3.6.3) lahendamine. Tuletame meelde ülesande (3.6.3):

$$\min_{w,a} E\big(Y-(w'X+a)\big)^2.$$

Eeldades $EX^2<\infty$ (loomulik eeldus, sest vastasel juhul poleks vähimruutude meetodil mõtet), saab näidata, et diferentseerimise ja integreerimise järjekorra ära vahetada. Seega

$$\frac{\partial}{\partial w_i}E\big(Y-(w'X+a)\big)^2=2E\Big((Y-w'X-a)X_i\Big),\quad i=1,\dots,d$$
$$\frac{\partial}{\partial a}E\big(Y-(w'X+a)\big)^2=-2E\big(Y-(w'X+a)\big)$$

Võrdsustates saadud osatuletised nulliga, saame võrrandisüsteemi w ja a leidmiseks:

$$\begin{aligned} E(XX')w+aEX &= E(XY)\\ w'EX+a &= EY. \end{aligned}$$

Järelikult

$$a=EY-w'EX$$

ning

$$E(XX')w+EYEX-EX(EX)'w=E(XY). \tag{3.6.9}$$

Et $E(XX')-EX(EX)'=\Sigma$, saame võrrandile (3.6.9) kuju

$$\Sigma w=E(XY)-EYEX, \tag{3.6.10}$$

millest

$$w=\Sigma^{-1}\Big(E(XY)-EYEX\Big).$$

Ülesanne 3.7 *Olgu klasse k. Defineerime $Y_i := I_{\{i\}}(Y)$. Seega $Y_i = 1$ parajasti siis, kui $Y = i$. Olgu*

$$(w_i, a_i) = \arg\min_{w,a} E\big(Y_i - (w'X + a)\big)^2.$$

Veendu, et

$$\sum_{i=0}^{k-1} w_i = 0, \quad \sum_{i=0}^{k-1} a_i = 1. \tag{3.6.11}$$

(Üldisust kitsendamata võid eeldada, et $EX = 0$).

Kaks klassi. Edaspidi vaatleme kahte klassi. Olgu $m_i := E[X|Y = i]$, $i = 0, 1$ ja

$$m := EX = \pi_1 m_1 + \pi_0 m_0.$$

Siis

$$E(XY) - EYEX = \pi_1 m_1 - \pi_1 m = \pi_0\pi_1(m_1 - m_0)$$

ning

$$w = \pi_0\pi_1\Sigma^{-1}(m_1 - m_0). \tag{3.6.12}$$

Vektori X kovariatsioonimaatriksi Σ_X lahutuse (3.1.4) ning seose (3.1.5) põhjal saame

$$\Sigma_X = \Sigma_W + \pi_0\pi_1(m_1 - m_0)(m_1 - m_0)'.$$

Seega seos (3.6.10) on

$$\big(\Sigma_W + \pi_0\pi_1(m_1 - m_0)(m_1 - m_0)'\big)w = \pi_0\pi_1(m_1 - m_0),$$

millest

$$\Sigma_W w = \pi_0\pi_1\big(1 - (m_1 - m_0)'w\big)(m_1 - m_0) = \alpha(m_1 - m_0),$$

kus

$$\alpha = \pi_0\pi_1\big(1 - (m_1 - m_0)'w\big) = \pi_0\pi_1\big(1 - \pi_0\pi_1(m_1 - m_0)'\Sigma^{-1}(m_1 - m_0)\big). \tag{3.6.13}$$

Siin viimane võrdus tuleneb seosest (3.6.12). Seega kahe klassi korral on (3.6.3) lahendid w ja a järgmised

$$w = \alpha\Sigma_W^{-1}(m_1 - m_0), \quad a = \pi_1 - w'm, \tag{3.6.14}$$

kus α on antud seosega (3.6.13), kusjuures on võimalik näidata, et $\alpha > 0$. Seega lahend w on samasihiline, mis funktsiooni J' miminiseeriv w (3.3.8).

Vektori $w = \alpha\Sigma_W^{-1}(m_1 - m_0)$ ja konstandi $a = \pi - w'm$ kaudu saame regressioonifunktsiooni abil saadud klassifitseerijale (3.6.1) kuju

$$g(x) = 1 \Leftrightarrow w'x + a \geq 0.5 \Leftrightarrow w'(x - m) + \pi_1 \geq 0.5 \Leftrightarrow w'x + (\pi_1 - w'm - 0.5) \geq 0. \tag{3.6.15}$$

Erijuht: Võrdsete kovariatsioonimaatrikistega normaaljaotus. Olgu tunnusvektori tinglikud jaotused normaaljaotusega ning $\Sigma_0 = \Sigma_1 = \Sigma$. Teame, et sellisel juhul Bayesi klassifitseerija lineaarne klassifitseerija kujul

$$g^*(x) = 1 \quad \Leftrightarrow \quad \underline{w}'x + w_0 \geq 0,$$

kus

$$\underline{w} = \Sigma^{-1}(m_1 - m_0) \quad \text{ja} \quad w_0 = \ln\frac{\pi_1}{\pi_0} - \frac{(m_1 + m_0)'}{2}\underline{w}.$$

Et w (3.6.14) on kujul $\alpha\underline{w}$ ja $\alpha > 0$, võime eeskirja (3.6.15) avaldada

$$g(x) = 1 \quad \Leftrightarrow \quad \underline{w}'x + \alpha^{-1}(\pi_1 - \alpha\underline{w}'m - 0.5) \geq 0. \tag{3.6.16}$$

Kui $\pi_1 = \pi_0$, on mõlemad reeglid samad (veendu!) ja seega g on Bayesi klassifitseerija, kuid üldiselt mitte – lineaarset klassifitseerijat defineeriv vektor w on küll sama, kuid konstant w_0 on üldiselt erinev.

Klassifitseerija g headus. Eelpool nägime, et väga spetsiifilisel erijuhul (mitmemõõtmeline normaaljaotus, $\Sigma_0 = \Sigma_1$ ja $\pi_1 = \pi_0$) on g Bayesi klassifitseerija ning seega $R(g)$ nii väike kui võimalik. Samas nägime ka, et g ei pruugi olla parim võimalik lineaarne klassifitseerija isegi kui klasside tinglikud jaotused on mitmemõõtmelise normaaljaotusega ja $\Sigma_0 = \Sigma_1$ (kuid $\pi_1 \neq \pi_0$). Siit küsimus: kui suur on üldiselt $R(g)$ ja kui palju see erineb parima lineaarse klassifitseerija riskist R. Vahe $R(g) - R$ sõltub (X, Y) jaotusest ja võib olla väga suur isegi kui $d = 1$. Nimelt kehtib:

$$\sup\big(R(g) - R\big) = 1, \tag{3.6.17}$$

kus supremum on võetud üle kõigi vektori (X, Y) jaotuste. Veendumaks selles, vaatame järgnevat lihtsat näidet.

Ülesanne 3.8 *Olgu (X, Y) jaotus järgmine:*

$$\mathbf{P}((X,Y) = (-\theta, 1)) = \mathbf{P}((X,Y) = (\theta, 0)) = \epsilon,$$
$$\mathbf{P}((X,Y) = (1,1)) = \mathbf{P}((X,Y) = (-1,0)) = \frac{1}{2} - \epsilon, \quad \theta > 0.$$

Leida R^* ja R. Tõestada, et

$$w = \frac{1 - 2\epsilon(1+\theta)}{4(\theta^2\epsilon + (\frac{1}{2} - \epsilon))}, \quad a = 0.5.$$

Veendu, et iga ϵ korral leidub piisavalt suur θ nii, et $w < 0$. Leida nii suure θ korral $R(g)$. Tõestada (3.6.17).

Klassifitseerija g_n. Võttes (X, Y) jaotuseks empiirilise jaotuse, saame lahendi ülesandele (3.6.2):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w'x_i - a)^2.$$

Lahendid on

$$\begin{aligned}
\hat{w} &= \hat{\alpha}\hat{\Sigma}_W^{-1}(\hat{m}_1 - \hat{m}_0) = (\hat{\alpha}n)S_W^{-1}(\hat{m}_1 - \hat{m}_0)\\
\hat{a} &= \hat{\pi}_1 - \hat{w}'\hat{m},\\
\hat{\alpha} &= \hat{\pi}_1\hat{\pi}_0\Big(1 - \frac{\hat{\pi}_1\hat{\pi}_0(\hat{m}_1 - \hat{m}_0)'\hat{\Sigma}_W^{-1}(\hat{m}_1 - \hat{m}_0)}{1 + \hat{\pi}_1\hat{\pi}_0(\hat{m}_1 - \hat{m}_0)'\hat{\Sigma}_W^{-1}(\hat{m}_1 - \hat{m}_0)}\Big).
\end{aligned}$$

Defineerides $x_i = (x_{i1}, \ldots, x_{id})'$ $i = 1, \ldots, n$ ja maatriksi

$$\mathbf{Z} = \begin{pmatrix} x_{11} & \cdots & x_{1d} & 1\\ x_{21} & \cdots & x_{2d} & 1\\ \cdots & \cdots & \cdots & \cdots\\ x_{n1} & \cdots & x_{nd} & 1 \end{pmatrix}, \quad y = \begin{pmatrix} y_1\\ y_2\\ \cdots\\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} w_1\\ \cdots\\ w_d\\ a \end{pmatrix}$$

saame lahenduse traditsioonilisel kujul

$$\begin{pmatrix} \hat{w}\\ \hat{a} \end{pmatrix} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'y.$$

Kui n kasvab, siis

$$\hat{m}_i \to m_i \quad \text{p.k.}, \quad \hat{\Sigma}_W^{-1} \to \Sigma_W^{-1} \quad \text{p.k.}, \quad \hat{\pi}_\text{i} \to \pi_\text{i} \quad \text{p.k.}$$
$$\hat{w} \to w, \quad \text{p.k.}, \quad \hat{a} \to a, \quad \text{p.k.}.$$

Kui $\mathbf{P}(w'X + a = 0.5) = 0$ (näiteks X on absoluutselt pideva jaotusega), siis domineeritud koondumise teoreemist järeldub, et

$$R(g_n) \to R(g), \quad \text{p.k.}. \tag{3.6.18}$$

Seega teatud jaotuste (sealhulgas sellised, mille korral X on pidev) puhul kehtib koondumine (3.6.18), kus g on (3.6.15) ja g_n on selle empiiriline versioon. Ülaltoodust teame, et juhul kui tunnusvektor X on mõlemas klassis normaalselt jaotatud, kusjuures jaotuste kovariatsioonimaatriks on sama ning kui $\pi_0 = \pi_1$, on g Bayesi klassifitseerija. Siit järeldub, et sellisel juhul on lineaarse regressiooni abil saadud klassifitseerija g_n mõjus. Samas piisab kui $\pi_1 \neq \pi_0$, et g poleks enam Bayesi klassifitseerija ja siis on suure n korral $R(g_n)$ ligikaudu võrdne riskiga $R(g)$; viimane aga võib olla oluliselt suurem kui Bayesi risk R^*. Veel enam, ülaltoodud kontranäitest järeldub, et lineaarne $R(g)$ võib olla väga suur ning koondumisest $R(g_n) \to R(g)$ p.k. järeldub siis, et ka $R(g_n)$ käitub väga halvasti. Sellest järeldib, et kui puudub täiendav informatsioon paari (X, Y) jaotuse kohta, võib klassifitseerimine lineaarse regressiooni põhjal viia väga halva tulemuseni kuitahes suure valimimahu n korral.

Kirjandus: Lineaarsest regressioonist loe [9] (sec 4.2.4 ja 4.3.4), [11] (sec 4.6), [6] (sec 5.8), [7] (sec 4.2 ja Ch 3), [10] (sec 7).

3.6.2 Logistiline regressioon

Tuletame meelde, et lineaarne regressioon püüab lähendada tinglikke tõenäosusi $p(0|x)$ ja $p(1|x)$ teatavas mõttes parimate lineaarsete funktsioonidega, $f_0(x)$ ja $f_1(x)$. Funktsioonid rahuldavad küll tingimust $f_1(x) + f_0(x) = 1$, kuid olles lineaarsed võtavad nad üldiselt ka negatiivseid väärtusi ning seetõttu on nende kasutamine tõenäosuste hinnangutena küsitav.
Logistiline regressioon, seevastu, ei modelleeri lineaarsena mitte tinglikke tõenäosusi vaid **logaritmilist tõepärasuhet:**

$$\ln \frac{p(1|x)}{p(0|x)}.$$

Oletame hetkeks, et kehtib seos

$$\ln \frac{p(1|x)}{p(0|x)} = w'x + w_o. \tag{3.6.19}$$

Et $p(0|x) = 1 - p(1|x)$ (vaatleme ikka kahte klassi), saame, et seose (3.6.19) kehtimisel

$$\frac{p(1|x)}{1 - p(1|x)} = \exp[w'x + w_o],$$

millest

$$p(1|x) = \frac{\exp[w'x + w_o]}{1 + \exp[w'x + w_o]} \tag{3.6.20}$$

$$p(0|x) = \frac{1}{1 + \exp[w'x + w_o]}. \tag{3.6.21}$$

Logistises regressioonis otsitaksegi vektorit $(w', w_o)'$ nii, et funktsioonid

$$\eta_1(x) := \frac{\exp[w'x + w_o]}{1 + \exp[w'x + w_o]} \tag{3.6.22}$$

$$\eta_0(x) := \frac{1}{1 + \exp[w'x + w_o]} \tag{3.6.23}$$

oleksid teatavas mõttes parimad hinnangud tõenäosustele $p(1|x)$ ja $p(0|x)$. Hinnangud (3.6.22) ja (3.6.23) on kahtlemata märksa realistlikumad kui linaarsed: nad on alati positiivsed ja summeeruvad igas punktis üheks. Hinnangute (3.6.22) ja (3.6.23) abil saadud klassifitseerija on

$$g(x) = \begin{cases} 1 & \text{kui } \eta_1(x) \geq \eta_0(x), \\ 0 & \text{mujal.} \end{cases} \tag{3.6.24}$$

Et aga $\eta_1(x) = \eta_0(x)$ parajasti siis, kui

$$\ln \frac{\eta_1(x)}{\eta_2(x)} = w'x + w_o = 0,$$

saame, et (3.6.24) määrab lineaarse klassifitseerija

$$g(x) = \begin{cases} 1 & \text{kui } w'x + w_o \geq 0, \\ 0 & \text{mujal.} \end{cases} \tag{3.6.25}$$

Nüüd on ka selge, et klassifitseerimise seisukohalt on täiesti ükskõik, kas me modelleerime lineaarselt logaritmilist tõepärasuhet $\frac{p(1|x)}{p(0|x)}$ või hoopis $\frac{p(0|x)}{p(1|x)}$.

Parameetrite hindamine. Kuidas andmete põhjal leida parimat vektorit w ning vabaliiget w_o? Levinuim meetod on suurima tõepära meetod. Et treeningvalim on *iid*, on fikseeritud tunnisvektorite $x_1, \ldots, x_n$ korral treeningvalimi (so klasside $y_1, \ldots, y_n$) saamise (tinglik) tõenäosus

$$L := \prod_{i:y_i=1} p(1|x_i) \prod_{i:y_i=0} p(0|x_i).$$

Asendades L avaldises (meile tundmatud) tõenäosused $p(i|x)$ funktsioonidega $\eta_i(x; w, w_o)$ (tuletame meelde, et funktsioonid $\eta_i(x)$ sõltuvad parameetritest (w, w_o)), saame **tingliku tõepärafunktsiooni:**

$$L(w, w_o) := \prod_{i:y_i=1} \eta_1(x_i; w, w_o) \prod_{i:y_i=0} \eta_0(x_i; w, w_o).$$

Suurima tõepära meetodis valitakse parameetrid w, w_o (ehk funktsioonid η_i) nii, et nad maksimiseeriksid $L(w, w_o)$. Selleks leitakse **logaritmiline tinglik tõepärafunktsioon:**

$$\begin{aligned} l(w, w_o) &:= \ln\Big(\prod_{i:y_i=1} \eta_1(x_i; w, w_o) \prod_{i:y_i=0} \eta_0(x_i; w, w_o) \Big) \\ &= \sum_{i:y_i=1} \ln \eta_1(x_i; w, w_o) + \sum_{i:y_i=0} \ln \eta_0(x_i; w, w_o) \\ &= \sum_{i:y_i=1} (w'x_i + w_o) - \sum_{i=1}^{n} \ln(1 + \exp[w'x_i + w_o]) \\ &= \sum_{i=1}^{n} \Big(y_i(w'x_i + w_o) - \ln(1 + \exp[w'x_i + w_o]) \Big). \end{aligned}$$

Selle maksimiseerimiseks leitkse osatuletiste vekor ehk gradient ja võrdsustakse see nulliga. Statistikas nimetetakse gradienti **talletiseks** (*score*) ning antud juhul avaldub see järgmiselt

$$\begin{pmatrix} \frac{\partial l(w,w_o)}{\partial w_o} \\ \frac{\partial l(w,w_o)}{\partial w_1} \\ \cdots \\ \frac{\partial l(w,w_o)}{\partial w_d} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \Big(y_i - \frac{\exp[w'x_i+w_o]}{1+\exp[w'x_i+w_o]} \Big) \\ \sum_{i=1}^n x_i^1 \Big(y_i - \frac{\exp[w'x_i+w_o]}{1+\exp[w'x_i+w_o]} \Big) \\ \cdots \\ \sum_{i=1}^n x_i^d \Big(y_i - \frac{\exp[w'x_i+w_o]}{1+\exp[w'x_i+w_o]} \Big) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} y_i - \eta_1(x_i; w, w_o) \\ x_i^1 \big(y_i - \eta_1(x_i; w, w_o) \big) \\ \cdots \\ x_i^d \big(y_i - \eta_1(x_i; w, w_o) \big) \end{pmatrix}.$$

Kokkuvõttes on meil $d+1$ mittelineaarsest võrrandist koosnev süsteem, mille lahendamiseks kasutatakse mitmesuguseid numbrilisi meetode. Raamatus [7] on esitatud algoritm võrrandite lahendamiseks Newton-Raphsoni meetodil.

Saadud hinnangute $\hat{w}$ ja $\hat{w}_o$ abil konstrueerutakse lineaarne klassifitseerija kujul (3.2.2):

$$g_n(x) = \begin{cases} 1 & \text{kui } \hat{w}'x + \hat{w}_o \geq 0, \\ 0 & \text{mujal.} \end{cases}$$

Rohkem kui kaks klassi. Lõpetuseks märgime, et enam kui kahe klassi korral modelleeritakse klasside tinglikud tõenäosused

$$\begin{aligned} p(i|x) &= \frac{\exp[w_i'x + w_{i0}]}{1 + \sum_{i=0}^{k-2} \exp[w_i'x + w_{i0}]}, \quad i = 0, \ldots, k-2 \\ p(k-1|x) &= \frac{1}{1 + \sum_{i=0}^{k-2} \exp[w_i'x + w_{i0}]}. \end{aligned} \tag{3.6.26}$$

Seega on k klassi korral $(k-1)(d+1)$ tundmatut parameetrit, mis kõik hinnatakse suurima tõepära meetodil. Hinnangute $\hat{w}_i, \hat{w}_{i0}, i = 0, \ldots, k-2$ abil saadakse tõenäosuste hinnangud

$$\begin{aligned} \hat{p}(i|x) &= \frac{\exp[\hat{w}_i'x + \hat{w}_{i0}]}{1 + \sum_{i=0}^{k-2} \exp[\hat{w}_i'x + \hat{w}_{i0}]}, \quad i = 0, \ldots, k-2 \\ \hat{p}(k-1|x) &= \frac{1}{1 + \sum_{i=0}^{k-2} \exp[\hat{w}_i'x + \hat{w}_{i0}]} \end{aligned} \tag{3.6.27}$$

ning klassifitseerija on seega

$$g_n(x) = \arg \max_{0=1,\ldots,k-1} \hat{p}(i|x).$$

3.6.3 Suurima tõepära klassifikaatorite mõjususest

Klassifitseerimine logistiline regressiooni abil on klassifitseerimine suurima tõepära meetodil: otsitav jaotus eeldatakse kuuluvat mingisse hulka (mudel) ning suurima tõepära printsiibil valitakse sobivaim jaotus. Erinevalt suurima tõepära meetodi klassikalisest rakendamisest ei modelleerita antud juhul mitte vektori (X, Y) jaotust vaid juhusliku suuruse Y tinglikku jaotust $F(y|x)$, tunnusvektori X jaotuse kohta mingeid kitsendusi ei seata. See asjaolu, iseenesestmõista, vaid suurendab meetodi kasutavust. Seega on logistiline regression meetod, kus Y tinglik jaotus ehk funktsioon $x \mapsto p(1|x)$ (vaatleme ikka vaid kaht klassi) kuulub mingisse jaotuste (funktsioonide) hulka $\mathcal{P}$. Statistikas nimetatakse hulka $\mathcal{P}$ *mudeliks*. Seega mudeli $\mathcal{P}$ elemendid on funktsioonid

$$\eta : \mathbb{R}^d \to [0, 1].$$

Suurima tõepära meetodil valitakse hulgast $\mathcal{P}$ andmetega sobivaim funktsioon $\hat{\eta}$ ning selle põhjal konstrueeritakse klassifitseerija g_n: $g_n(x) = 1$ parajasti siis kui $\hat{\eta}(x) \geq 0.5$.

Kas selline meetod on mõjus st kas $R(g_n) \to R^*$ p.k.? Sõltub

- mudeli kompleksusest;
- mudeli kehtivusest (korrektsusest): kas $p(1|x) \in \mathcal{P}$.

On selge, et mida suurem (kompleksem) mudel, seda suurem on võimalus, et ta on õige, st Bayesi klassifitseerija kuulub hulka

$$\mathcal{G} := \{g(x) = I_{\{\eta(x) \geq 0.5\}} : \eta \in \mathcal{P}\}$$

(lähendamisviga on 0). Kui samas pole mudel liiga kompleksne, siis hindamisviga läheneb nullile ja suurima tõepära meetod on mõjus.

Kehtiv mudel. Olgu mudel kehtiv, st $p(1|x) \in \mathcal{P}$. Seega hidamisviga on null ja mõjusus sõltub klassi $\mathcal{P}$ kompleksusest. Pane tähele, et oluline on klassi $\mathcal{P}$ mitte selle abil defineeritud klassifitseerijate klassi $\mathcal{G}$ kompleksus: võib ju olla, et $\mathcal{P}$ on väga kompleksne, kuid sellele vastav klassifitseerijate klass $\mathcal{G}$ mitte, kuid suurima tõepära meetod valib ju hinnangu klassist $\mathcal{P}$.
Teoreem 15.2 raamatus [1] annab piisavad tingimused klassi $\mathcal{P}$ kompleksusele (lõplik meetriline entroopia), mis garanteerivad mõjususe. Seega selliste klasside korral on suurima tõepära hinnangul leitud klasifitseerija mõjus. Saab näidata ([1], ptk 15.3), et klass

$$\mathcal{P} := \left\{ \frac{\exp[w'x + w_o]}{1 + \exp[w'x + w_o]} : w \in \mathbb{R}^d, w_o \in \mathbb{R} \right\}$$

rahuldab teoreemi 15.2 tingimusi. Seega kui (3.6.19) kehtib (logaritmiline tõepärasuhe on lineaarne), on logistilise regressiooni abil saadud klassifitseerija g_n mõjus, ehk $R(g_n) \to R^*$ p.k. Seega sellisel juhul on logistilise regressiooni kasutamine klassifitseerimises õigustatud.

Millal aga kehtib (3.6.19)?

Ülesanne 3.9 *Olgu klasside tinglikud tihedused kujul*

$$f_i(x) = c_i u(x) \exp[-\frac{1}{2}(x - m_i)'\Sigma(x - m_i)].$$

Tõestada, et (3.6.19) kehtib. Veenduda, et kui X on mõlemas klassis normaalselt jaotatud, kusjuures kovariatsioonimaatriks on sama, kehtib (3.6.19). Veendu, et sellisel juhul logistiline regressioon (suurima tõepära meetodil) annab mõjusa reegli.

Vale mudel. On selge, et kui tegelik jaotus $p(1|\cdot)$ ei kuulu mudelisse $\mathcal{P}$ (mudel on vale), siis üldiselt mõjususest rääkida ei saa, sest lähendamisviga võib olla suur. Sel juhul ootame, et kehtiks koondumine

$$R(g_n) \to R \text{ p.k.?} \tag{3.6.28}$$

Olgu $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid valim. Olgu iga η korral logaritmiline tõepärafunktsioon

$$l_n(\eta) = \sum_{i=1}^{n} \Big(\ln \eta(X_i) I_{\{1\}}(Y_i) + \ln(1 - \eta(X_i)) I_{\{0\}}(Y_i) \Big).$$

Suurte arvude seadusest järeldub, et iga $\eta \in \mathcal{P}$ korral

$$\frac{1}{n} \sum_{i=1}^{n} \ln \eta(X_i) I_{\{1\}}(Y_i) \to E\big(\ln(\eta(X)) I_{\{1\}}(Y) \big) = E[p(1|X) \ln \eta(X)] \quad \text{p.k.}.$$

Siin $p(1|x)$ on tegelik tõenäosus. Viimane võrdus: keskmista enne üle Y ja siis üle X. Samamoodi

$$\frac{1}{n} \sum_{i=1}^{n} \ln(1 - \eta(X_i)) I_{\{0\}}(Y_i) \to E[\big(1 - p(1|X)\big) \ln(1 - \eta(X))] \quad \text{p.k.}.$$

Seega fikseeritud η korral

$$\frac{1}{n} l_n(\eta) \to E\big[p(1|X) \ln \eta(X) + \big(1 - p(1|X)\big) \ln(1 - \eta(X))\big] =: l(\eta) \quad \text{p.k.}.$$

See koondumine kehtib p.k. iga η korral. Kui $\mathcal{P}$ on loenduv hulk, siis toodud koondumine kehtib iga η korral p.k.(mis seal vahet on?).

Ülesanne 3.10 *Olgu $\mathcal{P}$ lõplik hulk ning funktsioonil $l(\eta)$ olgu ühene maksimum üle $\mathcal{P}$. Tõestada, et*

$$\hat{\eta} := \arg\max_{\eta \in \mathcal{P}} l_n(\eta) \to \arg\max_{\eta \in \mathcal{P}} l(\eta) =: \eta' \quad \textit{p.k.}. \tag{3.6.29}$$

Seega lõpliku $\mathcal{P}$ ja piisavalt suure n korral (viimane sõltub muidugi valimist) on $l_n(\eta)$ maksimeerimine ekvivalentne $l(\eta)$ maksimeerimisega. Viimane on aga ekvivalentne funktsiooni

$$D(p(1|\cdot)||\eta(\cdot)|X) := E\Big(p(1|X) \ln \frac{p(1|X)}{\eta(X)} + (1 - p(1|X)) \ln \frac{1 - p(1|X)}{1 - \eta(X)}\Big)$$

minimiseerimisega üle η (veendu selles). Suurust $D(p(1|\cdot)||\eta(\cdot)|X)$ nimetatakse tinglikuks **Kullback-Leibleri kauguseks**, see on alati positiivne ning seda võib vaadelda kui tegeliku jaotuse $p(1|\cdot)$ kaugust tinglike jaotuste hulgast $\mathcal{P}$. Seega η' on teatavas (K-L) mõttes parim valik hulgast $\mathcal{P}$ ning kui (3.6.29) kehtib, siis suurima tõepära meetod valib piisavalt suure n korral selle välja. Paraku ei pruugi aga K-L-mõttes parima valiku η' põhjal konstrueeritud klassifitseerija g' olla riski R mõttes parim klassifitseerija hulgast $\mathcal{G}$ – võib juhtuda, et $R(g') > R$. Järgnev ülesanne annab lihtsa kontranäite, näide on seda üllatavam, et $|\mathcal{P}| = 2$. See tähendab, et vale mudeli korral pole suurima tõepära meetod võimaline valima parimat klassifitseerijat isegi kahe klassifitseerija hulgast!

Ülesanne 3.11 *Olgu $d = 1$ ning (X, Y) olgu järgmine nelja punkti $\{0,1\} \times \{0,1\}$ konstrueeritud jaotus:*

$$\mathbf{P}(X = 0, Y = 0) = \mathbf{P}(X = 1, Y = 0) = \frac{2}{9},\ \mathbf{P}(X = 0, Y = 1) = \frac{1}{9},\ \mathbf{P}(X = 1, Y = 1) = \frac{4}{9}.$$

- *Leida R^* ja Bayesi klassifitseerija.*
- *Koosnegu klass $\mathcal{P}$ kahest funktsioonist: $\eta_1(x)$ ja $\eta_2(x)$, kusjuures $\eta_1(x) \equiv 0.45$, $\eta_2(x) \equiv 0.95$. Leida g_1, g_2, $R(g_1)$ ja $R(g_2)$.*
- *Leida $x \mapsto p(1|x)$, selle abil leida $l(\eta)$ ja η'. Kas kehtib koondumine (6.2.6)?*
- *Leida*
$$\arg\min_{i=1,2} E\big(Y - \eta_i(X)\big)^2.$$

Tuletame meelde ülesandes 3.8 vaadeldud näidet. See näide käsitles lineaarset regressiooni, kuid põhimõtteliselt samasuguse näite saab teha ka logisitlise regressiooni kohta. Tõepoolest, olgu $d = 1$,

$$p(1|-m) = p(1|1) = p(0|m) = p(0|-1) = 1$$

ja

$$\mathbf{P}(X = -m) = \epsilon, \quad \mathbf{P}(X = -1) = \frac{1}{2} - \epsilon, \quad \mathbf{P}(X = m) = \epsilon, \quad \mathbf{P}(X = 1) = \frac{1}{2} - \epsilon,$$

kus $\epsilon < \frac{1}{4}$. Parim lineaarne (ühepunktiline) klassifitseerija on arusaadavalt järgmine

$$g_t(x) = 1 \quad \Leftrightarrow \quad x \geq t,$$

where $t \in (-1, 1]$. Seega $R = 2\epsilon$. Mudel:

$$\mathcal{P} = \left\{ \frac{\exp[wx + a]}{1 + \exp[wx + a]} : w, a \in \mathbb{R} \right\}$$

st η on kujul

$$\eta(x) = \frac{\exp[wx + a]}{1 + \exp[wx + a]}.$$

Tinglik logaritmiline tõepära on

$$\begin{aligned} l(\eta) = l(w, a) &= \ln \eta(-m)\epsilon + \ln(1 - \eta(-1))(0.5 - \epsilon) + \ln \eta(1)(0.5 - \epsilon) + \ln(1 - \eta(m))\epsilon \\ &= \big(- wm + a - \ln(1 + \exp[-wm + a])\big)\epsilon \\ &\quad - \ln(1 + \exp[-w + a])(0.5 - \epsilon) \\ &\quad + \big(w + a - \ln(1 + \exp[w + a])\big)(0.5 - \epsilon) \\ &\quad - \ln(1 + \exp[wm + a])\epsilon. \end{aligned}$$

Maksimeerime $l(w,a)$ üle w ja a, olgu a^* ja w^* selle ülesande lahendid. Pole raske veenduda, et $a^*=0$ ja kui m on piisavalt suur, siis $w^*<0$. Näiteks kui $m=1000$ ja $\epsilon=0.1$, siis $w^*\approx -0.0037$. Nüüd

$$\eta'(x)=\frac{\exp[w^*x]}{1+\exp[w^*x]}$$

ja vastav reegel on parimaga vastupidine: $g'(x)=1$ parjasti siis kui $x<0$. Selle reegli risk on $R(g')=1-2\epsilon$.

Pane veel tähele, et kui otsime parimat funktsiooni η vähimruutude meetodil, siis logistilise regressiooni korral meie kontranäide ei tööta. Tõepoolest, vähimruutude meetodil otsime funktsiooni $\eta\in\mathcal{P}$ nii, et $E(Y-\eta(X))^2$ oleks minimaalne. Meie näites

$$E(Y-\eta(X))^2=(1-\eta(-m))^2\epsilon+\Big(\eta^2(-1)+\big(1-\eta(1)\big)^2\Big)(\frac{1}{2}-\epsilon)+\eta^2(m)\epsilon.$$

On lihtne veenduda, et kui ϵ on piisavalt väike, siis ülaltoodud avaldist minimeeriv η on selline, et sõltumata m valikust $w>0$. See on kooskõlas asjaoluga, et antud juhul ruutkaofunktsiooni minimiseerimine on ligikaudu ERM printsiibi rakendamine.

Kokkuvõtteks: isegi siis kui suurima tõepära meetod valib klassist $\mathcal{P}$ K-L mõttes parima võimaliku hinnangu, ei pruugi selle põhjal moodustatud klassifitseerija olla parim võimalik. Ülalöeldust järeldub, et kui mudel (3.6.19) ei kehti ja regressioonifunktsioon on valitud suurima tõepära meetodil, võib klassifitseerimine logistilise regressiooni abil viia väga halva tulemuseni.

Kirjandus: Logistilisest refressioonist loe [9] (sec 4.4), [11] (sec 10.7, 10.8), [7] (sec 4.4.), [10] (sec 8).

3.6.4 Fisheri lineaarne diskriminantanalüüs (LDA)

Tuleta meelde, et iga $w\in\mathbb{R}^d$ korral juhusliku suuruse $w'X$ klassisisesed keskväärtused ja dispersioonid on

$$E[w'X|Y=i]=w'm_i,\quad D[(w'X)|Y=i]=w'\Sigma_i w.$$

Meenutame, et

$$D(w'X)=E(D[w'X|Y])+\mathrm{D}(E[w'X|Y])=w'\Sigma_W w+w'\Sigma_B w.$$

Fisheri lineaarne diskriminantanalüüs (teinekord ka lihtsalt lineaarne diskriminantanaüüs, edaspidi LDA) on lihtne meetod lineaarse klassifikaatori leidmiseks. Selle meetodi korral püütakse andmete põhjal leida lahendada järgmine optimeerimisülesanne:

$$\max_{w:\|w\|=1}\frac{w'\Sigma_B w}{w'\Sigma_W w}. \tag{3.6.30}$$

Saadud w põhjal konstrueeritakse lineaarne klassifikaator (3.2.2), kus w_0 on mingisugune konstant. Osast 3.3.2 teame, et (3.6.30) maksimiseerimine on ekvivalentne funktsiooni

$$J'(w) = \frac{(w'(m_1 - m_0))^2}{w'\Sigma_W w}$$

maksimiseerimisega (üle ühikvektorite) ja selle ülesande lahend on kujul

$$w = c\Sigma_W^{-1}(m_1 - m_0), \tag{3.6.31}$$

kus c on mingi konstant. Meid huvitab ainult vektori w siht, seega võime lahendiks võtta

$$w = \Sigma_W^{-1}(m_1 - m_0). \tag{3.6.32}$$

LDA teoreetilised põhjendused võiksid olla järgmised:

1. Vektor w valitakse selline, et tema sihis olevale ühemõõtmelisele alamruumile projekteerituna oleks tunnus teatavas mõttes hästi eralduv: tinglikud keskväärtused $E[w'X|Y = i] = w'm_i$ on teineteisest suhteliselt kaugel ja tinglike dispersioonide kaalutud summa $ED[w'X|Y] = w'\Sigma_W w$ suhteliselt väike.

2. Juhul kui klassidesisesed dispersioonid on võrdsed, st $\Sigma_1 = \Sigma_0$, on J' maksimeerimine ekvivalentne funktsiooni

   $$J(w) = \frac{w'(m_1 - m_0)}{(w'\Sigma_0 w)^{\frac{1}{2}} + (w'\Sigma_1 w)^{\frac{1}{2}}}$$

   maksimeerimisega. Lemma 3.1 põhjal aga on see ekvivalentne parima klassifitseerija riski R ülemise tõkke minimiseerimisega.

3. Kui klassisisesed jaotused on normaalsed ja $\Sigma_1 = \Sigma_0$, on (3.6.32) Bayesi klassifitseerijat määrav vektor.

Andmete põhjal hindamine. Et maatriks Σ_W^{-1} ning vektor $m_1 - m_0$ on meile üldiselt teadmata, asendatakse need andmete põhjal saadud hinnangutega ja nii saame vektori (tähistused vaata jaotusest 3.1.2)

$$\hat{w} = \hat{\Sigma}_W^{-1}(\hat{m}_1 - \hat{m}_0) = nS_W^{-1}(\hat{m}_1 - \hat{m}_0). \tag{3.6.33}$$

Vektor $\hat{w}$ maksimiseerib suhet

$$\hat{J}(w) := \frac{w'S_B w}{w'S_W w}$$

ning, et (seos (3.1.5))

$$S_B = \frac{n_1 n_0}{n}(\hat{m}_1 - \hat{m}_0)(\hat{m}_1 - \hat{m}_0)',$$

siis $\hat{J}(w)$ maksimiseerimine on ekvivalentne järgmise funktsiooni maksimiseerimisega:

$$\frac{w'(\hat{m}_1-\hat{m}_0)(\hat{m}_1-\hat{m}_0)'w}{w'S_Ww}=\frac{w'(\hat{m}_1-\hat{m}_0)(\hat{m}_1-\hat{m}_0)'w}{w'(S_0+S_1)w}=\frac{\big(w'(m_1-m_0)\big)^2}{s_0^2+s_1^2}, \quad (3.6.34)$$

kus

$$s_i^2:=w'S_iw=\sum_{j:y_j=i}(w'x_j-w'\,\hat{\hat{m}}_i)^2, \quad i=0,1. \quad (3.6.35)$$

Ülaltoodud teoreetiline põhjendus 1 sellisel juhul oleks järgmine: olgu $w\in\mathbb{R}^d$, $\|w\|=1$ (see nõue ei kitsenda üldisust) ja vaatleme ühemõõtmelist valimit

$$(w'x_1,y_1),\ldots,(w'x_n,y_n). \quad (3.6.36)$$

Ühemõõtmeline valim on hea – reaalteljel on segamini klassi 1 ja klassi 0 kuuluvad punktid. Muutes vektorit w, muudame nende punktide paiknemist. Oletame, et leidub selline w, millele vastavas ühemõõtmelises valimis on klassid hästi eristuvad: ühel telje osal on klassi 1 kuuluvad punktid kenasti kobaras koos, telje teises osas on klassi 0 kuuluvate punktide hulk. Sellisel juhul on enamasti suhteliselt lihtne leida (ühemõõtmelist) klassifitseerimisreegli g ning reegel tundmatu klassiga punkti $x\in\mathbb{R}^d$ klassifitseerimiseks oleks: projekteeri see punkt vektorit w läbivale alamruumile ja rakenda g, ehk $g(w'x)$. Hea alamruumi annab seega selline w, mille korral valimis (3.6.36) on erinevad klassid teineteisest "võimalikult kaugel". Sellisel juhul on teineteisest võimalikult kaugel ka klasside empiirilised keskmised $w'\hat{m}_0$ ja $w'\hat{m}_1$. Teisest küljest aga on selge, et vahe $|w'\hat{m}_0-w'\hat{m}_1|$ üksinda ei näita klasside eristatavust, kui sellega ei kaasne väike klassisisene hajuvus. Tõepoolest, otsime ju alamruumi millele projekteerituna oleksid klassid teineteisest võimalikult kaugel ($|w'\hat{m}_0-w'\hat{m}_1|$ võimalikult suur) ning samal ajal võimalikult kobaras koos. Viimase mõõtmiseks kasutame suurust $s_1^2+s_0^2$ (inglise keeles on see *scatter*), mis mõõdab valimi (3.6.36) klassidesisest hajuvust. Niisiis on loomulik, et keskmiste vahe (ruudu) $(w'\hat{m}_0-w'\hat{m}_1)^2$ maksimiseerimise asemel maksimiseerime hoopis suhet (3.6.34).

Praktiline (ja ilmselt ka peamine) põhjus LDA kasutamiseks on tema lihtsus (ei eelda keeruliste funktsioonide optimeerimist ega suuri arvutusi). LDA oli üks esimesi klassifitseerimismeetode (Fisher , 1936) ning seetõttu on ta ka üks tuntumaid.

Kuidas peale $\hat{w}$ leidmist ühemõõtmelise valimi

$$(\hat{w}'x_1,y_1),\ldots,(\hat{w}'x_n,y_n) \quad (3.6.37)$$

abil leida parim (lineaarne) klassifikaator, see sõltub ülesande püstitusest. Tuletame meelde, et kui X on mõlemas klassis normaalselt jaotatud ning kovariatsioonimaatsiksid on võrdsed, siis Bayesi reegel avaldub

$$g^*(x)=\begin{cases}1, & \text{kui } \underline{w}'x>-w_o;\\ 0, & \text{kui } \underline{w}'x\le -w_o,\end{cases}$$

kus

$$\underline{w} = \Sigma^{-1}(m_1 - m_0), \quad w_o = \ln\frac{\pi_1}{\pi_0} - \frac{(m_0 + m_1)'}{2}\underline{w}.$$

Samuti nägime, et selle reegli võib esitada kujul: $g^*(x) = 1$ parajasti siis, kui

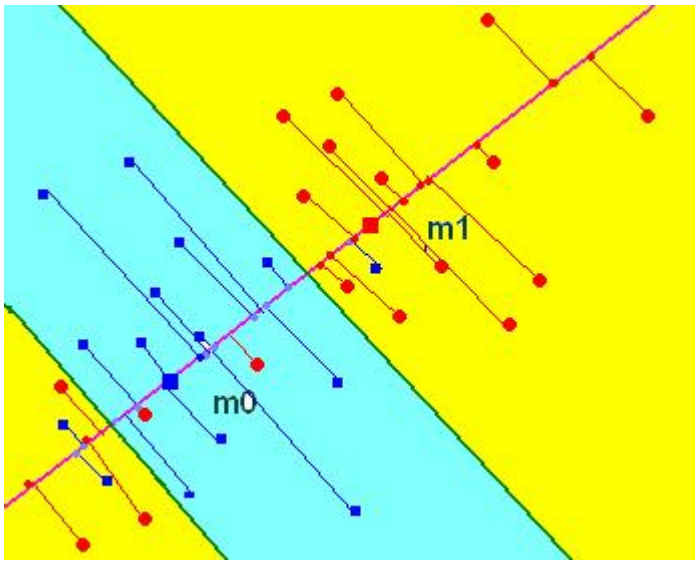
$$|\underline{w}'x - \underline{w}'m_1| < |\underline{w}'x - \underline{w}'m_0| - 2\ln\frac{\pi_1}{\pi_0}.$$

Järelikult, kui on alust eeldada, X on mõlemas klassis normaalselt jaotatud ning kovariatsioonimaatsiksid on võrdsed, siis arvestades, et vektor $\hat{w}$ on vektori $\underline{w}$ hinnang, saame üheks võimalikuks klassifitseerimisreegliks: $g_n(x) = 1$ parajasti siis, kui

$$|\hat{w}'x - \hat{w}'\hat{m}_1| < |\hat{w}'x - \hat{w}'\hat{m}_0| - 2\ln\frac{\hat{\pi}_1}{\hat{\pi}_0}, \tag{3.6.38}$$

Erijuhul $\hat{\pi}_1 = 0.5$, on (3.6.38) lihtne: punkt x määra klassi 1 parajsti siis, kui tema projektsioon $\hat{w}'x$ paikneb keskmisele $\hat{w}'\hat{m}_1$ lähemal kui keskmisele $\hat{w}'\hat{m}_0$.

Üldiselt võib aga (LDA mõttes) parima ühemõõtmelise valimi (3.6.37) leidmist vaadelda kui klassifitseerimisülesande vaheetappi: ühedimensionaalsete andmete klassifitseerimine on üldiselt lihtsam kui d-dimensionaalsete andmete klassifitseerimine ning seetõttu võib valimi (3.6.37) klassifitseerimisel kasutada meetode, mis d-dimensionaalses ruumis on mingil põhjusel (näiteks suure arvutusmahu tõttu) mitterakenduvad. Näiteks raamatu [7] autorid soovitavad valimi (3.6.37) klassifitseerimisel kasutada empiirilise riski minimiseerimist, üldiselt väga töömahukat kuid teoreetiliselt hästi põhjendatud protseduuri, millest räägime järgmises peatükis. Muidugi tuleb silmas pidada, et dimensiooni redutseerimine läbi LDA (või mistahes teise lineaarse teisenduse) viib väga limiteeritud klassifikaatoriteni – kui ühemõõtmelisel valimil leitud klassifikaator on lineaarne (ühepunktiline), on seda loomulikult ka lõplik (so ruumis $\mathbb{R}^d$ defineeritud) klassifikator. Kuid kui ühemõõtmelisel valimil leitud klassifikaator on mittelineaarne, on lõplik klassifikaator siiski vaid tükati lineaarne ja piltlikult esitatav hüpertasandite vahel olevate "ribadena".



Paneme tähele, et LDA pakutud vektor $\hat{w} = \hat{\Sigma}_W^{-1}(\hat{m}_1 - \hat{m}_0)$ on sama(suunaline), mis lineaarse regressiooni abil saadud vektor. Seega kahe klassi korral projekteerib lineaarse regressiooni abil saadud klassifitseerija andmed samale alamruumile, mis LDA (vektor $\hat{w}$ on mõlemal juhul sama(suunaline)). Saadud lineaarsed klassifitseerijad on seega lõikumatud (kollineaarsed), kas nad on võrdsed või mitte, sõltub vabaliikme w_o valikust. Peatükis 3.6.1 nägime, et juhul kui kui $n_0 = n_1$, on lineaarse regressiooni abil saadud klassifitseerija kujul

$$g_n(x) = 1 \quad \Leftrightarrow \quad \hat{w}'x \geq \hat{w}'\hat{m} \quad \Leftrightarrow \quad |\hat{w}'x - \hat{w}'\hat{m}_1| < |\hat{w}'x - \hat{w}'\hat{m}_0| \tag{3.6.39}$$

ning eespool nägime (3.6.38), et sellisel juhul (st $n_0 = n_1$) on (3.6.39) teatud mõttes loomulik tulem ka LDA korral.

Kokkuvõttes: kahe klassi korral on lineaarse regressiooni ja LDA abil saadud klassifikaatorite omadused sarnased, sarnased on ka nende (väär)kasutamisel tekkinud ohud, eelkõige tuleb jällegi meeles pidada, et kui puudub informatsioon paari (X, Y) jaotuse kohta, võib LDA viia väga halva tulemuseni.

Rohkem kui kaks klassi. Lõpetuseks lühidalt LDA kasutamisest enam kui kahe klassi korral. Enam kui kahe klassi korral otsib LDA enamasti sellist $k-1$-dimensionaalset alamruumi, millele projekteeridas oleksid k klassi võimalikult hästi eristuvad. (Muidugi võib ka enam kui kahe klassi korral projekteerida andmed ühemõõtmelisele alamruumile; samamoodi võib ka kahe klassi korral projekteerida andmeid näiteks 2-dimensionaalsele

alamruumile). Seega otsime $k-1$ vektorit $w_1, \ldots, w_{k-1}$ nii, et $k-1$ dimensionaalne valim

$$(w_1'x_1, \ldots, w_{k-1}'x_1, y_1), (w_1'x_2, \ldots, w_{k-1}'x_2, y_2), \ldots, (w_1'x_n, \ldots, w_{k-1}'x_n, y_n) \tag{3.6.40}$$

oleks teatavas mõttes hästi klassifitseeritav. Olgu $\mathbf{W}$ $d \times (k-1)$-dimensionaalne maatriks, mille veerud on $w_1, \ldots, w_{k-1}$. Valim (3.6.40) on maatrikstähistuses

$$(\mathbf{W}'x_1, y_1), (\mathbf{W}'x_2, y_2), \ldots, (\mathbf{W}'x_n, y_n). \tag{3.6.41}$$

Selle valimi keskmine on $\mathbf{W}'\hat{m}$, klasside keskmised on $\mathbf{W}'\hat{m}_i$ ning *scatter matrix* on $\mathbf{W}'S\mathbf{W}$, kus $S = \sum_{j=1}^{n}(x_j - \hat{m})(x_j - \hat{m})'$.

Ülesanne 3.12 *Tõestada, et valimi (3.6.41) klassidesisene hajuvus on* $\mathbf{W}'S_W\mathbf{W}$ *ning klassidevaheline hajuvus on* $\mathbf{W}'S_B\mathbf{W}$. *Seega maatriksi* $\mathbf{W}'S\mathbf{W}$ *lahutus (3.1.6) on*

$$\mathbf{W}'S\mathbf{W} = \mathbf{W}'S_W\mathbf{W} + \mathbf{W}'S_B\mathbf{W}, \tag{3.6.42}$$

Kahe klassi korral on nii $\mathbf{W}'S_B\mathbf{W}$ kui ka $\mathbf{W}'S_W\mathbf{W}$ reaalarvud, nende suhe aga maksimiseeritav funktsioon $\hat{J}$. Juhul kui $k > 2$, on mõlemad maatriksid. Optimiseeritav funktsioon on nüüd maatriksite $\mathbf{W}'S_B\mathbf{W}$ ja $\mathbf{W}'S_W\mathbf{W}$ determinantide suhe, st

$$\hat{J}(\mathbf{W}) = \frac{|\mathbf{W}'S_B\mathbf{W}|}{|\mathbf{W}'S_W\mathbf{W}|}.$$

Determinantide kasutamise sisuline põhjendus on järgmine: sümmeetrilise maatriksi determinant on omaväärtuste korrutis ning seega (mingis mõttes) ruutvormi hajuvuskarakteristikute korrutis. Praktilisest küljest on determinantide kasutamine mugav sellepärast, et funktsiooni $\hat{J}(\mathbf{W})$ maksimiseeriva maatriksi $\mathbf{W}$ veerud w_i on üldisatud omaväärtusülesande

$$S_Bw_i = \lambda_iS_Ww_i, \quad i = 1, \ldots, k-1 \tag{3.6.43}$$

lahendid. Et S_B astak on ülimalt $k-1$, on ülimalt $k-1$ omaväärtust nullist erinevad ning neile vastavad omavektorid (*canonical variates*) tekitavadki otsitava (ülimalt $k-1$ dimensionaalse) alamruumi, millele projekteeritud valimile rakendatakse edasist analüüsi. Märgime, et isegi kui nullist erinevaid omaväärtusi on enam kui 1 (või enam kui 2), võib soovi korral valimi ikkagi projekteerida vaid suurimale (või kahele suuremale) omaväärtusele vastavale ühedimensionaalsele (kahedimensionaalsele) alamruumile.
Paneme tähele, et omavektoriteni (3.6.43) jõuame ka järgmiselt arutledes: vektor w_1 määrab parima (LDA-mõttes) ühemõõtmelise alamruumi, st ta on maksimiseerimisülesande

$$\max_{w:\|w\|=1} \hat{J}(w)$$

lahend. Vektor w_2 maksimiseerib funktsiooni $\hat{J}(w)$ üle nende ühikvektorite, mis rahuldavad tingimust: $w'S_Ww_1 = 0$ ehk on S_W-ortogonaalsed vektoriga w_1; vektor w_3 maksimiseerib $\hat{J}(w)$ üle nende ühikvektorite, mis on S_W-ortogonaalsed nii vektoriga w_1 kui w_2 jne.

Klassifitseerimine pärast andmete sobivale alamruumile projekteerimist sõltub jällegi andmetest ja ülesande püstitusest, LDA siin täpseid ettekirjutusi ei tee. Kui on alust arvata, et tunnuse X jaotus klassides on normaalne, kovariatsioonimaatriksid on võrdsed ja ka tõenäosused $\mathbf{P}(Y = i)$ on võrdsed, on loomulik klassifitseerida x klassi i parajasti siis kui $\mathbf{W}$ tekitatud alamruumis on talle lähim klassi i keskmine ehk

$$g_n(x) = \arg \min_{i=1,\ldots,K} \|\mathbf{W}'x - \mathbf{W}'\hat{m}_i\|.$$

Üldiselt on LDA aga eelkõige meetod andmete dimensiooni vähendamiseks. Samas peab olema ettevaatlik, sest dimensiooni redutseerimise läbi võib kaduma minna nii mõndagi olulist.

Kirjandus: LDA rakendamisest mitme klassi korral võid täpsemalt lugeda raamatutest [7, 9] (mõlemates ptk. 4.3), [6] (ptk. 4.11), [10] (8.6.3).

Peatükk 4

Tugivektormasinad

4.1 Meeldetuletus: Lagrange'i määramate kordajate meetod

Sadulpunktist. Olgu $L(x, \alpha)$ suvaline kahemuutuja funktsioon. Defineerime

$$\tilde{f}(x) := \sup_{\alpha} L(x, \alpha), \quad \theta(\alpha) := \inf_{x} L(x, \alpha).$$

Olgu

$$f^* := \inf_{x} \tilde{f}(x), \quad \theta^* := \sup_{\alpha} \theta(\alpha).$$

Ülesanne 4.1 *1. Tõesta, et*

$$f^* = \inf_{x} \tilde{f}(x) = \inf_{x} \sup_{\alpha} L(x, \alpha) \geq \sup_{\alpha} \inf_{x} L(x, \alpha) = \sup_{\alpha} \theta(\alpha) = \theta^*.$$

2. Olgu

$$\tilde{f}(x^*) = \min_{x} \tilde{f}(x), \quad \theta(\alpha^*) = \max_{\alpha} \theta(\alpha). \tag{4.1.1}$$

Tõesta, et kui $\tilde{f}(x^*) = \theta(\alpha^*)$ – **tugev duaalsus** *(ik* strong duality*)– siis* (x^*, α^*) *on* **sadulpunkt***:*

$$L(x, \alpha^*) \geq L(x^*, \alpha^*) \geq L(x^*, \alpha), \quad \forall x, \alpha$$

ja

$$\tilde{f}(x^*) = L(x^*, \alpha^*) = \theta(\alpha^*). \tag{4.1.2}$$

3. Tõesta, et kui (x^*, α^*) *on sadulpunkt, siis kehtivad (4.1.2) ja (4.1.1).*

Esialgne ja duaalne ülesanne. Vaatleme optimeerimisülesannet:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} f(x) \\ &\text{nii et } g(x) \le 0, \end{aligned} \tag{4.1.3}$$

kus

$$f : \mathbb{R}^d \to \mathbb{R}, \quad g : \mathbb{R}^d \to \mathbb{R}^m.$$

Seega $g(x) \le 0$ tähendab, et $g_i(x) \le 0$ iga $i = 1, \dots, m$ korral.

Selleks defineerime **Lagrange'i funktsionaali**:

$$L(x, \alpha) = f(x) + \alpha' g(x),$$

kus $\alpha \in [0, \infty)^m$ ning paneme tähele, et (4.1.3) on ekvivalentne probleemiga (lihtne veenduda)

$$\min_{x\in\mathbb{R}^d} \max_{\alpha\ge 0} L(x, \alpha) = \min_{x\in\mathbb{R}^d} \tilde{f}(x), \tag{4.1.4}$$

kus

$$\tilde{f}(x) := \max_{\alpha\ge 0} L(x, \alpha).$$

Ekvivalentsus tähendab seda, et (4.1.3) lahend on ka (4.1.4) lahend ning miinimumid on võrdsed.

Vahetame minimiseerimise ning maksimiserimise ning vaatleme **duaalset ülesannet**:

$$\max_{\alpha\ge 0} \min_{x\in\mathbb{R}^d} L(x, \alpha) = \max_{\alpha\ge 0} \theta(\alpha), \tag{4.1.5}$$

kus

$$\theta(\alpha) := \min_{x\in\mathbb{R}^d} L(x, \alpha).$$

Olgu x^* ülesande (4.1.3) ning α^* ülesande (4.1.5) lahend, st

$$\tilde{f}(x^*) = \min_{x\in\mathbb{R}^d} \tilde{f}(x) = f^*, \quad \theta(\alpha^*) = \max_{\alpha\ge 0} \theta(\alpha) = \theta^*.$$

Seega

$$f^* \ge \theta^*.$$

Kui $f^* = \theta^*$ (tugev duaalsus), siis (x^*, α^*) on sadulpunkt ja kehtib (4.1.2):

$$\tilde{f}(x^*) = f(x^*) + \alpha^{*\prime} g(x^*) = L(x^*, \alpha^*) = \theta(\alpha^*) = \min_x L(x, \alpha^*). \tag{4.1.6}$$

Viimasest võrdusest (või ka sellest, et (x^*, α^*) on sadulpunkt) järeldub, et

$$x^* = \arg\min_x L(x, \alpha^*). \tag{4.1.7}$$

Samuti kehtib vastupidine: kui leidub L sadulpunkt (x^*, α^*), siis kehtib $f^* = \theta^*$ ning x^* ja α^* on vastavalt esialgse ja duaalse ülesande lahendid.

Seosest (4.2.7) järeldub, tugeva duaalsuse korral võib esialgse ülesande lahendi x^* leida duaalse ülesande kaudu järgmiselt:

- iga $\alpha \geq 0$ korral leia x_α nii, et $L(x_\alpha, \alpha) = \min_x L(x, \alpha) = \theta(\alpha)$. Nii saad duaalse ülesande (4.1.5);
- leia duaalse ülesande (4.1.5) lahend α^*;
- arvule α^* vastav x_{α^*} on ülesande (4.1.3) lahend x^*.

Ülesanne 4.2 *Veendu, et duaalne ülesanne on alati nõgus (ka siis, kui f ja g pole kumerad ega nõgusad).*

Seega duaalse ülesande lahendamine on enamasti kergem, tal on ka vähem kitsendusi.

KKT tingimused. Iga esialgse probleemi lahend x^* minimiseerib ka funktsiooni $\tilde{f}$ kusjuures $f(x^*) = \tilde{f}(x^*)$. Võrdustest (4.1.6) järeldub, et tugeva duaalsuse korral

$$f(x^*) = \tilde{f}(x^*) = f(x^*) + \alpha^{*\prime} g(x^*) = L(x^*, \alpha^*)$$

ehk $\alpha^{*\prime} g(x^*) = 0$. Et aga $g(x^*) \leq 0$ ja $\alpha^* \geq 0$, siis $\alpha^{*\prime} g(x^*) = 0$ parajasti siis, kui

$$\alpha_i^* g_i(x^*) = 0 \quad i = 1, \ldots, m.$$

Seega tugeva duaalsuse korral korral rahuldab optimaalne paar (α^*, x^*) nn **KKT (Karush-Kuhn-Tucker)** tingimusi:

$$\begin{aligned} \alpha_i^* &\geq 0, \quad i = 1, \ldots, m \\ g_i(x^*) &\leq 0, \quad i = 1, \ldots, m \\ \alpha_i^* g_i(x^*) &= 0, \quad i = 1, \ldots, m. \end{aligned}$$

Võrdustega antud tingimused. Tihti on osa tingimustest antud võrdustega:

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} f(x) \\ \text{nii et } &g(x) \leq 0, \\ &e(x) = 0, \qquad \text{kus} \end{aligned} \tag{4.1.8}$$

$$f : \mathbb{R}^d \to \mathbb{R}, \quad g : \mathbb{R}^d \to \mathbb{R}^m, \quad e : \mathbb{R}^d \to \mathbb{R}^l$$

ja $e(x) = 0$ tähendab, et $e_i(x) = 0$ iga $i = 1, \ldots, l$ korral.

Formaalselt võib toodud ülesande esitada kujul (4.1.3) järgmiselt

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} f(x) \\ \text{nii et } &g(x) \leq 0, \\ &e(x) \leq 0 \\ &-e(x) \leq 0. \end{aligned} \tag{4.1.9}$$

Seega ülesande (4.1.9) Lagrange'i funktsionaal on

$$L(x,\alpha,\beta^+,\beta^-) = f(x) + \alpha' g(x) - \beta'_+ e(x) + \beta'_- e(x) = f(x) + \alpha' g(x) + (\beta_- - \beta_+)' e(x).$$

Defineerides $\beta := \beta_- - \beta_+$ (ei ole ilmtingimata negatiivne), saame

$$L(x,\alpha,\beta) = f(x) + \alpha' g(x) + \beta' e(x),$$

kus $\alpha \in [0,\infty)^m$ ning $\beta \in \mathbb{R}^l$. Duaalne ülesanne on seega

$$\max_{\alpha\geq 0,\beta} \theta(\alpha,\beta), \tag{4.1.10}$$

kus $\theta(\alpha,\beta) := \min_{x\in\mathbb{R}^d} L(x,\alpha,\beta)$. Pane tähele, et ülesannet (4.1.10) maksimiseeritakse üle positiivsete α-de ja *kõikide* β-de.

Millal kehtib tugev duaalsus? Duaalsele ülesandele üleminek ning x^* leidmine ülaltoodud eeskirja põhjal on võimalik siis, kui kehtib $f^* = \theta^*$. Viimane pole garanteeritud, võib juhutuda, et $f^* > \theta^*$. Optimiseerimisteooria annab tingimused võrduse $f^* = \theta^*$ kehtimiseks. Saab näidata, et $f^* = \theta^*$ kehtib, kui on täidetud nn. **Slateri tingimused:**

1. Funktsioonid f ja g_i on kumerad ja e_i affiinsed;
2. leidub $x_0 \in \mathbb{R}^d$ nii, et $g_i(x_o) < 0$ iga $i = 1,\ldots,m$ korral ja $e_i(x_o) = 0$ iga $i = 1,\ldots,l$ korral.

Ruutplaneerimine (ik *quadratic programming*)**.**

$$\begin{aligned} &\min_x \frac{1}{2} x' K x + c' x \\ \text{nii et } &Ax + d \leq 0 \\ &Bx + e = 0 \end{aligned} \tag{4.1.11}$$

Kui K on positiivselt poolmääratud (sümmeetriline) maatriks, siis see on kumer optimeerimisprobleem. Enamasti on sellisel juhul täidetud Slateri tingimused. Kuid saab näidata (Dorne, 1960, vt [12], Thm 5.20), et kui see ka nii ei ole, siis lahendi olemasolul kehtib alati $f^* = \theta^*$ ning duaalsele ülesandele üleminek on seega võimalik.

Näide. Olgu optimeerimisülesanne järgmine:

$$\begin{aligned} &\min_x \frac{1}{2}x'Kx + c'x \\ &\text{nii et} \quad Ax + d \leq 0, \end{aligned} \tag{4.1.12}$$

kus K on positiivselt määratud, st leidub pöördmaatriks K^{-1}. Lagrange'i funktsioon:

$$L(x, \alpha) = \frac{1}{2}x'Kx + c'x + \alpha'(Ax + d).$$

Gradiendid

$$\nabla_x L(x, \alpha) = 0 \quad \Leftrightarrow \quad Kx + A'\alpha + c = 0 \quad \Rightarrow \quad x_\alpha = -K^{-1}(A'\alpha + c).$$

Pannes sinna x_α, saame

$$\begin{aligned} &\frac{1}{2}\big(A'\alpha + c)'K^{-1}KK^{-1}(A'\alpha + c) - c'K^{-1}(A'\alpha + c) + \alpha'\big(- AK^{-1}(A'\alpha + c) + d\big) = \\ &\frac{1}{2}\alpha'AK^{-1}A'\alpha + \frac{1}{2}c'K^{-1}c + \alpha'AK^{-1}c - c'K^{-1}A'\alpha - c'K^{-1}c - \alpha'AK^{-1}A'\alpha - \alpha'AK^{-1}c + \alpha'd = \\ &- \frac{1}{2}\alpha'AK^{-1}A'\alpha - \frac{1}{2}c'K^{-1}c - c'AK^{-1}A'\alpha + d'\alpha. \end{aligned}$$

Seega duaalne probleem on

$$\begin{aligned} &\max_\alpha -\frac{1}{2}\alpha'AK^{-1}A'\alpha + [d' - c'K^{-1}A']\alpha \\ &\text{nii et} \quad \alpha \geq 0, \end{aligned} \tag{4.1.13}$$

Duaalset ülesannet võib lahendada gradient-meetoodil või arvuti abil (nt *Matlab*). Olgu α^* duaalse ülesande lahend. siis esialgse ülesande lahend on

$$x^* = -K^{-1}(A'\alpha^* + c).$$

Ülesanne 4.3 *Vaatle ülesannet (4.1.11) kus K on positiivselt määratud nii, et K^{-1} leidub. Näita, et duaalne ülesanne on*

$$\begin{aligned} &\max_\alpha -\frac{1}{2}(\alpha'A + \beta'B)K^{-1}(A'\alpha + B'\beta) + [d' - c'K^{-1}A']\alpha + [e' - c'K^{-1}B']\beta - \frac{1}{2}c'K^{-1}c \\ &\textit{nii, et} \quad \beta, \quad \alpha \geq 0 \end{aligned} \tag{4.1.14}$$

ning kui (α^*, β^*) on duaalse probleemi lahend, siis esialgse probleemi lahend on

$$x^* = -K^{-1}(A'\alpha^* + B'\beta^* + c).$$

4.2 Marginaalmeetodid

4.2.1 Lineaarselt eralduv valim (hard margin)

Eeldus: Olgu alljärgnevas klasside märgistus -1 ja +1.

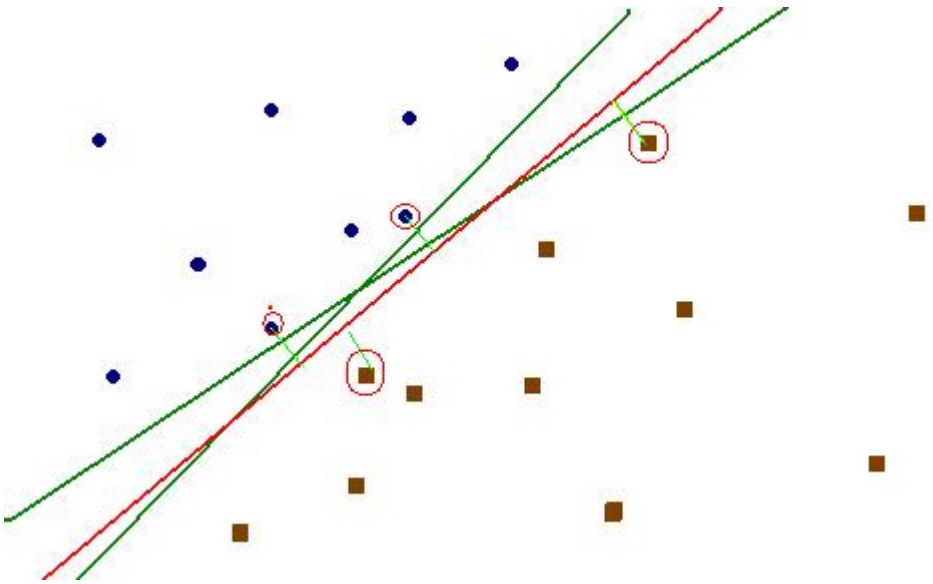
Olgu valim selline, et klassid on *lineaarselt eralduvad*. Seega leidub vähemalt üks selline lineaarne klassifitseerija kujul $g(x) = \text{sgn}(w'x + w_0)$, mis klassifitseerib kõik treeningvalimi punktid korrektselt. Teisisõnu, leidub vähemalt üks vektor w ning konstant w_o nii, et iga treeningvalimi paari (x_i, y_i) korral $w'x_i + w_o \geq 0$, kui $y_i = 1$ ja $w'x_i + w_o < 0$, kui $y_i = -1$. Ekvivalentselt,

$$y_i(w'x_i + w_o) \geq 0, \quad i = 1, \ldots, n. \tag{4.2.1}$$

Paneme tähele, et kui $\|w\| = 1$ (kuid w võime alati sellise võtta), on (4.2.1) vasak pool punkti x_i kaugus klassifitseerivast hüpertasandist. Üldjuhul on aga selliseid klassifikaatoreid mitu. Siit küsimus millist klasse eraldavat hüpertasandit (klassifikaatorit) valida? Margimaalmeetodite idee on lihtne: valime sellise lineaarse hüpertasandi, mis asub kahe klassi vahel võimalikult "keskel". Teisisõnu, valime sellise klasse eraldava hüpertasandi, mille kaugus lähimast punktist oleks maksimaalne. Kui hüpertasand eristab klassid, siis klasse eraldava hüpertasandi kaugust lähimast punktist nimetatakse **valimi (geomeetriliseks) marginaaliks** (*geometrical margin of the training set*). Kui $\|w\| = 1$, siis valimi geomeetriline marginaal on

$$\min_{i=1,\ldots,n} y_i(w'x_i + w_o)$$

ning me otsime sellist tasandit (sellist w ja w_o), mille korral see oleks suurim.



Siit saame optimeerimisülesande:

$$\begin{gathered}\max_{\gamma,w,w_o} \gamma \\ \text{nii et } y_i(w'x_i + w_o) \geq \gamma, \quad i = 1, \ldots, n \\ \|w\| = 1.\end{gathered} \tag{4.2.2}$$

Siin γ on marginaal. Pole raske näha, et ülesanne (4.2.2) on ekvivalentne järgmise optimiseerimisülesandega

$$\begin{gathered}\max_{w,w_o} \frac{1}{\|w\|} \\ \text{nii et } y_i(w'x_i + w_o) \geq 1, \quad i = 1, \ldots, n.\end{gathered} \tag{4.2.3}$$

Ülesanne 4.4 *Olgu w, w_o ülesande (4.2.2) lahend ning olgu w^*, w_o^* ülesande (4.2.3) lahend. Näita, et nad defineerivad sama hüpertasandi, st $w'x + w_o = 0$ parajasti siis kui $w^{*\prime}x + w_o^* = 0$. Näita, et selle hüpertasandi marginaal on $\|w^*\|^{-1}$.*

Märkus. Teine võimalus ülesandeni (4.2.3) jõudmiseks on järgmine: nõuame, et w ja w_o oleksid sellised, et $y_i(w'x_i + w_o) \geq 1$ iga x korral, kusjuures $\min_i y_i(w'x_i + w_o) = 1$. See tähendab, et punktid x_i, mille korral $y_i(w'x_i + w_o) = 1$, asuvad hüpertasandile kõige lähemal. Nende kaugus hüpertasandist on võrdne ja ning see võrdub marginaaliga. Olgu x_1 selline, et $w'x_1 + w_o = 1$ ja $w'x_2 + w_o = -1$. Siis $w'(x_2 - x_1) = 2$ ehk

$$\frac{1}{\|w\|} w'(x_2 - x_1) = \frac{2}{\|w\|}.$$

Suurus

$$\frac{1}{\|w\|} w'(x_2 - x_1)$$

on aga $(x_2 - x_1)$ projektsioon vektorit w läbivale sirgele. Selle projektsiooni pikkus on kaks marginaali, millest $\frac{1}{\|w\|}$ ongi marginaal.

Ülesanne (4.2.3) on aga ekvivalentne järgmise optimeerimisülesandega

$$\begin{gathered}\min_{w,w_o} \frac{1}{2}\|w\|^2 \\ \text{nii et } y_i(w'x_i + w_o) \geq 1, \quad i = 1, \ldots, n.\end{gathered} \tag{4.2.4}$$

Ülesanne 4.5 *Veendu, et ülseanne (4.2.4) on kujul (4.1.11), kuid maatriks K pole positiivselt määratud vaid on positiivselt poolmääratud.*

Ülesande (4.2.4) lahendamine. Ülesande (4.2.4) korral on Slateri tingimus täidetud, seega saab seda lahendada Lagrange'i määramata kordajate meetodil. Siin otsitavad on $w \in \mathbb{R}^d$ ja $w_o \in \mathbb{R}$. Tingimus

$$y_i(w'x_i + w_o) \geq 1, \quad i = 1, \ldots, n$$

on

$$g_i(w, w_o) = 1 - y_i(w'x_i + w_o) \leq 0, \quad i = 1, \ldots, n$$

Lagrange'i funktsionaal on järgmine:

$$L(w, w_o, \alpha) = \frac{\|w\|^2}{2} + \sum_{i=1}^{n} \alpha_i \big(1 - y_i(w'x_i + w_o)\big) = \frac{\|w\|^2}{2} + \sum_{i=1}^{n} \alpha_i (1 - y_i w'x_i) - w_o \sum_{i=1}^{n} \alpha_i y_i.$$

Leiame $(w, w_o)_\alpha$, nii, et

$$L\big((w, w_o)_\alpha, \alpha\big) = \min_{w, w_o} L(w, w_o, \alpha).$$

Pane tähele, et funktsioon $w \mapsto L(w, w_o, \alpha)$ on kumer iga α ja w_o korral. Pannes gradiendi (osatuletised) võrduma nulliga, saame

$$\frac{\partial}{\partial w_j} L(w, w_o, \alpha) = w^j - \sum_{i=1}^{n} \alpha_i y_i x_i^j = 0, \quad j = 1, \ldots d,$$

millest

$$w_\alpha^j = \sum_{i=1}^{n} \alpha_i y_i x_i^j, \quad i = 1, \ldots, d \quad \Leftrightarrow \quad w_\alpha = \sum_{i=1}^{n} \alpha_i y_i x_i. \tag{4.2.5}$$

Veendume, et kehtib ka järgmine võrrand

$$\frac{\partial}{\partial w_o} L(w, w_o, \alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0.$$

On selge, et kui α on selline, et $\sum_{i=1}^{n} \alpha_i y_i \neq 0$, siis $\min_{w, w_o} L(w, w_o, \alpha) = -\infty$ ning selline α ei saa maksimiseerida duaalset funktsiooni $\theta(\alpha)$. Seega vaatleme vaid vektoreid α, mille korral $\sum_i \alpha_i y_i = 0$. Samuti tasub meelde tuletada, et otsime sadulpunkti (α^*, w^*, w_o^*) ja see tähendab, et osatuletised w ja w_o järgi punktis α^* peavad võrduma nulliga.
Asetades w_α Lagrange'i funktsionaali ning arvestades, et $\sum_i \alpha_i y_i = 0$, saame

$$\theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j.$$

Duaalne ülesanne on seega

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j \tag{4.2.6}$$

$$\text{nii, et } \alpha_i \geq 0, \quad \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Olgu $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)'$ duaalse ülesande (4.2.6) lahend. Seosest (4.2.5) saame, et (4.2.4) lahendvektor w^* on kujul

$$w^* = w_{\alpha^*} = \sum_{i=1}^{n} \alpha_i^* y_i x_i. \tag{4.2.7}$$

KKT tingimused antud juhul:

$$\begin{aligned} \alpha_i^* &\geq 0, \quad i = 1, \ldots, n \\ y_i(w^{*\prime} x_i + w_0^*) &\geq 1, \quad i = 1, \ldots, n \\ \alpha_i^* \big(1 - y_i(w^{*\prime} x_i + w_0^*)\big) &= 0, \quad i = 1, \ldots, n \end{aligned}$$

Viimasest tingimusest järeldub, et $\alpha_i^* > 0$ vaid siis, kui $y_i(w^{*\prime} x_i + w_0^*) = 1$ ehk punkti x_i kaugus optimaalsest hüpertasandist on minimaalne (võrdne marginaaliga). Valimi punkte, millele vastav α_i^* on nullist erinev nimetatakse **tugivektoriteks** (*support vectors*). Paneme tähele, et w^* sõltub vaid tugivektoritest ehk x_i, mis pole tugivektor, pole kaasatud summasse (4.2.7). Teisisõnu,

$$w^* = \sum_{i \in \mathrm{SV}} \alpha_i^* y_i x_i, \tag{4.2.8}$$

kus SV on tugivektorite indeksite hulk. Elimineerides valimist kõik teised punktid peale tugivektorite, saaksime sama lahendi w^*. Selleks, et tugivektoreid määrata, vajame aga teisi valimi punkte ka.

Eeskiri konstandi w_0^* määramiseks on nüüd lihtne: võta x_i nii, et $\alpha_i^* > 0$. Selline x_i peab olema tugivektor, millest $y_i(w^{*\prime} x_i + w_0^*) = 1$ ehk

$$w_0^* = y_i - w^{*\prime} x_i.$$

Seega oleme taandanud optimeerimisülesande (4.2.4) duaalse ülesande (4.2.6) lahendamisele. Viimane on kumer optimeerimisülesanne, mille võib lahendada näiteks gradientmeetodite või *Matlab* abil.

Olles lahendanud duaalse ülesande, saame vektori w^* (4.2.8) ning skalaari w_0^*, mis määravad optimaalse hüpertasandi. Klassifitseerimiseeskiri on seega

$$g(x) = \operatorname{sgn}\Big(w^{*\prime} x + w_o^*\Big) = \operatorname{sgn}\Big(\sum_{i \in \mathrm{sv}} y_i \alpha_i^* x_i' x + w_o^*\Big). \tag{4.2.9}$$

Tugivektorid ja marginaal. Tugeva duaalsuse tõttu $\theta(\alpha^*) = f(x^*)$. Seega

$$\sum_i \alpha_i^* - \frac{1}{2} \sum_{i,j}^{n} \alpha_i^* \alpha_j^* y_i y_j x_i' x_j = \frac{1}{2} \|w^*\|^2 = \frac{1}{2} \sum_{i,j}^{n} \alpha_i^* \alpha_j^* y_i y_j x_i' x_j.$$

Järelikult

$$\sum_i \alpha_i^* = \sum_{i,j}^{n} \alpha_i^* \alpha_j^* y_i y_j x_i' x_j = \|w^*\|^2$$

ja marginaali saab avaldada α^* kaudu järgmiselt

$$\gamma = \frac{1}{\|w^*\|} = \Big(\sum_{i \in \mathrm{sv}} \alpha_i^*\Big)^{-\frac{1}{2}}. \tag{4.2.10}$$

Mida väiksemad on α_i^*'s,seda suurem marginaal.

Tugivektorid pole alati üheselt defineeritud. Tuletame meelde, et definitsiooni kohaselt x_i on tugivektor, kui vastav kordaja $\alpha_i^* > 0$. Sellise vektori kaugus eraldavast hüpertasandist on alati minimaalne (võrdne marginaaliga), kuid mitte kõik sellised vektorid ei ole ilmtingimata tugivektorid. Veel enam, kui duaalse ülesande lahend pole ühene, pole üheselt defineeritud ka tugivektor – kas x_i on tugivektor või mitte sõltub lahendist. Järgmine näide veenab meid selles.

Ülesanne 4.6 *Olgu* $d = 2$,

$$x_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, x_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, x_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, x_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$y_1 = -1$, $y_2 = y_3 = y_4 = 1$.
Veendu, et tegemist on lineaarselt eristuva valimiga, kusjuures marginaal on $\frac{1}{2}$.
Veendu, et maatriks $(x_i' x_j)_{ij}$ *on*

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

ning

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j x_i' x_j = \sum_i \alpha_i - \frac{1}{2}(\alpha_1^2 + \alpha_2^2 + \alpha_4^2) + \alpha_2 \alpha_4 = \theta(\alpha).$$

Olgu $\alpha^* = (\alpha_i^*)$ *duaalse ülesande (4.2.6) lahend,* w^* *olgu vastav vektor. Veendu, et lahend* α^* *on selline, et* $\|w^*\| = 2$, $\sum_i \alpha_i^* = 4$ *ning* $\theta(\alpha^*) = 2$. *Veendu, et* α^* *pole ühene ning järgmised vektorid on kõik optimeerimisülesande lahendid:*

$$(2, 0, 2, 0), \quad (2, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}), \quad (2, 1, 0, 1).$$

Tugivektorite geomeetriline interpretatsioon. Järgnev ülesanne annab tugivektoritele geomeetrilise interpretatsiooni.

Ülesanne 4.7 *Olgu valim* $x_1, \ldots, x_n$ *lineaarselt eralduv. Olgu* K_+ *ja* K_- *osavalimite kumerad katted:*

$$K_+ := \{ \sum_{i:y_i=+1} c_i x_i \,:\, c_i \geq 0, \sum_i c_i = 1 \}, \quad K_- := \{ \sum_{i:y_i=-1} c_i x_i \,:\, c_i \geq 0, \sum_i c_i = 1 \}.$$

Näidata, et optimaalne hüpertasand on risti hulki K_+ ja K_- ühendava lühima lõiguga. Selleks vaatame optimeerimisprobleemi

$$\min_{c_1,\ldots,c_n} \| \sum_{i:y_i=+1} c_i x_i - \sum_{i:y_i=-1} c_i x_i \| \tag{4.2.11}$$

$$\sum_{i:y_i=+1} c_i = 1, \quad \sum_{i:y_i=-1} c_i = 1, \quad c_i \geq 0. \tag{4.2.12}$$

Olgu c_i^* ülesande (4.2.11) lahend,

$$v^* := c^+ - c^-, \quad c^+ := \sum_{i:y_i=+1} c_i^* x_i, \quad c^- := \sum_{i:y_i=-1} c_i^* x_i.$$

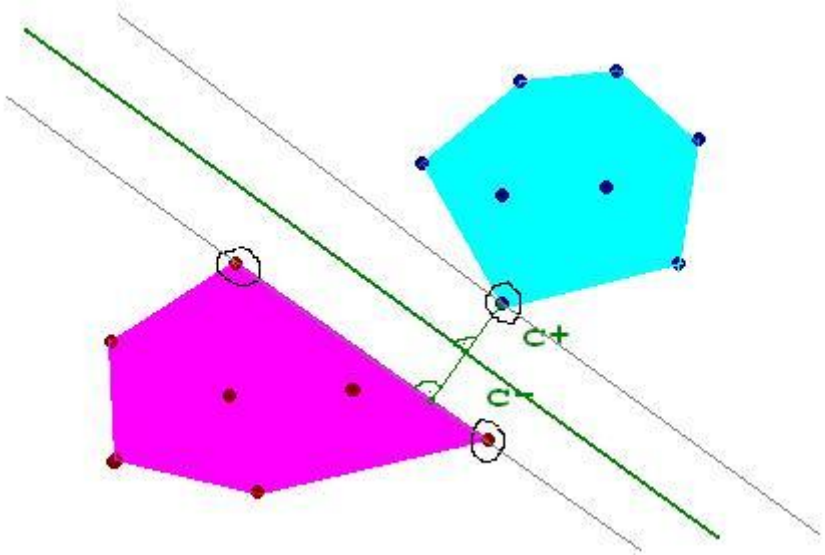
Tõesta, et ülesande (4.2.11) lahendid on kujul

$$c_i^* = \frac{2\alpha_i^*}{\sum_i \alpha_i^*},$$

kus α_i^* on duaalse ülesande (4.2.6) lahendid. Järeldada sellest, et

$$v^* = \frac{2w^*}{\|w^*\|^2}, \quad \textit{kus} \quad w^* = \sum_i \alpha_i^* y_i x_i.$$

Seega v^* ja w^* on samasihilised ning w^* on hulki K_+ ja K_- ühendava lühima lõigu sihis. Veenduda, et marginaal γ avaldub $2\gamma = \|v^*\|$. Veenduda, et konstant w_o on selline, et optimaalne hüpertasand poolitab hulki K_+ ja K_- ühendava lühima lõigu. (Selleks tõesta, et $w^{*'}(c^+ + c^-) + 2w_o = 0$ ja näita, et sellest piisab).



4.2.2 Optimiseerimisülesande (4.2.2) otse lahendamine

Tegelikult on võimalik otse (ilma ülesandele (4.2.3) üleminemata) lahendada ka optimeerimisülesanne (4.2.2):

$$\begin{gathered}\min_{w,w_o,\gamma} -\gamma \\ \text{nii et } y_i(w'x_i + w_o) \geq \gamma, \quad i = 1,\ldots,n \\ \|w\| = 1\end{gathered}$$

See ülesanne on kujul (4.1.8), sest osa tingimusi on antud võrdustega. Veendu, et üldisust kitsendamata võib aga sama ülesande esitada kujul (4.1.3)

$$\begin{gathered}\min_{w,w_o,\gamma} -\gamma \\ \text{nii et } y_i(w'x_i + w_o) \geq \gamma, \quad i = 1,\ldots,n \\ \|w\| \leq 1\end{gathered}$$

Nüüd veendu, et Slateri tingimused on täidetud ja et selle ülesande Lagrange'i funktsionaal on

$$\begin{aligned}L(w,w_o,\gamma,\beta,\lambda) &= -\gamma - \sum_i \beta_i\big(y_i(w'x_i + w_o) - \gamma\big) + \lambda\big(\|w\|^2 - 1\big) \\ &= \gamma(\sum_i \beta_i - 1) - \sum_i \beta_i\big(y_i(w'x_i + w_o)\big) + \lambda\big(\|w\|^2 - 1\big).\end{aligned}$$

Ülesanne 4.8 *Leia ülesande (4.2.2) duaalne ülesanne. Selleks näita, et Lagrange'i funktsionaali $L(w,w_o,\gamma,\beta,\lambda)$ minimiseeriv w rahuldab võrdust*

$$\sum_i \beta_i y_i x_i = 2\lambda w.$$

Samuti veendu, et sadulpunkti korral peab vektor β rahuldama tingimusi

$$\begin{gathered}\sum_i \beta_i = 1 \\ \sum_i \beta_i y_i = 0.\end{gathered}$$

Seejärel veendu, et duaalne ülesanne on

$$\begin{gathered}\max_{\beta,\lambda} \quad -\Big(\frac{1}{4\lambda}\sum_{i,j}\beta_i\beta_j y_i y_j x_i' x_j + \lambda\Big) \\ \textit{nii et} \quad \sum_i \beta_i = 1, \quad \sum_i \beta_i y_i = 0, \quad \beta_i \geq 0, \quad \lambda \geq 0\end{gathered}$$

ja näita, et duaalse ülesande lahendid on λ^* ja β^*, kus β^* on järgmise ülesande lahend:

$$min \quad \sum_{i,j} \beta_i \beta_j y_i y_j x_i' x_j$$

$$nii\ et \quad \sum_i \beta_i = 1, \quad \sum_i \beta_i y_i = 0, \quad \beta_i \geq 0$$

ja

$$\lambda^* = \frac{1}{2}\sqrt{\sum_{i,j} \beta_i^* \beta_j^* y_i y_j x_i' x_j}.$$

Veendu, et

$$w^* = \frac{v}{\|v\|} = \frac{v}{2\lambda^*}, \tag{4.2.13}$$

kus

$$v := \sum_i \beta_i^* y_i x_i.$$

KKT tingimusi kasutades leia marginaal

$$\gamma^* = \|v\| = 2\lambda^*.$$

Tugivektoreid kasutades leia, et võrrand optimaalse konstandi w_o määramiseks on

$$w_o = y_i \gamma^* - w^{*'} x_i,$$

kus x_i on tugivektor.
Lõpuks kontrolli, et ülesande (4.2.2) lahend määrab sama hüpertasandi, mis (4.2.3) lahend. Selleks pane tähele, et $\beta_i^* = \frac{c_i^*}{2}$ kus c_i^* on (4.2.11) lahend. Sellest järelda, et

$$\beta_i^* = \frac{\alpha_i^*}{\sum_i \alpha_i^*} \qquad ning \qquad w^* = \frac{w_\alpha^*}{\|w_\alpha^*\|},$$

kus $w_\alpha^* = \sum_i \alpha_i^* y_i x_i$ on (4.2.7). Samuti veendu, et

$$w_o = \frac{w_o^\alpha}{\|w_\alpha^*\|},$$

kus w_o^α on ülesande (4.2.3) optimaalne konstant.

4.2.3 Lineaarselt mitteeralduv valim (soft margin)

Juhul, kui valim on lineaarselt mitteeralduv, pole ülesanne (4.2.2) lahenduv. Sellisel juhul üldistatakse geomeetrilise marginaali mõistet järgmiselt. Vaatleme hüpertasandit

$$H = \{x : w'x + w_o = 0\},$$

kus $\|w\| = 1$ ja olgu $\gamma > 0$ mittenegatiivne konstant, mida endiselt nimetame marginaaliks. Iga valimi punkti (x_i, y_i) korral **marginaalviga** (*margin error*) on

$$\big(\gamma - y_i(w'x_i + w_o)\big)_+ \quad \text{kus} \quad (x)_+ = \begin{cases} x, & \text{if } x > 0; \\ 0, & \text{if } x \le 0. \end{cases}$$

Seega punkti (x_i, y_i) marginaalviga on null, kui x_i kaugus hüpertasandist H on vähemalt γ ja H klassifitseerib punkti (x_i, y_i) õigesti. Vastasel juhul marginaalviga on positiivne isegi kui (x_i, y_i) on õigesti klassifitseeritud. Kui punkt (x_i, y_i) on valel pool hüpertasandit, siis marginaalviga on suurem kui γ. Kui valim pole lineaarselt eralduv, siis pole ühtegi hüpertasandit (ühtegi paari w, w_o) ja ühtegi marginaali γ nii, et kõik marginaalvead oleksid võrdsed nulliga. Marginaalvigade korral pole γ enam treeningvalimi geomeetriline marginaal, seda nimetatakse **pehmeks marginaaliks** (*soft margin*).

Lineaarselt mitteralduva valimi korral on mitu võimalust parima klassifitseerija defineerimiseks. Enamasti on nad kujul

$$\min_{w:\|w\|=1, w_o, \gamma \ge 0} \Big[h(\gamma) + D \sum_{i=1}^{n} u\big((\gamma - y_i(w'x_i + w_o))_+\big) \Big], \tag{4.2.14}$$

Kus h in kahanev funktsioon ja u in kasvav, $u(0) = 0$. Seega eesmärk on samaaegselt minimiseerida marginaalvigade summat ja maksimiseerida marginaali. Need probleemid on omavahel vastukäivad (ühe kasvades teine väheneb) ning nende vahekorda reguleerib nn *regulariseerimiskonstant* $D > 0$. Kui D on suur, siis eelkõige minimiseeritakse marginaalvigu, kui D on väike, siis eelkõige kasvatatakse (pehmet) marginaali. Fikseeritud u ja h korral sõltub lahend suuresti D valikust.

Optimiseerimisprobleemidest kujul (4.2.14)

Paneme tähele, et ülesande (4.2.14) võib esitada kujul

$$\min_{\gamma \ge 0} \big(h(\gamma) + DT(\gamma) \big), \tag{4.2.15}$$

kus

$$T(\gamma) := \min_{w, w_o : \|w\|=1} \sum_{i=1}^{n} u\big(\big(\gamma - y_i(w'x_i + w_o)\big)_+ \big). \tag{4.2.16}$$

Teisisõnu, probleemi võib lahendada kahes osas: iga marginaali γ korral leida parim hüpertasand ja siis leida ülesande (4.2.15) lahendi γ^*. Kui valim on lineaarselt eralduv, siis

leidub geomeetriline marginaal γ_o nii, et $T(\gamma) = 0$ iga $\gamma \leq \gamma_o$. Kui D on piisavalt suur (kui suur peaks ta vähemalt olema?), siis (4.2.15) lahend on γ_o, st lineaarselt eralduva valimi korral eelmises peatükis vaadeldud optimiseerimisülesanne (4.2.2) on ülesande (4.2.15) erijuht.

Vaatleme ülessannet kujul (4.2.15). Olgu h kahanev diferentseeruv funktsioon; olgu T mittekahanev pidev funktsioon, millele on igas punktis mõlemapoolsed tuletised. Olgu $\gamma^* > 0$ ülesande (4.2.15) lahend. Oletades hetkeks, et T on punktis γ^* diferentseeruv (h on igas punktis diferentseeruv), saame, et $h'(\gamma^*) + DT'(\gamma^*) = 0$ ehk $-h'(\gamma^*) = DT'(\gamma^*)$. Kui T pole selles punktis diferentseeruv, siis peab olema täidetud järgmine tingimus:

$$-\frac{h'(\gamma^*)}{D} \in [T'(\gamma^*-), T'(\gamma^*+)] =: \partial T(\gamma^*). \tag{4.2.17}$$

Kui h ja T on kumerad, siis on (4.2.17) ka piisav tingimus miinimumiks. Siit lause.

Lause 4.1 *Olgu valim selline, et T on kumer. Olgu h ja g diferentseeruvad rangelt kahanevad kumerad funktsioonid. Olgu $D > 0$ ja olgu $\gamma^* \geq 0$ ülesande (4.2.15) lahend. Siis leidub $C > 0$ nii, et γ^* on järgmise ülesande lahend:*

$$\min_{\gamma \geq 0} \Big(g(\gamma) + CT(\gamma) \Big). \tag{4.2.18}$$

Tõestus. Olgu $\gamma^* > 0$. Et g ja h on rangelt kahanevad ja diferentseeruvad, siis $-g'(\gamma^*) > 0$ ning $-h'(\gamma^*) > 0$. Et $\gamma^* > 0$ on (4.2.15) lahend, siis kehtib (4.2.17). Võttes

$$C := \frac{Dg'(\gamma^*)}{h'(\gamma^*)},$$

saame

$$\frac{-g'(\gamma^*)}{C} \in \partial T(\gamma^*) = [T'(\gamma^*-), T'(\gamma^*+)].$$

Kui g ja T on mõlemad kumerad, siis on seda ka $g + CT$, millest järeldub, et γ^* on $g + CT$ miinimum.
Olgu nüüd $\gamma^* = 0$. Et $h + DT$ on kumer, siis see funktsioon peab olema kasvav, st $h'(0+) + DT'(0+) \geq 0$. Nüüd võta

$$C = \frac{g'(0+)}{h'(0+)} D,$$

veendu, et $C > 0$ ja $g'(0+) + CT'(0+) \geq 0$. ■

Funktsioonid h mis huvi pakuvad on enamasti kumerad ja rangelt kahanevad, näiteks $h(\gamma) = -\gamma$, $h(\gamma) = \gamma^{-p}$, (kus $p \geq 1$), $h(\gamma) = -\gamma^p$ (kus $0 < p \leq 1$), $h(\gamma) = \exp[-\gamma]$. Ülaltoodud lausest järeldub, et kui T on kumer ja kasvav, siis kõik need funktsioonid annavad ühe ja sama lahenduse. Vaatleme nüüd funktsiooni (4.2.16). Kuigi selline T on kumerate funktsioonide miinimum, ei järeldu sellest T kumerus; veel enam, teatud valimite korral

ei pruugi T kumer olla. Selgub, et sellisel juhul ei kehti ka lause väide.

Kontranäide: Olgu valim järgmine (d=1): punktis -19 on märgiga +1 tunnus, punktis 0 on 100 märgiga -1 tunnust, punktis 4 on 4 märgiga +1 tunnust. Et punktis 0 on sada märgiga -1 tunnust on iga marginaali γ korral minimaalse marginaavigadega klassifikaatorit kerge leida: punkti 0 marginaalviga peab olema 0. Seega antud marginaali γ korral võrdleme kaht klassifikaatorit: $\mathrm{sgn}(x-\gamma)$ ning $\mathrm{sgn}(-x-\gamma)$. Neist esimene klassifitseerib punkti x kui +1, parajasti siis $x \geq \gamma$. Seega punkt -19 klassifitseeritakse alati valesti, marginaalviga on $-19+2\gamma$. Punkt 4 jääb õigele poole marginaali (st klassifitseeritakse korrektselt ja piisava marginaaliga), kui $2\gamma \leq 4$. Vastasel juhul on margnaalviga $(4-2\gamma)$ ning et selles punktis on neli valimi elementi, tuleb see arv neljaga läbi korrutada. Seega vaadeldava klassifitseerija korral on marginaalvigade summa, olgu see $T^+(\gamma)$, järgmiselt

$$T^+(\gamma) = \begin{cases} 2\gamma+19, & \text{kui } 0 \leq 2\gamma \leq 4; \\ 2\gamma+19+4(2\gamma-4) = 10\gamma+3, & \text{kui } 2\gamma > 4. \end{cases}$$

Teine klassifikaator klassifitseerib punkti x kui -1 kui $x \geq -\gamma$. Seega kui $2\gamma \leq 19$ siis punkt -19 klassifitseeritakse piisava marginaaliga. Samas marginaalviga punktis 4 on $4+2\gamma$. See arv tuleb korrutada neljaga. Seega vaadeldava klassifitseerija korral on marginaalvigade summa juhul kui $2\gamma \leq 19$

$$T^-(\gamma) = 4(2\gamma+4) = 8\gamma+16.$$

Funktsioon T on seega

$$T(\gamma) = \min\{T^+(\gamma), T^-(\gamma)\} = \begin{cases} 8\gamma+16, & \text{kui } 0 \leq 2\gamma \leq 1; \\ 2\gamma+19, & \text{kui } 1 \leq 2\gamma \leq 4; \\ 10\gamma+3, & \text{kui } 13 \geq 2\gamma > 4; \\ 8\gamma+16, & \text{kui } 19 \geq 2\gamma > 13; \\ 10\gamma-3, & \text{kui } 19 < 2\gamma > 13; \end{cases}$$

See funktsioon pole kumer. Olgu $h(\gamma) = -\gamma$. Seega vaatleme funktsiooni

$$f_D(\gamma) := DT(\gamma) - \gamma.$$

See on tükati lineaarne funktsioon, mistõttu miinimum (kui ta on ühene) asub alati ühes järgmistest punktidest $\{0, 0.5, 2, 6.5, 9.5\}$. Veendume, et 0.5 pole ühegi D korral f_D miinimum. Kui 0.5 oleks miinimum, peaksid kehtima võrratused $f_D(0.5) \leq f_D(0)$ ja $f_D(0.5) \leq f_D(2)$. Esimene võrratus annab tingimuse $DT(0.5) - 0.5 = 20D - 0.5 \leq DT(0) - 0 = 16D$ ehk $D \leq \frac{1}{8}$. Teisest võrratusest saame, et $DT(0.5) - 0.5 = 20D - 0.5 \leq DT(2) - 2 = 23D - 2$ ehk $D \geq 0.5$. Seega need tingimused on vastuolus ning funktsiooni f_D miinimumkohad on kas 0 või 2. Teisalt aga on kerge leida mittekahanevat kumerat funktsiooni g nii, et $T(\gamma) + g(\gamma)$ miinimumkoht pole ei 0 ega 2. Näiteks olgu

$$-g(\gamma) = \begin{cases} 10\gamma, & \text{kui } \gamma \leq 0.5; \\ \frac{2}{3}\gamma + (5 - \frac{1}{3}), & \text{kui } 0.5 \leq \gamma. \end{cases}$$

Seega g on tükati lineaarne, $g(0) = 0, g(0.5) = -5, g(2) = -6$ ja g on kumer. Seega

$$T(0.5) + g(0.5) = 20 - 5 < 16 - 0 = T(0) + g(0), \quad T(0.5) + g(0.5) < 23 - 6 = T(2) + g(2),$$
$$T(0.5) + g(0.5) \leq T(6.5) + g(6.5) = 68 - 9, \quad T(0.5) + g(0.5) \leq T(9.5) + g(9.5) = 92 - 11.$$

ning seega 0.5 on miinimum.

Ülesanne (4.2.15) kui tingimustega planeerimisülesanne. Olgu h ja T suvalised funktsioonid mingil hulgal Γ ja $D > 0$. Vaatleme ülesannet

$$\min_{\gamma \in \Gamma} \big(h(\gamma) + DT(\gamma)\big) \tag{4.2.19}$$

ja olgu γ^* (üks paljudest kui neid on mitu) selle ülesande lahend. Siis leidub konstant C (mis võib sõltuda valitud γ^*-st kui neid on mitu) nii, et γ^* on ka järgmise ülesande lahend:

$$\min_{\gamma \in \Gamma} h(\gamma) \tag{4.2.20}$$
$$\text{nii, et} \quad T(\gamma) \leq C.$$

Tõepoolest, võtame $C := T(\gamma^*)$ ja veendume, et γ^* on ka (4.2.20) lahend. Kui see ei ole nii, siis leidub $\gamma_o \in \Gamma$ nii, et $h(\gamma_o) < h(\gamma^*)$ ning $T(\gamma_o) \leq T(\gamma^*) = C$. Aga sellisel juhul

$$h(\gamma_o) + DT(\gamma_o) < h(\gamma^*) + DT(\gamma^*),$$

mis on vastuolus sellega, et γ^* on (4.2.19) lahend.

Kas aga kehtib ka vastupidine väide: iga ülesande (4.2.20) lahend γ_o on mingi sobiva konstandi $D > 0$ korral ka (4.2.19) lahend? See on nii, kui kehtib tingimus: leidub $D > 0$ nii, et ülesande (4.2.19) (mingi) lahend γ^* (sõltub D-st) rahuldab tingimust $T(\gamma^*) = C$. Tõepoolest, sellisel juhul γ_o on ka (4.2.19) lahend, sest vastasel juhul

$$h(\gamma_o) + DT(\gamma_o) > h(\gamma^*) + DT(\gamma^*) = h(\gamma^*) + DC.$$

Et aga $T(\gamma_o) \leq C$, siis ülaltoodu võrratusest järeldub, et

$$h(\gamma_o) - h(\gamma_D) > D\big(C - T(\gamma_o)\big) \geq 0,$$

mis oleks aga vastuolus eeldusega, et γ_o on (4.2.20) lahend.

Kontranäide: Vaatleme veelkord ülatoodud kontranäidet. Et $T(0.5) = 20$, siis $\gamma_o = 0.5$ on järgmise ülesande lahend

$$\min_{\gamma \geq 0} -\gamma$$
$$\text{nii et} \quad T(\gamma) \leq 20.$$

Eelpool aga nägime, et 0.5 pole ühegi $D > 0$ korral ülesande (4.2.15) lahend.

Seega Üldiselt pole ülesanded (4.2.19) ja (4.2.20) ekvivalentsed: ülesande (4.2.19) iga lahend on (sobiva C korral) mingi kujul (4.2.20) oleva ülesande lahend, kuid vastupidine ei pruugi alati kehtida.

Ülesanne **4.9** *Olgu T kumer ja rangelt kasvav funktsioon ning olgu h diferentseruv, kumer ja rangelt kahanev funktsioon. Olgu $\gamma_o > 0$ järgmise ülesande lahend:*

$$\min_{\gamma \geq 0} h(\gamma)$$
$$\text{nii et} \quad T(\gamma) \leq C.$$

Tõestada, et leidub $D > 0$ nii, et γ_0 on ka (4.2.15) lahend.

4.2.4 1-norm soft margin

Üks võimalik valik funktsioonideks u ja h on $u(x) = x$ ja $h(\gamma) = -\gamma$. Sellisel juhul (4.2.14) on

$$\min_{w:\|w\|=1, w_o, \gamma \geq 0} \Big(-\gamma + D \sum_{i=1}^{n} \big(\gamma - y_i(w'x_i + w_o)\big)_+ \Big). \tag{4.2.21}$$

Pane tähele: kui $D < \frac{1}{n}$ siis $\gamma^* = \infty$. Teisest küljest aga, kui valim pole lineaarselt eralduv, siis leidub $D_o \leq 1$ nii, et $\gamma^* = 0$ iga $D > D_o$ korral. Peatükis 4.2.6 näitame, et ülesanne (4.2.21) on tihedalt seotud järgmise ülesandega

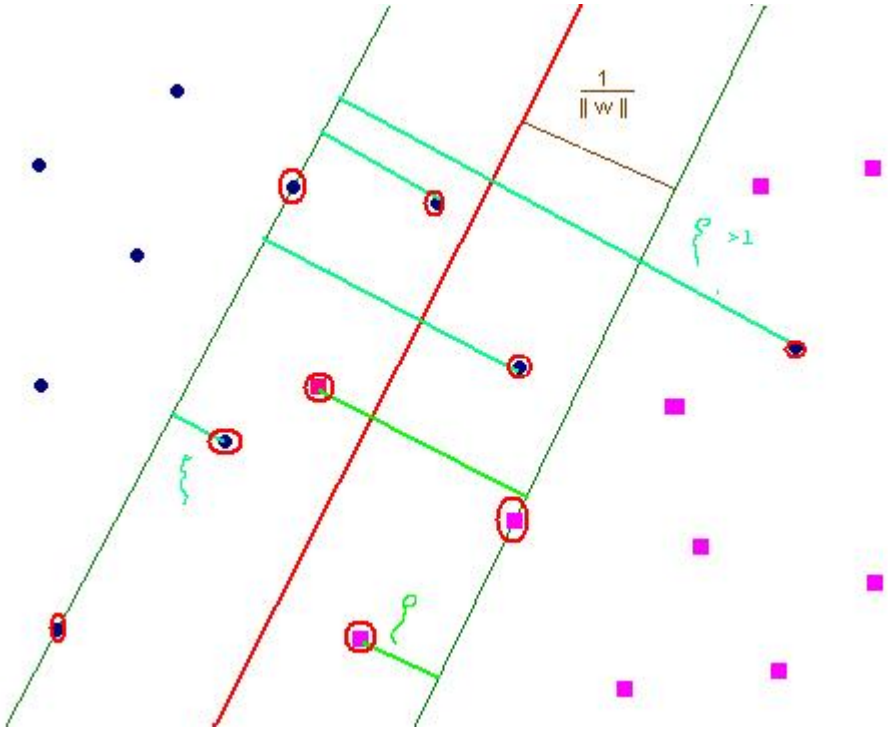
$$\min_{w:\|w\|=1, w_o, \gamma \geq 0} \Big(\frac{1}{2}\frac{1}{\gamma^2} + \frac{C}{\gamma} \sum_{i=1}^{n} \big(\gamma - y_i(w'x_i + w_o)\big)_+ \Big). \tag{4.2.22}$$

Täpsemalt: iga C korral leidub konstant D (sõltub andmetest) nii, et (4.2.21) lahend defineerib sama hüpertasandi kui (4.2.22) lahend. Kirjanduses vaadeldaksegi enamasti on ülesannet (4.2.22). Pane tähele, et seee ülesanne minimiseerib relatiivsete marginaalvigade summat. Täpselt nii nagu lineaarselt eralduva valimi korral lubab loobumine tingimusest $\|w\| = 1$ asendada γ suurusega $\frac{1}{\|w\|}$ nii, et probleem (4.2.22) on ekvivalentne järgmise nn **1-norm soft margin SVM** probleemiga:

$$\min_{w, w_o} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \big(1 - y_i(w'x_i + w_o)\big)_+. \tag{4.2.23}$$

Selle ülesande lahendamiseks defineeritakse abimuutujad (ik*slack variables*) ξ_i, $i = 1, \ldots, n$, mille kaudu (4.2.23) esitub

$$\min_{w, w_o, \xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{4.2.24}$$
$$\text{nii et } y_i(w'x_i + w_o) \geq 1 - \xi_i \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$



Lagrange'i funktsionaal on

$$\begin{aligned} L(w, w_o, \xi, \alpha, \gamma) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\gamma_i\xi_i + \sum_{i=1}^{n}\alpha_i\big(1 - \xi_i - y_i(w'x_i + w_o)\big) \\ &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\xi_i(C - \alpha_i - \gamma_i) + \sum_{i=1}^{n}\alpha_i\big(1 - y_i(w'x_i + w_o)\big). \end{aligned}$$

Võttes tuletised w w_o ning ξ järgi ja võrdsustades need nulliga (otsime ikka sadulpunkti), saame

$$\begin{aligned} w &= \sum_{i=1}^{n}\alpha_i y_i x_i \\ 0 &= \sum_{i=1}^{n}\alpha_i y_i \\ C &= \alpha_i + \gamma_i. \end{aligned}$$

Pannes saadud avaldised $L(w, w_o, \xi, \alpha, \gamma)$, saame duaalse funktsiooni

$$\theta(\alpha, \gamma) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j,$$

mis on sama, mis (4.2.1). Erinevus ülesande (4.2.4) duaalsest ülesandest (4.2.6) on tingimustes. Ülesande (4.2.23) duaaalne ülesanne on

$$\max_{\alpha,\gamma} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j, \tag{4.2.25}$$

$$\text{nii, et } \alpha_i \geq 0, \quad \gamma_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i + \gamma_i = C, \quad i = 1, \ldots, n.$$

Et γ ei ole maksimiseeritavas funktsioonis, on saadud ülesanne ekvivalentne

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j, \tag{4.2.26}$$

$$\text{nii, et } \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C.$$

Kokkuvõttes erineb ülesande (4.2.23) duaalne ülesanne ülesande (4.2.4) duaalsest ülesandest vaid lisatingimuse $\alpha_i \leq C$ lisamise läbi (*box constraint*).

Olgu α^* ja γ^* (4.2.26) lahendid. Siis optimaalne w^* esitub kujul

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$

Optimaalne marginaal on ikka $\frac{1}{\|w^*\|}$.

KKT tingimused on

$$\begin{aligned} &\gamma_i^* \xi_i^* = 0, \quad \forall i \\ &\alpha_i^* (1 - \xi_i^* - (w^{*\prime} x_i + w_o^*) y_i) = 0, \quad \forall i. \end{aligned} \tag{4.2.27}$$

ja lisaks kehtib

$$\gamma^* + \alpha^* = C$$

Tugivektorid on need valimi punktid, mille korral $\alpha_i^* > 0$. Seega seosest (4.2.27) saame

$$1 - \xi_i^* - (w^{*\prime} x_i + w_o^*) y_i = 0 \quad \Leftrightarrow \quad (w^{*\prime} x_i + w_o^*) y_i \leq 1.$$

Järelikult tugivektor x_i on marginaalveaga: ta asub kas hüpertasandi valel pool, (st $(w^{*\prime} x_i + w_o^*) y_i < 0$) või mitte kaugemal kui marginaal (st $0 \leq (w^{*\prime} x_i + w_o^*) y_i \leq 1$).

Võrdusest $\alpha_i^* + \gamma_i^* = C$ saame, et kui $0 < \alpha_i^* < C$, siis $\gamma_i^* > 0$ ja seega $\xi_i^* = 0$; seosest (4.2.27) omakorda saame $(w^{*\prime}x_i + w_o^*)y_i = 1$.
Kui $(w^{*\prime}x_i + w_o^*)y_i < 1$, siis $\xi_i^* > 0$, nii, et $\gamma_i^* = 0$ ja $\alpha_i^* = C$.
Kui $\alpha_i^* = 0$, siis $\gamma_i^* = C$ nii, et seosest (4.2.27) saame: $\xi_i^* = 0$ ja $(w^{*\prime}x_i + w_o^*)y_i \geq 1$.

Võttes ülalöeldu kokku, saame

$$\begin{aligned}(w^{*\prime}x_i + w_o^*)y_i > 1 \Rightarrow \alpha_i^* = 0 &\Rightarrow (w^{*\prime}x_i + w_o^*)y_i \geq 1\\ 0 < \alpha_i^* < C &\Rightarrow (w^{*\prime}x_i + w_o^*)y_i = 1\\ (w^{*\prime}x_i + w_o^*)y_i < 1 \Rightarrow \alpha_i^* = C &\Rightarrow (w^{*\prime}x_i + w_o^*)y_i \leq 1.\end{aligned}$$

Teinekord nimetatakse tugivektoreid, mille korral $0 < \alpha_i^* < C$ *in-bound* tugivektoriteks. Needsinased asuvad alati õigel pool oleval tugitasandil st nende kaugus hüpertasandist on täpselt marginaal. Tugivektoreid mille korral $\alpha_i^* = C$ nimetatakse aga *bound*-tugivektoriteks. Nemad on enamasti valel pool tugitasandit (kuid võivad olla õigel pool klassifitseerivat hüpertasandit).

Konstandi w_o^* saame nüüd määrata *in-bound* tugivektorite abil, st w_o^* peab olema selline, et $(w^{*\prime}x_i + w_o^*)y_i = 1$ iga i korral, millele vastav $\alpha_i^* \in (0, C)$. Klassifitseerimisreegel on endiselt (4.2.9).

Märkus: Sõltuvalt ülesandest ei pruugi *in-bound* tugivektoreid alati leida.

Ülesanne 4.10 *Olgu $d = 1$, $x_1 = y_1 = -1$, $x_2 = y_2 = 1$. Leida (4.2.23) lahend. Veendu, et teatud C korral $\alpha_1 = \alpha_2 = C$. Veendu, et sellisel juhul w_o^* pole ühene.*

Märkus kirjandusest: Enamikes raamatutes [15, 9, 12, 13] on 1-norm soft margin SVM optimeerimisülesanne kujul (4.2.24) ja me teame, et see on ekvivalentne ülesandega (4.2.22). Raamatus [14] on ülesanne kujul (4.2.21). Raamatus [7] on proleem defineeritud kujul (4.2.20), nimelt:

$$\min_{w,w_o,\xi} \|w\| \tag{4.2.28}$$

$$\text{nii et } y_i(w'x_i + w_o) \geq 1 - \xi_i \quad \xi_i \geq 0, \quad \sum_{i=1}^{n} \xi_i \leq C_o, \quad i = 1, \ldots, n.$$

Samas (lk 420) on kirjutatud "Computationally it is convenient to re-express (4.2.28) in the equivalent form (4.2.24)". See, nagu teame, pole aga päris nii, sest need ülesanded pole alati ekvivalentsed.

4.2.5 2-norm soft margin SVM

Eelmises peatükis käsitletud ülesande saime üldisest ülesandest (4.2.14) võttes $h(\gamma) = -\gamma$ ja $u(x) = x$. Teine populaarne valik on ruutfunktsioon $u(x) = x^2$ ja nii same ülesande

$$\min_{w:\|w\|=1,w_o,\gamma\geq 0} \Big(-\gamma + D\sum_{i=1}^{n} \big(\gamma - y_i(w'x_i + w_o)\big)_+^2 \Big). \tag{4.2.29}$$

Nagu 1-norm ülesanneta korral, on ka (4.2.29) seotud järgmise ülesandega:

$$\min_{w:\|w\|=1,w_o,\gamma\geq 0}\Big(\frac{1}{2}\frac{1}{\gamma^2}+\frac{C}{2\gamma^2}\sum_{i=1}^{n}\big(\gamma-y_i(w'x_i+w_o)\big)_+^2\Big). \tag{4.2.30}$$

Täpsemalt, iga (4.2.30) lahend on ka (4.2.29) lahend. Ning (täpselt nii nagu eelpoolvaadeldud ülesannete korralgi) loobudes nõudest $\|w\|=1$ saame ülaltoodud probleemile ekvivalentse kuju

$$\min_{w,w_o}\Big(\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{n}\big(1-y_i(w'x_i+w_o)\big)_+^2\Big). \tag{4.2.31}$$

Abimuutujate abil esitub see ülesanne (4.2.31)

$$\begin{aligned}&\min_{w,w_o,\xi}\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2\\ &\text{nii et } y_i(w'x_i+w_o)\geq 1-\xi_i\quad \xi_i\geq 0,\quad i=1,\dots,n.\end{aligned}$$

Kitsendusest $\xi_i\geq 0$ võime loobuda, sest ülesanne sellest ei muutu (veendu!), mistõttu ülaltoodud ülesanne on ekvivalentne ülesandega

$$\begin{aligned}&\min_{w,\xi}\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2\\ &\text{nii et } y_i(w'x_i+w_o)\geq 1-\xi_i\quad i=1,\dots,n.\end{aligned}$$

Lagrange'i funktsionaal on

$$\begin{aligned}L(w,w_o,\xi,\alpha)&=\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2+\sum_{i=1}^{n}\alpha_i\big(1-\xi_i-y_i(w'x_i+w_o)\big)\\ &=\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2-\sum_{i=1}^{n}\xi_i\alpha_i+\sum_{i=1}^{n}\alpha_i\big(1-y_i(w'x_i+w_o)\big)\\ &=\frac{1}{2}\|w\|^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2-\sum_{i=1}^{n}\xi_i\alpha_i+\sum_{i=1}^{n}\alpha_i\big(1-y_iw'x_i\big)-w_o\sum_{i=1}^{n}\alpha_iy_i.\end{aligned}$$

Võttes tuletised w w_o ning ξ järgi ja võrdsustades need nulliga, saame

$$\begin{aligned}w&=\sum_{i=1}^{n}\alpha_iy_ix_i\\ 0&=\sum_{i=1}^{n}\alpha_iy_i\\ 0&=C\xi_i-\alpha_i\quad\Leftrightarrow\quad \xi_i=\frac{\alpha_i}{C}\quad i=1,\dots,n.\end{aligned}$$

Asendades funktsionaalis $L(w, w_o, \xi, \alpha)$ saadud avaldised, saame duaalse funktsiooni

$$\begin{aligned}\theta(\alpha) &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j + \frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2 - \frac{1}{C}\sum_{i=1}^{n}\alpha_i^2 \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}^{n}\alpha_i\alpha_j y_i y_j x_i' x_j - \frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2 \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}^{n}\alpha_i\alpha_j y_i y_j \big(x_i' x_j + \frac{1}{C}\delta_{ij}\big),\end{aligned}$$

kus δ_{ij} on Kroneckeri delta. Seega duaalne ülesanne on

$$\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}^{n}\alpha_i\alpha_j y_i y_j \big(x_i' x_j + \frac{1}{C}\delta_{ij}\big) \tag{4.2.32}$$

$$\text{nii, et } \alpha_i \geq 0, \quad \sum_{i=1}^{n} y_i\alpha_i = 0. \tag{4.2.33}$$

ja see erineb duaalsest ülesandest (4.2.6) vaid seeläbi, et ruutvormis $\sum_{i,j}^{n}\alpha_i\alpha_j y_i y_j (x_i' x_j)$ on maatriksi $(x_i' x_j)_{ij}$ peadiagonaalile liidetud konstant $\frac{1}{C}$.

KKT tingimus on

$$\sum_{i=1}^{n}\alpha_i^*\big(1 - \xi_i^* - y_i(w^{*\prime} x_i + w_o^*)\big) = 0,$$

mistõttu $\alpha_i^* > 0$ vaid siis, kui $y_i(w^{*\prime} x_i + w_o^*) \leq 1$. Tugivektor, nagu ikka, defineeritakse kui selline x_i, mille korral $\alpha_i^* > 0$. Määramaks konstanti w_o^*, kasutame seost $\xi_i = \frac{\alpha_i^*}{C}$: kui $\alpha_i^* > 0$, siis

$$1 - \frac{\alpha_i^*}{C} = y_i(w^{*\prime} x_i + w_o^*)$$

ja siit saame w_o^*.

Ülesanne 4.11 *Tõestada, et*

$$\|w^*\|^2 = \sum_i \alpha_i^* - \frac{1}{C}\sum_i (\alpha_i^*)^2.$$

Seega marginaal

$$\gamma = \Big(\sum_i \alpha_i^* - \frac{1}{C}\sum_i (\alpha_i^*)^2\Big)^{-\frac{1}{2}}.$$

Ülesanne 4.12 *Üldistada ülesandeid (4.2.23) ja (4.2.31) juhule, kus igal valimi punktil on oma penalty-konstant C_i.*

Ülesanne 4.13 *Leida probleemi*

$$\min_{w,w_o,\xi} \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{n} \xi_i + \frac{C_2}{2} \sum_{i=1}^{n} \xi_i^2$$
$$nii\ et\ y_i(w'x_i + w_o) \geq 1 - \xi_i \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

duaalne ülesanne.

Kuidas valida konstanti C? Ühest vastust ei ole, enamasti proovitakse mitmeid erinevaid konstante ning seejärel valitakse parim. Proovimiseks jagatakse valim kaheks: ühe osa peal treenitakse (antud konstandiga) kassifitseerija, teise osa peal kontollitakse saadud klassifitseerija headust.

4.2.6 Ülesannete (4.2.23) ja (4.2.21) omavaheline seos*

Tuletame meelde probleemi (4.2.23) lahendid:

$$w^* = \sum_i \alpha_i^* y_i x_i, \quad w_o^* = -\frac{1}{2}(w^{*\prime} x_i + w^{*\prime} x_j), \quad \text{kus } 0 < \alpha_i^*, \alpha_j^* < C \quad \text{and } y_i = 1, y_j = -1.$$

Siin α^* on duaalse ülesande (4.2.26) lahend:

$$\max_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j,$$
$$\text{nii et } \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{ja } \alpha_i \in [0, C], \quad \forall i.$$

Alljärgnves veendume, et iga $C > 0$ leiduvad konstandid D ja $c > 0$ nii, et cw^* ja cw_o^* on järgmise ülesande (4.2.21) lahendid:

$$\min_{u:\|u\|=1,b,\gamma\geq 0} \Big(-\gamma + D \sum_{i=1}^{n} \big(\gamma - y_i(u'x_i + b)\big)_+ \Big).$$

Sellisel juhul määravad mõlemad ülesanded ühe ja sama hüpertasandi. Olgu $A = \sum_{i=1}^{n} \alpha_i^*$ ning defineerime $D := \frac{C}{A}$. Abimuutujate kaudu avaldub (4.2.21) järgmiselt:

$$\min_{u,b,\gamma,\xi} \quad -\gamma + D \sum_{i=1}^{n} \xi_i$$
$$\text{nii et } y_i(u'x_i + b) \geq \gamma - \xi_i \quad \xi_i \geq 0, \quad i = 1, \ldots, n, \quad \|u\| = 1.$$

Lagrange'i funktsionaal:

$$L(u, b, \gamma, \xi, \beta, \rho, \lambda) := -\gamma + D \sum_i \xi_i - \sum_i \beta_i \big(y_i(u'x_i - b) - \gamma + \xi_i\big) - \sum_i \rho_i \xi_i + \lambda\big(\|u\|^2 - 1\big),$$

kus $\rho_i \geq 0, \beta_i \geq 0$. Paneme tähele, kui $\lambda \leq 0$, siis $\theta(\beta, \rho, \lambda) = -\infty$.
[Selles veendumiseks vali γ, ξ, u ja b nii, et kõik n võrratust kehtiksid, selline komplekt kindlasti leidub. See tähendab, et

$$\sum_i \beta_i y_i u' x_i \geq \sum_i \beta_i (\gamma - \xi_i - b y_i)$$

ning see võrratus kehtib ka siis kui u korrutada läbi kuitahes suure absoluutväärtusega (positiivse, kui vasak pool on positiivne ja negatiivse, kui vasaka pool on negatiivse) konstandiga ja nii läheneb keskmine tegur ülaltoodud avladise piirväärtusele $-\infty$. Seeläbi kasvab $\|u\|^2$ aga kui $\lambda \leq 0$, siis ka see tegur ei kasva.]
Seetõttu ei saa $\lambda \leq 0$ maksimiseerida duaalset funktsiooni ning seega vaatleme duaalses ülesandes ainult olukorda $\lambda > 0$. Duaalne ülesanne on (vt ka [14], ch 7.2 või ülesanne 4.4)

$$\max_{\beta \geq 0, \lambda > 0} \quad -\frac{1}{4\lambda} \sum_{i,j}^{n} \beta_i \beta_j y_i y_j x_i' x_j - \lambda, \tag{4.2.34}$$

$$\text{nii, et } \sum_{i=1}^{n} \beta_i y_i = 0, \quad \beta_i \in [0, D], \quad \sum_{i=1}^{n} \beta_i = 1.$$

Seda saab maksimiseerida kahes osas: lahendada ülesanne

$$\min_{\beta \geq 0} \sum_{i,j}^{n} \beta_i \beta_j y_i y_j x_i' x_j, \tag{4.2.35}$$

$$\text{nii, et } \sum_{i=1}^{n} \beta_i y_i = 0, \quad \beta_i \in [0, D], \quad \sum_{i=1}^{n} \beta_i = 1$$

ning saadud lahendi, olgu see β^*, kaudu avalda optimaalne

$$\lambda^* = \frac{1}{2} \sqrt{\sum_{i,j}^{n} \beta_i^* \beta_j^* y_i y_j x_i' x_j}.$$

Seega esialgse probleemi lahendid on

$$u^* = \frac{1}{2\lambda^*} \sum_i \beta_i^* y_i x_i, \quad b^* = -\frac{1}{2}\big(u^{*\prime} x_j + u^{*\prime} x_i\big), \quad \text{kus } 0 < \beta_i^*, \beta_j^* < D \quad \text{ja } y_i = 1, y_j = -1.$$

Näitame nüüd, et duaalsete probleemide (4.2.26) ja (4.2.35) lahendid on omavahel seotud: iga i korral $\frac{\alpha_i^*}{A} = \beta_i^*$. Veendume selles. Et

$$\sum_i \frac{\alpha_i^*}{A} = 1, \quad \frac{\alpha_i^*}{A} \in [0, \frac{C}{A}] = [0, D],$$

siis tingimused on täidetud. Oletades vastuväiteliselt, et see pole nii, siis leiduksid konstandid $\beta_i \in [0, D]$ nii, et $\sum_i \beta_i = 1$ ja järgmine võrratus kehtiks:

$$\sum_{i,j}^{n} \beta_i \beta_j y_i y_j x_i' x_j < \frac{1}{A^2} \sum_{i,j}^{n} \alpha_i^* \alpha_j^* y_i y_j x_i' x_j.$$

Defineerides $\alpha_i := A\beta_i$, me saame, et $\alpha_i \in [0, C]$ ja

$$\sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j < \sum_{i,j}^{n} \alpha_i^* \alpha_j^* y_i y_j x_i' x_j,$$

mis on vastuolus eeldusega, et α^* on (4.2.26) lahend. Seega $\frac{\alpha_i^*}{A} = \beta_i^*$ millest järeldub, et

$$\frac{1}{A} \sum_i \alpha_i^* y_i x_i = \frac{w^*}{A} = \sum_i \beta_i^* y_i x_i = 2\lambda^* u^*.$$

Kui $0 < \alpha_i^*, \alpha_j^* < C$ ja $y_i = 1, y_j = -1$, siis $0 < \beta_i^*, \beta_j^* < D$. Seega

$$w_o^* = -\frac{1}{2}(w^{*\prime} x_i + w^{*\prime} x_j) = -2\lambda^* A \frac{1}{2}(u^{*\prime} x_i + u^{*\prime} x_j) = 2A\lambda^* b^*.$$

Järelikult

$$w^* = (2A\lambda^*)u^*, \quad w_o^* = (2A\lambda^*)b^*.$$

Et $\|u^*\| = 1$, saame $\|w^*\| = 2A\lambda^*$. Tuletame meelde, et $\|w^*\|^{-1} = \gamma^*$, kus γ^* on marginaal, nii, et $u^* = \gamma^* w^*$ ja $b^* = \gamma^* w_o^*$.

Kokkuvõttes: Iga $C > 0$ korral leiduvad konstandid $D := \frac{C}{A}$ ja $c := 2A\lambda^* > 0$ nii, et $w^* = cu^*$ ja $w_o^* = cb^*$, mistõttu nende poolt tekitatud hüpertasandid on võrdsed ja et $c > 0$, siis on ka nende tekitatud klassifitseerijad võrdsed.

Konstantidest D ja C. Et $u^* = \gamma^* w^*$, $b^* = \gamma^* w_o$, saame et $(w^{*\prime} x_i + w_o^*) y_i = 1$ parajasti siis kui $(u^{*\prime} x_i + b^*) y_i = \gamma^*$. Et $\frac{\alpha_i^*}{A} = \beta_i^*$, kordajate α_i^* omadustest saame

$$\begin{aligned}
(u^{*\prime} x_i + b^*) y_i > \gamma^* &\Rightarrow \beta_i^* = 0 \Rightarrow (w^{*\prime} x_i + b^*) y_i \geq \gamma^* \\
0 < \beta_i^* < D &\Rightarrow (u^{*\prime} x_i + b^*) y_i = \gamma^* \\
(u^{*\prime} x_i + b^*) y_i < \gamma^* &\Rightarrow \beta_i^* = D \Rightarrow (u^{*\prime} x_i + b^*) y_i \leq \gamma^*.
\end{aligned}$$

Märgime, et toodud seosed järelduvad ka vahetult (ilma α_i-dele üleminekute) KKT tingimustest. Kuid erinevalt kordajatest α^*, kehtib $\sum_i \beta_i^* = 1$. Seetõttu saame konstandile D teatava interpretatsiooni. Tõepoolest, et $\beta_i^* \leq D$, saame $D \geq \frac{1}{n}$, sest vastasel juhul tingimus $\sum_i \beta_i^* = 1$ ei saa kehtida. Tihti on konstandil D järgmine kuju: $D = \frac{1}{\nu n}$, kus $\nu \in (0, 1]$. Sellisel juhul parameeter ν kontrollib *bound* tugivektorite arvu: nende vektorite korral $\beta_i^* = \frac{1}{\nu n}$ ja seosest $\sum_i \beta_i^* = 1$ järeldub, et selliseid tugivektoreid ei saa olla rohkem

kui νn. Iga marginaalvea korral vastav kordaja on *bound*-tugivektor ja seega ν kontrollib marginaalvigade proportsiooni. Samamoodi järeldub, et tugivektorite arv peab olema vähemalt νn, sest vähemalt niipalju kordajaid peab olema positiivsed. Seega D valimine aitab kontrollida tugivektorite ja marginaalvigade arvu.
Konstandil C pole sellist ilusat interpretatsiooni, sest kuigi $C = DA$, ei tea me A väärtust enne (duaalse) ülesande lahendamist. Kahtlemata sõltub regulariseerimiskonstandist C lahend ja seetõttu on selle valik oluline. Praktikas kasutatakse selleks erinevaid ristvalideerimisi (*cross-validation*).

Üksühene vastavus. Näitamaks, et konstandile $D > 0$ vastab mingi teine konstant $C > 0$ nii, et (4.2.21) lahendid oleksid proportsionaalsed (4.2.23) lahenditega, võib kasutada sama argumenti. Argument eeldab aga, et leidub C nii, et $D = \frac{C}{A}$ (tuleta meelde, et A sõltub ka konstandist C). Tuletame meelde, et

$$\|w^*\|^2 = \sum_{i,j}^{n} \alpha_i^* \alpha_j^* y_i y_j x_i' x_j,$$

millest esialgse ja duaalse miinimumi võrdusest saame

$$\theta(\alpha^*) = A - \frac{1}{2}\|w^*\|^2 = \frac{1}{2}\|w^*\|^2 + C\sum_i \xi_i^* \quad \Rightarrow \quad A = \|w^*\|^2 + C\sum_i \xi_i^*.$$

4.2.7 Näide ülesannete (4.2.23) ja (4.2.21) ekvivalentsusest*

Olgu $d = 1$, $x_1 = -2$, $x_2 = x_3 = 0$; $y_1 = +1, y_2 = y_3 = -1$, kus $k > 4$. Pole raske veenduda, et

$$\min_{w:\|w\|=1, w_o} \Big(\frac{1}{2\gamma^2} + \frac{C}{\gamma} \sum_{i=1}^{n} \big(\gamma - y_i(w'x_i + w_o)\big)_+ \Big) = \frac{1}{2\gamma^2} + \frac{C}{\gamma}(2\gamma - 2)_+.$$

Kui $\gamma \leq 1$, siis optimaalne klassifitseerija on punktis $-\gamma$ (veendu!). Seega ülesannet on võimalik lihtsalt lahendada ilma duaalsele ülesandele üle minemata: funktsiooni

$$\gamma \mapsto \frac{1}{2\gamma^2} + C\frac{(2\gamma - 2)_+}{\gamma}$$

minimiseeriv $\gamma^*(C)$ on järgmine

$$\gamma^*(C) = \begin{cases} 1, & \text{kui } C \geq \frac{1}{2}; \\ \frac{1}{2C}, & \text{kui } C < \frac{1}{2}. \end{cases}$$

Vaadeldava ülesande duaalne ülesanne on

$$\begin{aligned} \max \quad & \sum_{i=1}^{3} \alpha_i - \frac{1}{2} 4\alpha_2^2 \\ \text{nii, et} \quad & \alpha_1 = \alpha_3 + \alpha_2, \quad \alpha_i \in [0, C]. \end{aligned}$$

Et

$$\sum_{i=1}^{3}\alpha_i - \frac{1}{2}4\alpha_2^2 = 2(\alpha_1 - \alpha_1^2),$$

siis lahend: $\alpha_2^* + \alpha_3^* = \alpha_1^*$ ja

$$\alpha_1^* = \begin{cases} \frac{1}{2}, & \text{kui } C > \frac{1}{2}; \\ C, & \text{kui } C \le \frac{1}{2}. \end{cases}$$

Kui $C \le \frac{1}{2}$, siis lahend: $w^* = -2C$, marginaal $\gamma^*(C) = \frac{1}{2C}$. Kui C on suurem, siis $w^* = -1$ ja $\gamma^* = 1$ (see kattub eelpool leitud funktsiooniga $\gamma^*(C)$) Nii või teisiti, $w^* = -\frac{1}{\gamma^*}$.
Kordajad α_2^* ja α_3^* pole üheselt määratud, võttes mõlemad võrdseks $\frac{C}{2}$, saame *in-bound* tugivektorid, millest $y_2 w_o = 1$ ehk $w_o = -1$. Optimaalne klassifitseerija

$$g^*(x) = \operatorname{sgn}(w^* x - 1) \quad \text{ehk} \quad g^*(x) = +1 \quad \Leftrightarrow \quad x \le -\gamma^*.$$

Seega

$$A(C) = 2\alpha_1^* = \begin{cases} 1, & \text{kui } C > \frac{1}{2}; \\ 2C, & \text{mujal.} \end{cases}$$

Vaatleme nüüd ülesannet

$$\min_{w:\|w\|=1, w_o, \gamma \ge 0} \Big(-\gamma + D\sum_{i=1}^{n} \big(\gamma - y_i(w' x_i + w_o)\big)_+ \Big).$$

Nagu enne, saame nüüd iga γ korral

$$\min_{w:\|w\|=1, w_o} \Big(-\gamma + D\sum_{i=1}^{n} \big(\gamma - y_i(w' x_i + w_o)\big)_+ \Big) = -\gamma + D(2\gamma - 2)_+$$

Seega

$$\gamma^*(D) = \begin{cases} 1, & \text{kui } D > \frac{1}{2}; \\ [1, \infty) & \text{kui } D = \frac{1}{2}; \\ \infty & \text{kui } D < \frac{1}{2}. \end{cases}$$

Seega $D = \frac{1}{2}$ korral pole marginaal (ja lahend) ühene. Iga klassifitseerija kujul $g(x) = +1$, kui $x \le -\gamma$, kus $\gamma \ge 1$ on optimaalne.
Duaalne ülesanne

$$\begin{aligned} &\min \quad \beta_1^2 \\ &\text{nii, et} \quad \beta_1 = \beta_2 + \beta_3, \quad \sum_{i=1}^{3} \beta_i = 1, \quad \beta_i \in [0, D]. \end{aligned}$$

Kui $D < \frac{1}{2}$, siis sellel ülesandel lahendid pole, kui $D \ge \frac{1}{2}$, siis $\beta_1^* = \frac{1}{2}$. Pane tähele, et iga $C > 0$ korral

$$\beta_1^* = \frac{\alpha_1^*}{A(C)} = \frac{\alpha_1^*}{2\alpha_1^*} = \frac{1}{2}.$$

Seega $\lambda^* = \frac{1}{2}$ ja $u^* = \frac{1}{2\lambda^*}\beta_i^* = -1$.

Kui $D > \frac{1}{2}$, siis x_1 ja x_2 on erinevate märkidega *in-bound* tugivektorid ja optimaalse konstandi saame seosest

$$b^* = -\frac{1}{2}\big(u^{*\prime}x_1 + u^{*\prime}x_2\big) = -1$$

ja optimaalse marginaali saame seosest

$$0 < \beta_i^* < D \Rightarrow (u^{*\prime}x_i + b^*)y_i = \gamma^*. \tag{4.2.36}$$

Võttes $x_2 = 0$, saame $\gamma^* = -1b^* = 1$. Järelikult, kui $D > \frac{1}{2}$, siis tõepoolest $u^* = \gamma^* w^* = w^*$ ja $b^* = \gamma^* w_o^* = w_o^*$, kus w^* ja w_o^* on ülesande (4.2.23) lahendid suvalise $C > \frac{1}{2}$ korral.

Kui $D = \frac{1}{2}$, siis pole kaht erineva märgiga *in-bound* tugivektorit, küll aga saame seosest (4.2.36) (võttes $x_i = x_2$), et $b_o^* = -\gamma^*$. Samuti teame, et iga arv hulgast $[1, \infty)$ võib olla marginaal. Näitame, et iga selline lahend on sobiva C korral ka ülesande (4.2.23) lahend. Olgu $\gamma^* > 1$. Optimaalne klassifitseerija sellisel juhul on

$$g^*(x) = \operatorname{sgn}(u^*x + b_o^*) = \operatorname{sgn}(-x - \gamma^*).$$

Selline klassifitseerija klassifitseerib punkti $-\gamma^*$ abil. Võttes nüüd $C = \frac{1}{2\gamma^*}$, saame ülesande (4.2.23) lahendid: $w^* = -2C$, $w_o^* = -1$. Et $\gamma^* = \frac{1}{2C}$, siis

$$u^* = -1 = \gamma^* w^*, \quad b_o^* = \gamma^* w_o^* = -\gamma^*.$$

Seega, tõepoolest, iga (4.2.21) lahend on mingi sobiva C korral ka ülesande (4.2.23) lahend. Samuti kehtib vastupidine – need ülesanded on tõepoolest ekvivalentsed.

4.3 Riski hinnangutest

Funktsionaalne marginaal. Lineaarne klassifitseerija:

$$g(x) = \operatorname{sgn}\big(f(x)\big) = \begin{cases} 1, & \text{kui } f(x) \geq 0; \\ -1, & \text{kui } f(x) < 0. \end{cases} \tag{4.3.1}$$

kus $f(x) = w'x + w_0$ on lineaarne funktsioon. Kui $\|w\| = 1$ ja g klassifitseerib x_i korrektselt, siis $y_i f(x_i)$ (mittenegatiivne) on (x_i, y_i) geomeetriline marginaal. Üldistame geomeetrilise marginaali mõistet.

Olgu $f : \mathbb{R}^d \to \mathbb{R}$ mingi (mitte ilmtingimata lineaarne) funktsioon ja olgu g selle abil defineeritud klassifitseerija (4.3.1). Seega g klassifitseerib (x_i, y_i) korrektselt kui $y_i f(x_i) > 0$ ja valesti kui $y_i f(x_i) < 0$. Arvu $y_i f(x_i)$ on punkti (x_i, y_i) **funktsionaalne marginaal**.

Seosega (4.3.1) defineeritud klassifitseerija g risk ja empiiriline risk on

$$R(g) = \mathbf{P}(Y \neq \operatorname{sgn}(f(X)) \leq EI_{\{f(X)Y \leq 0\}}, \quad R_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\operatorname{sgn} f(x_i) \neq y_i\}} \leq \frac{1}{n}\sum_{i=1}^{n} I_{\{y_i f(x_i) \leq 0\}}.$$

Suurus $A_n(f)$. Olgu iga f korral

$$A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi_-(-f(x_i)y_i), \quad \phi_-(z) = \begin{cases} 1, & \text{if } z \geq 0; \\ 1 + \frac{z}{\gamma}, & \text{if } -\gamma \leq z \leq 0; \\ 0, & \text{if } z < -\gamma \end{cases}$$

Arvu $A_n(f)$ saab ülalt hinnata järgmiselt:

- Et $\phi_-(z) \leq I_{(-\gamma,\infty)}(z)$, siis

$$A_n(f) \leq \frac{1}{n}\sum_{i=1}^{n} I_{\{-f(x_i)y_i > -\gamma\}} = \frac{1}{n}\sum_{i=1}^{n} I_{\{f(x_i)y_i < \gamma\}} =: R_n^{\gamma}(f).$$

  Arv R_n^{γ} on marginaalvigade proportsioon. Tuleta meelde, et paari (x_i, y_i) loetakse marginaalveaks isegui kui $\operatorname{sgn} f(x_i) = y_i$ (st $f(x_i)y_i > 0$) kuid "klassifitseerimise usaldusväärsus" $f(x_i)y_i$ on väiksem kui γ.

- Kehtib

$$\phi_-(z) \leq (1 + \frac{z}{\gamma})_+, \quad \text{kus} \quad (x)_+ = \begin{cases} x, & \text{kui } x > 0; \\ 0, & \text{kui } x \leq 0. \end{cases}$$

  Seega,

$$A_n(f) \leq \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{y_i f(x_i)}{\gamma})_+ = \frac{1}{n\gamma}\sum_{i=1}^{n}(\gamma - y_i f(x_i))_+ = \frac{1}{n\gamma}\sum_{i=1}^{n}\xi_i,$$

  kus

$$\xi_i := (\gamma - y_i f(x_i))_+ \geq 0.$$

  Arv ξ_i on meile juba tuttav, ta mõõdab marginaalvea suurust.

Marginaaliga antud riski hinnangud. Tuletame meelde (4.2.8):

$$w^* = \sum_{i \in \mathrm{SV}} \alpha_i^* y_i x_i.$$

Vaatleme lineaarseid funktsioone

$$\mathcal{F}_1 := \big\{ f(x) = w'x : \quad \|w\| \leq 1 \big\}.$$

Seega lineaarne funktsioon $x \mapsto w^{*\prime}x$ kus $\|w^*\| \leq 1$ kuulub hulka $\mathcal{F}_1$.

Olgu $\gamma > 0$ fikseeritud ja olgu $f_n \in \mathcal{F}_1$ valimi $(X_1, Y_1), \ldots, (X_n, Y_n)$ põhjal valitud (riski hinnangute korral vaatleme valimit alati juhuslikuna). Olgu $g_n = \operatorname{sgn} f_n$ vastav klassifitseerija. Siis järgmised riski hinnanagud kehtivad ([14], p. 103): tõenäosusega $1 - \delta$

$$R(g_n) \leq A_n(f_n) + \frac{4}{n\gamma}\sqrt{\sum_{i=1}^{n} X_i'X_i} + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}. \tag{4.3.2}$$

Et $A_n(f) \le R_n^\gamma(g)$, seosest (4.3.2) saame (vaata ka [3] Cor 4.3): tõenäosusega vähemalt $1-\delta$,

$$R(g_n) \le R_n^\gamma(g_n) + \frac{4}{n\gamma}\sqrt{\sum_{i=1}^n X_i'X_i} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}. \tag{4.3.3}$$

See tõke on teatav teoreetiline põhjendus eesmärgile maksimiseerida marginaali ja samaaegselt minimiseerida marginaalvigade arvu.

Hinnang

$$A_n(f) \le \frac{1}{n\gamma}\sum_{i=1}^n \xi_i$$

koos seosega (4.3.2) annab veel ühe tõkke 1-normiga SVM tüüpi lineaarse klassifitseerija riskile ([14], Thm 4.17): tõenäosusega $1-\delta$,

$$R(g_n) \le \frac{1}{n\gamma}\sum_{i=1}^n \xi_i + \frac{4}{n\gamma}\sqrt{\sum_{i=1}^n X_i'X_i} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}. \tag{4.3.4}$$

See tõke on teatav teoreetiline põhjendus eesmärgile maksimiseerida marginaali ja samaaegselt minimiseerida marginaalvigade summat (nagu 1-norm SVM korral tehakse).

Asendades ϕ funktsiooniga ϕ^2 saame analoogiliselt 2-norm tüüpi riski hinnangud ([14], (7.13)): tõenäosusega vähemalt $1-\delta$,

$$R(g_n) \le \frac{1}{n\gamma^2}\sum_{i=1}^n \xi_i^2 + \frac{8}{n\gamma}\sqrt{\sum_{i=1}^n X_i'X_i} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}. \tag{4.3.5}$$

See tõke on teatav teoreetiline põhjendus eesmärgile maksimiseerida marginaali ja samaaegselt minimiseerida marginaalvigade ruutude summat (nagu 2-norm SVM korral tehakse).

Märkused:

1 Ülaltoodud võrratused kehtisid funktsioonidele ruumist $\mathcal{F}_1$ ja need funktsioonid ei sisalda konstanti. Seega funktsioon kujul $w'x + w_o$ sinna hulka ei kuulu isegi kui $\|w\| \le 1$. Samas võib konstandi lisada tunnusvektorile, suurendades selle dimensiooni ühe võrra. Kuid sellisel juhul peab tingimuse $\|w\| \le 1$ asendama tingimusega $\|w\|^2 + w_o^2 \le 1$ ja suuruse $\sqrt{\sum_{i=1}^n X_i'X_i}$ peab asendama suurusega $\sqrt{\sum_{i=1}^n X_i'X_i + 1}$. Seega formaalselt ülaltoodud riski hinnaguid vahetult SVM abil saadud klassifitseerijate rakendada üldjuhul ei saa. Saamaks funktsiooni klassist $\mathcal{F}_1$ tuleb funktsiooni $w'x+w_o$ kordajad (nii vektor w kui konstant w_o) läbi jagada suurusega $\sqrt{\|w\|^2 + w_o^2}$, kuid seejuures tuleb silmas pidada, et ka esialgne marginaal γ tuleb sama arvuga läbi jagada (ja $\sqrt{\sum_{i=1}^n X_i'X_i}$ asemel on nüüd $\sqrt{\sum_{i=1}^n X_i'X_i + 1}$).

2 Kõik ülalnimatatud tõkked kehtivad vaid fikseeritud γ korral, kuid mitte ühtlaselt (üle kõikide γ-de); ühtlase tõkke korral oleks paremal pool veel lisaliige. Seega formaalselt need tõkked ei kehti, kui γ on andmete põhjal valitud nagu näiteks SVM korral. Kuigi seda on kirjanduses märgitud (näiteks Remark 7.17 raamatus [14]), kasutatakse neid hinnanguid siiski ka SVM korral (näiteks sealsamas raamatus [14]).

3 Tuleta meelde, et klassi VC-dimensioon oli maksimaalne arv, mida see klass on võimeline tükeldama. Nõudes tükeldust kindla marginaaliga, saame VC dimensiooni teatava üldistuse *fat-shattering* dimensioon. Juhul kui ruum on tõkestatud st kõik punktid asuvad keras rasadiusega $r < \infty$, siis hüpertasandite *fat-shattering* dimensioon võib olla väiksem kui VC dimensioon (tõkestamata ruumi korral nad on võrdsed). *Fat-shattering* dimensiooni abil saab defineerida omaette klassi riski hinnang marginaalide kaudu. Kõik need hinnangud eeldavad, et tunnusvektor on tõkestatud, st mingi $r < \infty$ korral $\mathbf{P}(\|X\| \leq r) = 1$. Selliseid riski hinnanguid võib leida näiteks [5], Ch 10, [12], Ch. 4.

4.4 Tugivektorklassifitseerijad ja teised tuumameetodid

4.4.1 Andmete teisendamine kõrgdimensionaalsesse ruumi

Tugivektorklassifitseerijate (ik *support vector machine (SVM) classifiers*) peamine idee on väga lihtne: teatud kujutise

$$\Phi : \mathbb{R}^d \to \mathcal{H}$$

abil projekteeritakse ruumis $\mathbb{R}^d$ antud tunnusvektorid ruumi $\mathcal{H}$. Nii saadakse valim

$$(\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n),$$

kus tunnusvaktorid on Hilberti ruumis $\mathcal{H}$. Sellele teisendatud valimile rakendatakse eelmises peatükis käsitletud lineaarseid klassifitseerimismeetode, enamasti marginaalmeetodi. Nii saadakse linaarne klassifitseerija ruumis $\mathcal{H}$, esialgses ruumis $\mathbb{R}^d$ ei pruugi aga saadud klassifitseerija olla üldsegi lineaarne. Võttes $\mathcal{H} = \mathbb{R}^d$ ja $\Phi(x) = x$, saame erijuhuna eelmises peatükis käsitletud lineaarsed klassifitseerijad.

Andmete analüüsiks sobivamale ruumile projekteerimise idee on sama vana kui statistika (meenuta LDAd), kuid tugivektormeetodite eripära on see, et (erinevalt klassikalisest mitmemõõtmelisest statistikast) tunnusvektoreid ei projekteerita mitte madalama dimensiooniga (alam)ruumile vaid tihti hoopis palju kõrgema dimensiooniga ruumi. Teinekord on ruumi $\mathcal{H}$ dimensioon isegi lõpmatu. Esmapilgul näib selline lähenemisviis mõttetuna, sest kõrge dimensiooniga ruumis töötamine on arvutuslikult keerukam. Selgub aga, et leidub küllatki lai klass kujutisi Φ, mis teevad tugivektorklassifitseerijad võimalikuks.

Vaatleme jällegi olukorda, kus kaks klassi on märgistatud +1 ja -1. Olgu $\Phi : \mathbb{R}^d \to \mathcal{H}$, kus $\mathcal{H}$ on Hilberti ruum. Hüpertasand Hilberti ruumis on kujul

$$\{z \in \mathcal{H} : \langle w, z\rangle + w_0 = 0\},$$

kus $w \in \mathcal{H}$ ja $w_0 \in \mathbb{R}^1$. Teisendatud andmetele $\Phi(x_1), \ldots, \Phi(x_n)$ lineaarse klassifitseerija rakendamisel saame (üldiselt mittelineaarse) klassifitseerija esialgses ruumis $\mathbb{R}^d$:

$$g(x) = \operatorname{sgn}(\langle w, \Phi(x)\rangle + w_0). \tag{4.4.1}$$

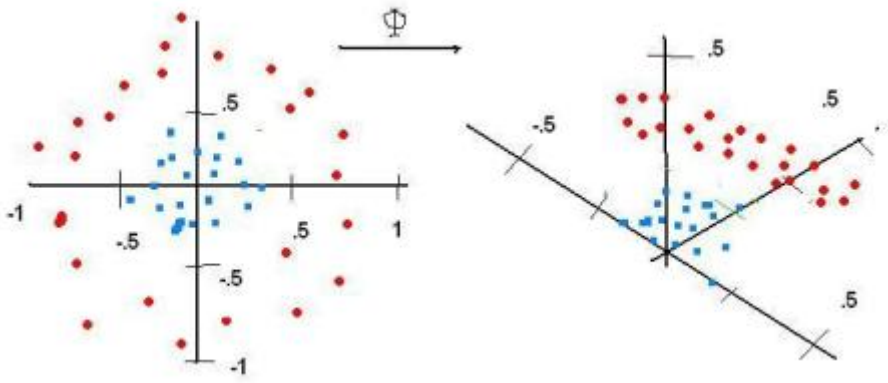
Näide. Olgu

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3, \quad \Phi(x) = \Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) = (z_1, z_2, z_3).$$

Lineaarne klassifitseerija ruumis $\mathbb{R}^3$ esitub hüpertasandi $w'z + w_0 = 0$ abil, esalgses ruumis on see klassifitseerija mittelineaarne, sest klassifitseerija (4.4.1) on

$$\operatorname{sgn}(w^1x_1^2 + w^2\sqrt{2}x_1x_2 + w^3x_2^2 + w_0), \quad \text{kus } w = (w^1, w^2, w^3).$$

Alljärgnev joonis illustreerib olukorda, kus tunnusvektorid pole lineaarselt eristuvad (esialgses ruumis $\mathbb{R}^2$), kuid teisendatud tunnused (ruumis $\mathbb{R}^3$) on.



1- ja 2-norm soft margin SVM. Lineaarse klassifitseerija leidmiseks ruumis $\mathcal{H}$ kasutame eelmises peatükis tutvustatud marginaalmeetode. 1-norm SVM probleem (4.2.23) on nüüd

$$\min_{w\in\mathcal{H}, w_o\in\mathbb{R}} \quad \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n}\big(1 - y_i(\langle w, \Phi(x_i)\rangle + w_o)\big)_+$$

Nii nagu eelmises peatükis, saab analoogiliselt (ja selle juurde tuleme hiljem tagasi, vt 4.6) näidata, et vastav duaalne ülesanne on

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j}^n \alpha_i\alpha_j y_i y_j \langle \Phi(x_i), \Phi(x_j)\rangle$$
$$\text{nii, et } C \geq \alpha_i \geq 0, \quad \sum_{i=1}^n y_i\alpha_i = 0.$$

2-norm SVM (4.2.31) probleem on nüüd

$$\min_{w\in\mathcal{H}, w_o\in\mathbb{R}} \quad \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^n \big(1 - y_i(\langle w, \Phi(x_i)\rangle + w_o)\big)_+^2.$$

ning selle ülesande duaalne ülesanne on kujul

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j}^n \alpha_i\alpha_j y_i y_j \big(\langle \Phi(x_i), \Phi(x_j)\rangle + \frac{1}{C}\delta_{ij}\big)$$
$$\text{nii, et } \alpha_i \geq 0, \quad \sum_{i=1}^n y_i\alpha_i = 0,$$

Kasutades duaalse ülesande lahendit α^*, saame mõlema probleemi lahendile kuju

$$w^* = w_{\alpha^*} = \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i). \tag{4.4.2}$$

Konstandid: 1-norm SVM konstandi w_o^* leiame võrrandist

$$y_i(\langle w^*, \Phi(x_i)\rangle + w_0^*) = y_i\big(\sum_{j=1}^n \alpha_j^* y_j \langle \Phi(x_j), \Phi(x_i)\rangle + w_o^*\big) = 1,$$

kus x_i on *in-bound* tugivektor st $0 < \alpha_i^* < C$.
2-norm SVM konstandi w_o^* saame võrdusest

$$1 - \frac{\alpha_i^*}{C} = y_i(\langle w^*, \Phi(x_i)\rangle + w_0^*) = y_i\big(\sum_{j=1}^n \alpha_j^* y_j \langle \Phi(x_j), \Phi(x_i)\rangle + w_o^*\big),$$

kus x_i on tugivektor $\alpha_i^* > 0$.

Klassifitseerija (4.4.1) on seega

$$g(x) = \operatorname{sgn}\big(\langle w^*, \Phi(x)\rangle + w_o^*\big) = \operatorname{sgn}\big(\sum_{i=1}^n \alpha_i^* y_i \langle \Phi(x_i), \Phi(x)\rangle + w_o^*\big). \tag{4.4.3}$$

Märkus. Kui $\mathcal{H}$ on jadaruum või $\mathbb{R}^m$, siis konstandi võib lisada tunnusvektorile ja klassifitseeriva hüpertasandi võib defineerida konstandita, st klassifitseerija (4.4.1) on $g(x) = \operatorname{sgn}\big(\langle w^*, \Phi(x)\rangle\big)$. Näiteks kui Φ on nagu ülaltoodud näites, st $\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, siis

$$\langle w, \Phi(x)\rangle + w_o = \langle u, \Phi^*(x)\rangle,$$

kus

$$\Phi^*(x^1, x^2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2, 1), \quad u' = (w', w_o) \in \mathbb{R}^4.$$

Enamasti käsitleme aga funktsioonide ruume ning sellisel juhul üldisust kitsendamata konstandist loobuda ei saa. Samas, nagu märgitakse raamatus ([13], lk 17), paljude praktikas kastutatavate kujutiste Φ korral ei ole konstandil erilist mõju klassifitseerijale, mistõttu (eriti teoreetilisemas kirjanduses) on sellest tihti loobutud.

4.5 Tuum

4.5.1 Tuumatrikk

Paneme tähele, et nii klassifitseerija (4.4.3) kui ka selleks vajaminevad konstandid α_i sõltuvad funktsioonist Φ vaid läbi skalaarkorrutise $\langle \Phi(x), \Phi(y)\rangle$. Seega klassifitseerija (4.4.3) leidmiseks pole vaja teada funktsiooni Φ, piisab kui teame funktsiooni

$$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad K(x, y) := \langle \Phi(x), \Phi(y)\rangle. \tag{4.5.1}$$

See tähelepanek teeb tugivektorklassifitseerimise võimalikuks, sest tihti on funktsiooni $K(x, y)$ võimalik arvutada teisendust Φ kasutamata. Seosega (4.5.1) antud funktsiooni K nimetame **tuumaks** (ik *kernel*). Funktsioon K on tuum, kui leidub (vähemalt üks) kujutis $\Phi : \mathbb{R}^d \to \mathcal{H}$, kus $\mathcal{H}$ on skalaarkorrutisega ruum nii, et (4.5.1) kehtib.

Näide. Eelpooltoodud näites on tuum

$$K(x, y) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{pmatrix} = (x_1y_1 + x_2y_2)^2 = (x'y)^2$$

ning selle arvutamiseks pole vaja kasutada teisendust Φ – piisab kui arvutame saklaarkorrutise esialgses ruumis ning siis võtame selle ruutu.
Paneme tähele, et sama tuuma defineerib ka näiteks funktsioon

$$\Phi' : \mathbb{R}^2 \to \mathbb{R}^4, \quad \Phi'(x_1, x_2) = (x_1^2, x_1x_2, x_1x_2, x_2^2).$$

Tuuma abil esitub klassifitseerija (4.4.1)

$$g(x) = \operatorname{sign}\Big(\sum_{i \in SV} \alpha_i^* y_i K(x_i, x) + w_o^*\Big) \tag{4.5.2}$$

kus kordajad α_i^* saadakse sobiva duaalse ülesande lahendina. Näiteks duaalne ülesanne (4.2.6) esitub tuuma abil

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j}^n \alpha_i\alpha_j y_i y_j K(x_i, x_j).$$

Ülaltoodust on selge, miks tugivektorklassifitseerimises ei pöörata tähelepanu mitte kujutisele Φ vaid tuumale K. Erinevad tuumad defineerivad erinevate omadustega klassifiseerijad kujul (4.5.2), mistõttu tuuma valimine on ekvivalentne klassifikatorite klassi valimisega. Kui tuum on valitud, leitakse saadud klassist (duaalse ülesande lahendamise läbi) teatud mõttes parim klassifitseerija.

Riski hinnangud. Riski hinnangud eelmisest peatükist kehtivad ka mittelineaarse SVM korral. Ruum $\mathcal{F}_1$ on praegu

$$\mathcal{F}_1 = \Big\{ f(x) = \langle w, \Phi(x)\rangle : \quad w \in \mathcal{H}, \quad \|w\| \le 1 \Big\}$$

ja iga klassifitseerija

$$g_n(x) = \operatorname{sgn}\big(\langle w, \Phi(x)\rangle\big) \quad \text{nii et } \langle \Phi(x), \Phi(x)\rangle = \|w\|^2 \le 1$$

ja *fikseeritud* γ korral on tõke (4.3.4) järgmine: tõenäosusega $1-\delta$

$$R(g_n) \le \frac{1}{n\gamma}\sum_{i=1}^n \xi_i + \frac{4}{n\gamma}\sqrt{\sum_{i=1}^n K(X_i, X_i)} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

Tõke (4.3.5) on nüüd: tõenäosusega $1-\delta$

$$R(g_n) \le \frac{1}{n\gamma^2}\sum_{i=1}^n \xi_i^2 + \frac{8}{n\gamma}\sqrt{\sum_{i=1}^n K(X_i, X_i)} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

4.5.2 Kuidas tuuma ära tunda?

Nägime, et kujutise Φ asendamine tuumaga K on igati loogiline, kuid siinkohal tekib oluline küsimus: kuidas tuuma ära tunda? Tõepoolest, vastavalt definitsioonile (4.5.1) on tuum defineeritud läbi kujutise Φ – funktsioon K on tuum parajasti siis, kui leidub sobivate omadustega Φ. Kuidas aga vaid funktsiooni K põhjal otsustada, et tegemist on tuumaga (st vastav Φ, mille kuju meid ehk ei huvitagi, leidub)? On selge, et olemaks tuum, peab K olema sümmeetriline, samuti peab ta rahuldama järgmist tingimust: iga lõpliku $x_1, \ldots, x_m$ korral peab maatriks $\big(K(x_i, x_j)\big)_{i,j}$ – **Grami maatriks** – olema positiivselt poolmääratud.

Ülesanne 4.14 *Tõesta, et* $\big(K(x_i, x_j)\big)_{i,j}$ *on positivselt poolmääratud.*

NB! Alljärgnevas $\mathcal{X}$ on suvaline hulk (mitte ilmtingimata $\mathbb{R}^d$).

Definitsioon 4.1 *Funktsioon*

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

on **positiivselt (pool)määratud**, *kui ta on sümmeetriline ning iga lõpliku $\mathcal{X}$ alamhulga korral on vastav Grami maatriks positiivselt (pool)määratud.*

Seega, kui K on tuum (st $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$, mingi $\Phi : \mathcal{X} \to \mathcal{H}$ korral), on ta positiivselt poolmääratud. Selgub, et positiivne poolmääratus on ka piisav selleks, et K oleks tuum.

Teoreem 4.2 *(***Moore-Aronszajn, 1950***) Funktsioon*

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

on tuum parajasti siis, kui kui ta on positiivselt poolmääratud.

Tõestus. Tarvilikkus on ilmne.
Piisavuse tõestus koosneb kahest osast. Esimeses osas tõestame, et leidub *skalaarkorrutisega ruum* $\mathcal{F}$ ja kujutis $\Phi : \mathcal{X} \to \mathcal{F}$ nii, et $\langle \Phi(x), \Phi(y) \rangle = K(x,y)$. Tõestuse teises osas täielikustame selle ruumi. Siis leidub *Hilberti ruum* $\mathcal{H}$ nii, et $\Phi : \mathcal{X} \to \mathcal{H}$ ja $\langle \Phi(x), \Phi(y) \rangle = K(x,y)$. Vastavalt definitsioonile on siis K tuum.

1) Olgu K positiivselt poolmääratud funktsioon. Defineerime kujutise

$$\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}, \quad \Phi(x) = K(x, \cdot). \tag{4.5.3}$$

Nüüd oleme defineerinud funktsioon Φ, kuid me peame veel defineerima sobiva kujutisruumi ning sobiva skalaarkorrutise. Vaatleme hulka

$$\mathcal{F} = \operatorname{span}\{K(x,\cdot) : x \in \mathcal{X}\} = \Big\{ \sum_{i=1}^{m} \alpha_i K(x_i, \cdot) : x_1, \ldots, x_m, m \in \mathbb{N}, \alpha_i \in \mathbb{R} \Big\}. \tag{4.5.4}$$

Hulk $\mathcal{F}$ on vektorruum. Defineerime seal skalaarkorrutise. Olgu

$$f = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot), \quad g = \sum_{i=1}^{n} \beta_i K(y_i, \cdot)$$

ning

$$\langle f, g \rangle = \Big\langle \sum_{i=1}^{m} \alpha_i K(x_i, \cdot), \sum_{i=1}^{n} \beta_i K(y_i, \cdot) \Big\rangle := \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j K(x_i, y_j). \tag{4.5.5}$$

Kõigepealt veendume, et (4.5.5) on korrektselt defineeritud. Selleks paneme tähele, et

$$\langle f, g \rangle = \sum_{i=1}^{m} \alpha_i g(x_i) = \sum_{j=1}^{n} \beta_j f(y_j), \tag{4.5.6}$$

mistõttu $\langle f, g \rangle$ on korrektselt defineeritud, st see ei sõltu f ja g esitusest. Veendume, et tegemist on skalaarkorrutisega. Selleks kontrollime aksioome:

- $\langle f, f\rangle \geq 0$
- $\langle f, g\rangle = \langle g, f\rangle$
- $\langle f + h, g\rangle = \langle f, g\rangle + \langle h, g\rangle$
- $\langle \lambda f, g\rangle = \lambda\langle f, g\rangle$
- $\langle f, f\rangle = 0 \Rightarrow f = 0$.

Esimene aksioom järeldub sellest, et K on positiivselt poolmääratud; teine aksioom järeldub sellest, et K on sümmeetriline; kolmas ja neljas järelduvad seosest (4.5.6). Seega on meil defineeritud poolskalaarkorrutis. See tähendab, et kehtib ka Cauchi-Schwartzi võrratus:

$$\langle f, g\rangle^2 \leq \langle f, f\rangle\langle g, g\rangle. \tag{4.5.7}$$

Viimase aksioomi tõestamiseks paneme eelkõige tähele, et iga f ja x korral

$$\langle f, K(x,\cdot)\rangle = \langle \sum_{i=1}^{m} \alpha_i K(x_i,\cdot), K(x,\cdot)\rangle = \sum_{i=1}^{m} \alpha_i K(x_i, x) = f(x). \tag{4.5.8}$$

Seega, kui $\langle f, f\rangle = 0$, siis iga $x \in \mathcal{X}$ korral

$$f^2(x) = |\langle K(x,\cdot), f\rangle|^2 \leq \langle K(x,\cdot), K(x,\cdot)\rangle\langle f, f\rangle = K(x,x)\langle f, f\rangle = 0,$$

st defineeritud poolskalaarkorrutis on tegelikult skalaarkorrutis.
Seega oleme defineerinud skalaarkorrutisega ruumi $\mathcal{F}$ nii, et iga x korral on $\Phi(x)$ selle element. Skalaarkorrutisega ruum on normeeritud (ja seega ka meetriline) ruum, norm on $\|f\| = \sqrt{\langle f, f\rangle}$ ning Cauchy-Schwartzi võrratus on

$$|\langle f, g\rangle| \leq \|f\|\|g\|.$$

2) Täielikustame selle ruumi. Olgu $\{f_n\}$ Cauchy jada. Iga $x \in \mathcal{X}$ korral

$$(f_n(x) - f_m(x))^2 = \langle f_n - f_m, K(x,\cdot)\rangle^2 \leq \|f_n - f_m\|^2\langle K(x,\cdot), K(x,\cdot)\rangle = \|f_n - f_m\|^2 K(x,x).$$

Seega on $\{f_n(x)\}$ Cauchy jada, mistõttu leidub $\lim_n f_n(x) := g(x)$. Laiendame hulka $\mathcal{F}$ funktsioonidega $x \mapsto g(x)$ – olgu see $\mathcal{H}$. Seega $\mathcal{F} \subset \mathcal{H}$ ning iga $h \in \mathcal{H}$ korral leidub vähemalt üks Cauchy jada $\{f_n\} \subset \mathcal{F}$ nii, et $f_n(x) \to h(x)$ iga x korral (kui $h \in \mathcal{F}$, siis võib võtta $f_n = h$ iga n korral). Laiendame skalaarkorrutist hulgale $\mathcal{H}$. Selleks defineerime

$$\langle h_1, h_2\rangle := \lim_n \langle f_n^1, f_n^2\rangle, \tag{4.5.9}$$

kus $\{f_n^1\} \subset \mathcal{F}$ on Cauchy jada nii, et $f_n^1(x) \to h_1(x)\ \forall x$ ja $\{f_n^2\} \subset \mathcal{F}$ on Cauchy jada nii, et $f_n^2(x) \to h_2(x)\ \forall x$. Veendume, toodud definitsioon on korrektne. Selleks veendume, et

- piirväärtus (4.5.9) leidub;

- piirväärtus (4.5.9) ei sõltu jadade f_n^1 ja f_n^2 valikust.

Piirväärtus leidub. Et $\{f_n^i\}$ on Cauchy jada, on ta alati tõkestatud, st $\exists K_i < \infty$ nii, et $\|f_n^i\| \le K_i$, $i = 0, 1$. Veendume, et jada $\{\langle f_n^1, f_n^2\rangle\}$ on Cauchy jada reaalarvude vallas. Olgu $\epsilon > 0$. Valime N nii suure, et iga $n, m > N$ korral

$$\|f_n^2 - f_m^2\| \le \frac{\epsilon}{2K_1}, \quad \|f_n^1 - f_m^1\| \le \frac{\epsilon}{2K_2}.$$

Kui $m, n > N$, siis

$$\begin{aligned}|\langle f_n^1, f_n^2\rangle - \langle f_m^1, f_m^2\rangle| &= |\langle f_n^1, f_n^2\rangle - \langle f_n^1, f_m^1\rangle + \langle f_n^1, f_m^1\rangle - \langle f_m^1, f_m^2\rangle| \\ &\le |\langle f_n^1, f_n^2 - f_m^2\rangle| + |\langle f_m^2, f_n^1 - f_m^1\rangle| \\ &\le \|f_n^1\|\|f_n^2 - f_m^2\| + \|f_m^2\|\|f_n^1 - f_m^1\| \\ &\le K_1\|f_n^2 - f_m^2\| + K_2\|f_n^1 - f_m^1\| \le \epsilon.\end{aligned}$$

Seega $\{\langle f_n^1, f_n^2\rangle\}$ on Cauchy jada, mistõttu piirväätus (4.5.9) leidub.

Veendume, et piirväärtus ei sõltu jadade valikust.
Kui erinevatele ekvivalentsiklassidele vastavad funktsioonid oleksid erinevad, siis lihtne: Olgu $\{g_n^i\}$ mingi teine Cauchy jada, mis koondub punktiviisiliselt funktsiooniks h_i. Analoogiliselt arutledes saame, et

$$|\langle f_n^1, f_n^2\rangle - \langle g_n^1, g_n^2\rangle| \le |\langle g_n^1, f_n^2 - g_n^2\rangle| + |\langle f_n^2, f_n^1 - g_n^1\rangle| \le \|g_n^1\|\|f_n^2 - g_n^2\| + \|f_n^2\|\|f_n^1 - g_n^1\|.$$

Üldiselt aga paneme tähele, et kui f_n ja g_n on kaks erinevat Cauchy jada, mis mõlemad defineerivad ühe ja sama funktsiooni h, siis iga $f \in \mathcal{F}$ korral $\lim_n\langle f, f_n\rangle = \lim_n\langle f, g_n\rangle$. Tõepoolest, kui $f = \sum_{i=1}^k \alpha_i K(x_i, \cdot)$, siis

$$\langle f, f_n\rangle - \langle f, g_n\rangle = \langle f, g_n - f_n\rangle = \sum_{i=1}^k \alpha_i\big(f_n(x_i) - g_n(x_i)\big) \to 0.$$

Seega, kui $f \in \mathcal{F}$, siis iga $h \in \mathcal{H}$ korral on skalaarkorrutis $\langle f, h\rangle$ korrektselt defineeritud. Veendume, et sama kehtib ka siis, kui mõlemad funktsioonid kuuluvad hulka $\mathcal{H}$. Olgu $f, g \in \mathcal{H}$. Olgu $\{f_n\}$ ja $\{g_n\}$ sellised Cauchy jadad, et $f_n(x) \to f(x)$ ja $g_n(x) \to g(x)$ iga x korral. Olgu $K < \infty$ nii, et $\|f_n\| < K$ ja $\|g_n\| < K$ iga n korral. Veendume järgmises:

$$\forall \epsilon > 0, \quad \exists N < \infty \quad \text{nii, et } |\langle f_n, g_m\rangle - \langle f_k, g_l\rangle| \le \epsilon, \quad \text{kui} \quad m, n, k, l > N. \qquad (4.5.10)$$

Tõepoolest, võttes N nii suure, et $\|g_m - g_l\| \le \frac{\epsilon}{2K}$ ja $\|f_n - f_k\| \le \frac{\epsilon}{2K}$, kui $m, n, k, l > N$, saame

$$\begin{aligned}|\langle f_n, g_m\rangle - \langle f_k, g_l\rangle| &\le |\langle f_n, g_m\rangle - \langle f_n, g_l\rangle| + |\langle f_n, g_l\rangle - \langle f_k, g_l\rangle| \\ &\le \|f_n\|\|g_m - g_l\| + \|g_l\|\|f_n - f_k\| \le \epsilon,\end{aligned}$$

kui $m,n,k,l > N$. Teame, et $\lim_n \langle f_n, g_m\rangle = \langle f, g_m\rangle$ ja $\lim_m \langle f_n, g_m\rangle = \langle f_n, g\rangle$. Seosest (4.5.10) järeldub aga, et jadad $\{\langle f, g_m\rangle\}$ ning $\{\langle f_n, g\rangle\}$ on Cauchy jadad (reaalarvude vallas, muidugi) ning mõlema jada piirväärtus on $\langle f, g\rangle$.
Tõestus, et $\{\langle f, g_m\rangle\}$ on Cauchy jada: võta suvaline $\epsilon > 0$ ja vali N nii suur, et iga $m,n,k,l > N$ korral

$$|\langle f_n, g_m\rangle - \langle f_k, g_l\rangle| < \frac{\epsilon}{3}. \tag{4.5.11}$$

Olgu $m,l > N$ ja vali $n,k > N$ nii, et

$$|\langle f, g_m\rangle - \langle f_n, g_m\rangle| < \frac{\epsilon}{3}, \quad |\langle f, g_l\rangle - \langle f_k, g_l\rangle| < \frac{\epsilon}{3}.$$

Nüüd

$$|\langle f, g_m\rangle - \langle f, g_l\rangle| \le |\langle f, g_m\rangle - \langle f_n, g_m\rangle| + |\langle f, g_l\rangle - \langle f_k, g_l\rangle| + |\langle f_n, g_m\rangle - \langle f_k, g_l\rangle| < \epsilon.$$

Olgu $\alpha_1 := \lim_m \langle f, g_m\rangle$. Veendume, et $\alpha_1 = \langle f, g\rangle$. Vali $\epsilon > 0$. Olgu N selline, et kehtib (4.5.11). Olgu $m > N$ selline, et $|\alpha_1 - \langle f, g_m\rangle| < \frac{\epsilon}{3}$. Olgu $n > N$ selline, et $|\langle f_n, g_m\rangle - \langle f, g_m\rangle| < \frac{\epsilon}{3}$. Nüüd iga $k > N$ korral

$$|\alpha_1 - \langle f_k, g_k\rangle| \le |\alpha_1 - \langle f, g_m\rangle| + |\langle f, g_m\rangle - \langle f_n, g_m\rangle| + |\langle f_n, g_m\rangle - \langle f_k, g_k\rangle| < \epsilon.$$

Seega $\alpha_1 = \lim_k \langle f_k, g_k\rangle = \langle f, g\rangle$.
Kokkuvõttes saame

$$\langle f, g\rangle = \lim_n \langle f_n, g_n\rangle = \lim_n \lim_m \langle f_n, g_m\rangle = \lim_n \langle f_n, g\rangle = \lim_m \langle f, g_m\rangle = \lim_m \lim_n \langle f_n, g_m\rangle. \tag{4.5.12}$$

Nüüd on lihtne näha, et $\langle f, h\rangle$ ei sõltu funktsioone f ja g defineeriva jada valikust. Tõepoolest, olgu f_n' mingi teine Cauchy jada nii, et $f_n'(x) \to f(x)$ iga x korral. Et iga g_m korral $\lim_n \langle f_n', g_m\rangle = \langle f_n', g_m\rangle$, saame seosest (4.5.12), et

$$\lim_n \langle f_n', g_n\rangle = \lim_m \lim_n \langle f_n', g_m\rangle = \langle f, g\rangle.$$

Seega on hulgal $\mathcal{H}$ korrektselt defineeritud kujutis $\langle \cdot, \cdot\rangle$, aksioomide vahetu kontroll näitab, et $\langle \cdot, \cdot\rangle$ on skalaarkorrutis. Seega $\mathcal{H}$ on skalaarkorrutisega ruum ja seega ka normeeritud ruum.

Paneme tähele, et hulk $\mathcal{F} \subset \mathcal{H}$ on kõikjal tihe. See tähendab, et iga $h \in \mathcal{H}$ korral leidub $f_n \in \mathcal{F}$ nii, et $\|f_n - h\| \to 0$. Tõepoolest, olgu $h \in \mathcal{H}$ ja $\{f_n\}$ seda defineeriv Cauchy jada. Seosest (4.5.12) saame aga, et

$$\|f_n - h\| = \langle f_n - h, f_n - h\rangle = \langle f_n, f_n\rangle - 2\langle h, f_n\rangle + \langle h, h\rangle \to 0. \tag{4.5.13}$$

Veendume nüüd, et $\mathcal{H}$ on täielik. Olgu $\{h_n\}$ hulga $\mathcal{H}$ elementidest moodustatud Cauchy jada. Et $\mathcal{F}$ on kõikjal tihe, leidub iga n korral jada $f_n \in \mathcal{H}$ nii, et $\|h_n - f_n\| < \frac{1}{n}$. On lihtne veenduda, et $\{f_n\}$ on Cauchy jada (sest $\{h_n\}$ on Cauchy jada). Seega leidub

h nii, et $f_n(x) \to h(x)$ iga x korral. Seosest (4.5.13) saame, et $\|f_n - h\| \to 0$, milest $\|h_n - h\| \le \|h_n - f_n\| + \|f_n - h\| \to 0$.

Omadus (4.5.8). Lõpuks paneme tähele, et ruumis $\mathcal{H}$ defineeritud skalaarkorrutis säilitab omadust (4.5.8). Kuulugu $h \in \mathcal{H}$. Seega leidub Cauchy jada $f_n \in \mathcal{F}$ nii, et $f_n(x) \to h(x)$. Vastavalt skalaarkorrutise definitsioonile $\langle h, K(x,\cdot)\rangle = \lim_n \langle f_n, K(x,\cdot)\rangle$. Et aga $\langle f_n, K(x,\cdot)\rangle = f_n(x)$, saame, et $\langle h, K(x,\cdot)\rangle = \lim_n f_n(x) = h(x)$. ■

Reprodutseeriva tuumaga Hilberti ruum. Ülaltoodud tõestuses defineeritud Hilberti ruumil $\mathcal{H}$ on omadus (4.5.8), st iga $f \in \mathcal{H}$ korral

$$\langle K(x,\cdot), f\rangle = f(x).$$

Selle kohta öeldakse, et tuumal K on **reprodutseeriv** omadus.

Definitsioon 4.3 *Olgu $\mathcal{H}$ Hilberti ruum, mille elmendid on funktsioonid hulgal $\mathcal{X}$, st $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. Funktsiooni $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ nimetatakse ruumi $\mathcal{H}$* **reprodutseerivaks tuumaks** *(*reproducing kernel*), kui*

i) *iga $x \in \mathcal{X}$ korral $K(x,\cdot) \in \mathcal{H}$;*

ii) *funktsioonil K reprodutseeriv omadus, st iga $x \in \mathcal{X}$ korral $\langle K(x,\cdot), f\rangle = f(x)$.*

Hilberti ruumi $\mathcal{H}$ nimetatakse sellisel juhul **RKHS ruumiks** *(*Reproducing Kernel Hilbert Space*).*

Iga reprodutseeriv tuum on tuum definitisiooni (4.5.1) mõttes, sest võttes nn **kanoonililise kujutise**

$$\Phi(x) = K(x,\cdot)$$

saame tuuma reprodutseerivast omadusest, et

$$K(y,x) = \langle K(x,\cdot), K(y,\cdot)\rangle = \langle \Phi(x), \Phi(y)\rangle = \langle \Phi(y), \Phi(x)\rangle.$$

Järgnev ülesanne näitab, et reprodutseeriv tuum on ühene.

Ülesanne 4.15 *Olgu K ja K' Hilberti ruumi $\mathcal{H}$ kaks reprodutseerivat tuuma. Näita, et $K = K'$, st iga $x, y \in \mathcal{X}$ korral $K(x,y) = K'(x,y)$.*

Saab näidata ka vastupidist: igale tuumale (määramispiirkonnaga $\mathcal{X}\times\mathcal{X}$) vastab ainult üks RKHS mille reprodutseeriv tuum on K ([13], Thm 4.21). Et Moore-Aronszajni teoreemi tõestuses ühe sellise ruumi konstrueerisime, siis see ongi see ainus RKHS. Pane tähele: ülaltoodust ei järeldu, et iga tuuma korral oleks ainult üks sakalaarkorrutisega ruum H_o ja kujutis $\Phi_o : \mathcal{X} \to H_o$ nii, et $K(x,y) = \langle \Phi_o(x), \Phi_o(y)\rangle$. Selliseid kujutisi võib olla (ja enamasti ongi) mitu, kuid vaid üks neist on RKHS. Veel enam, saab näidata ([13], Thm 4.21), et iga sellise kujutise Φ_o korral vastava RKHS saab defineerida järgmiselt:

$$\mathcal{H} := \{h : \mathcal{X} \to \mathbb{R} : \exists w \in H_o : h(x) = \langle w, \Phi_o(x)\rangle, \quad \forall x \in \mathcal{X}\}. \tag{4.5.14}$$

Veendu, et funktsioonid kujul $\sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$ kuuluvad hulka $\mathcal{H}$. Seosest (4.5.14) järeldub samuti, et kui H_o on m-dimensionaalne ruum, siis ka vastav RKHS on ülimalt m-dimensionaalne. [Tõepoolest, kui leiduvad $w_1, \ldots, w_m$ nii, et iga $w \in H_0$ esitub summana $w = \sum_i \alpha_i w_i$, siis iga $h \in \mathcal{H}$ esitub summana $h = \sum_i \alpha_i h_i$, kus $h_i = \langle w_i, \Phi_o \rangle$]. Samuti järeldub, et mistahes (tuumale vastava) Φ korral (olgu see siis mingi Φ_o või ka kujutis $K(x, \cdot)$) funktsioon $\langle w, \Phi(x) \rangle$ on kujul $h(x)$, kus h on mingi RKHS element. Seega klassifikaatori $g(x) = \mathrm{sgn}(\langle w, \Phi(x) \rangle + w_o)$ võib alati esitada kujul $g(x) = \mathrm{sgn}(h(x) + w_o)$, kus h on mingi RKHS element. Veel enam, saab näidata, et seosega (4.5.14) defineeritud RKHS korral kehtib:

$$\|h\| = \inf\{\|w\|_{H_o} : w \in H_o \text{ nii, et } h = \langle w, \Phi_o \rangle.\}$$

Seega, kui meil on SVM klassifitseerimisülesanne

$$\min_{w, w_o} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \big(1 - y_i(\langle w, \Phi_o(x_i) \rangle + w_o)\big)_+^p, \tag{4.5.15}$$

kus Φ_o ei ole ilmtingimata kanooniline kujutis RKHS'i, siis ülesande (4.5.15) võib alati esitada ekvivalentsel kujul:

$$\min_{h \in \mathcal{H}, w_o} \frac{1}{2}\|h\|^2 + C \sum_{i=1}^{n} \big(1 - y_i(h(x_i) + w_o)\big)_+^p. \tag{4.5.16}$$

Siin $\mathcal{H}$ on RKHS ja ekvivalents tähendab, et igale ülesande (4.5.15) abil saadud klassifitseerijale vastab (4.5.16) abil saadud klassifitseerija ja vastupidi. Seega edaspidi vaatleme SVM klassifitseerimisülesandeid vaid kujul (4.5.16), kus $\mathcal{H}$ on RKHS.

Näide. Vaatleme eelmises näites defineeritud kujutist

$$\Phi_o : \mathbb{R}^2 \to \mathbb{R}^3, \quad \Phi(x) = \Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2). \tag{4.5.17}$$

Siin H_o on $\mathbb{R}^3$ ja vastavalt seosele (4.5.14),

$$\mathcal{H} = \{h : \mathbb{R}^2 \to \mathbb{R}, h(x) = w_1x_1^2 + w_2\sqrt{2}x_1x_2 + w_3x_2^2\}.$$

Seega RKHS moodustavad funktsioonid kujul

$$h(x) = ax_1^2 + bx_1x_2 + cx_2^2, \quad (a, b, c)' \in \mathbb{R}^3.$$

On selge, et ruumi H_o ja $\mathcal{H}$ vahel on üksühene vastavus: igale vektorile $(w_1, w_2, w_3)' \in H_0$ vastab üks funktsioon $h(x) = w_1x_1^2 + w_2\sqrt{2}x_1x_2 + w_3x_2^2$ ja igale funktsioonile $h(x) = ax_1^2 + bx_1x_2 + cx_2^2$ vastab üks vektor $(w_1, w_2, w_3)'$, kus $w_1 = a$, $w_2 = b/\sqrt{2}$ ja $w_3 = c$. Et seos on üksühene, saame $\|h\|_{\mathcal{H}} = \|w\|_{H_o}$ ning $\langle h_1, h_2 \rangle_{\mathcal{H}} = \langle w_1, w_2 \rangle_{H_o}$ (sest muidu ei kehtiks $\|h\|_{\mathcal{H}} = \|w\|_{H_o}$). Nii on H_o ja $\mathcal{H}$ vahel isomeetriline isomorfism. Järelikult, kui

$$h_1(x) = a_1x_1^2 + b_1x_1x_2 + c_1x_2^2, \quad h_2(x) = a_2x_1^2 + b_2x_1x_2 + c_2x_2^2,$$

siis

$$\langle h_1, h_2\rangle_{\mathcal{H}} = a_1a_2 + \frac{b_1b_2}{2} + c_1c_2.$$

Nüüd $K(x,\cdot)$ on seega funktsioon (polünoom), mille kordajad (a,b,c) on $(x_1^2, 2x_1x_2, x_2^2)$; võttes $h(x) = ax_1^2 + bx_1x_2 + cx_2^2$, saame

$$\begin{aligned}\langle K(x,\cdot), h\rangle_{\mathcal{H}} &= \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2)', (w_1,w_2,w_3)'\rangle_{H_o} \\ &= \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2)', (a, \frac{b}{\sqrt{2}}, c)'\rangle_{H_o} = ax_1^2 + bx_1x_2 + cx_2^2 = h(x).\end{aligned}$$

Tuletame aga meelde, et sama tuuma defineerib ka kujutis

$$\Phi_o : \mathbb{R}^2 \to \mathbb{R}^4, \quad \Phi(x) = \Phi(x_1,x_2) = (x_1^2, x_1x_2, x_1x_2x_2^2).$$

Vastaval seosele (4.5.14),

$$\mathcal{H} = \{h : \mathbb{R}^2 \to \mathbb{R}, h(x) = w_1x_1^2 + w_2x_1x_2 + w_3x_1x_2 + w_4x_2^2 = w_1x_1^2 + (w_2+w_3)x_1x_2 + w_4x_2^2\}.$$

Seega ruum $\mathcal{H}$ koosneb jälle funktsioonidest kujul $a_1^2 + bx_1x_2 + cx_2^2$. Näeme, et $\mathcal{H}$ on ikka üks ja seesama hulk. Aga praegu pole enam vektorite (w_1,w_2,w_3,w_4) ja funktsioonide h vahel üksühest vastavust.

Lõpetuseks paneme tähele, et kui

$$\Phi_o : \mathbb{R}^2 \to \mathbb{R}^3, \quad \Phi(x) = \Phi(x_1,x_2) = (x_1^2, x_1x_2, x_2^2), \tag{4.5.18}$$

siis ruum $\mathcal{H}$ on ikka funktsioonid kujul $a_2x_1^2 + b_2x_1x_2 + c_2x_2^2$ aga Hilberti ruum (RKHS) on teine, sest skalaarkorrutis ja seega ka norm on teised. Tõepoolest, nüüd on funktsioonide

$$h_1(x) = a_1x_1^2 + b_1x_1x_2 + c_1x_2^2, \quad h_2(x) = a_2x_1^2 + b_2x_1x_2 + c_2x_2^2$$

skalaarkorrutis

$$\langle h_1, h_2\rangle_{\mathcal{H}} = a_1a_2 + b_1b_2 + c_1c_2$$

ja seega kui $h(x) = ax_1^2 + bx_1x_2 + cx_2^2$, siis $\|h\|_{\mathcal{H}}^2 = a^2 + b^2 + c^2$. Kokkuvõttes: kujutised (4.5.17) ja (4.5.18) on erinevad ja erinevad on ka vastavad tuumad. Järelikult on erinevad ka vastavad RKHS: hulk $\mathcal{H}$ on küll sama, aga norm ja skalaarkorrutis erinevad.

Alternatiivne definitsioon. Olgu $\mathcal{H}$ RKHS reprodutseeriva tuumaga K ning olgu $h_n, h \in \mathcal{H}$ koonduv jada: $\|h_n - h\| \to 0$. Siis tuuma K reprodutseerivast omadusest ning Cauchy-Schwartzi võrratusest järeldub, et iga x korral

$$|h_n(x) - h(x)| = |\langle K(x,\cdot), (h_n - h)\rangle| \le \|K(x,\cdot)\|\|h_n - h\| \to 0.$$

Teisisõnu, iga x korral on kujutis

$$\delta_x : \mathcal{H} \to \mathbb{R}, \quad \delta_x(h) = h(x)$$

pidev, sest argumentide koondumisest järeldub väärtuste koondumine. Selgub, et see omadus on ka tarvilik selleks et Hilberti ruum $\mathcal{H}$, (mille elemendid on funktsioonid) oleks RKHS. Selles veenab meid järgmine ülesanne.

Ülesanne 4.16 *Rieszi teoreemi kasutades tõesta, et Hilberti ruum $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ on RKHS parajasti siis, kui iga x korral δ_x on pidev funktsionaal. Kas järgmised Hilberti ruumid on RKHS: $L_2[0,1]$? l_2? $\mathbb{R}^d$? Kui ja, siis leia reprodutseeriv tuum.*

Enamasti defineeritaksegi RKHS ülaltoodud omaduse kaudu: Hilberti ruum $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ on RKHS kui iga x korral funktsioon δ_x in pidev (ekvivalentselt: tõkestatud, sest lineaarne funktsioon oon pidev parajasti siis, kui ta on tõkestatud).

RKHS omadustest. RKHS (või sinna kuuluvate funktsioonide) omadused sõltuvad hulga $\mathcal{X}$ ja tuuma K omadustest. Kui $\mathcal{X}$ on loenduv, siis RKHS on separaabel Hilberti ruum; samuti on RKHS separaabel, kui $\mathcal{X}$ on separaabel meetriline ruum ja K pidev ([13], Thm 4.33). RKHS omadustest täpsemalt loe ([13], Ch 4.3).

Ülesanne 4.17 *Olgu K selline, et $\|K\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)} < \infty$. Tõesta, et iga x, y korral $K(x,y) \le \|K\|_\infty^2$. Tõesta, et kõik RKHS elemendid on tõkestatud funktsioonid.*

Ülesanne 4.18 *Olgu $\mathcal{X}$ meetriline ruum ja $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ RKHS. Tõestada, et kui reprodutseeriv tuum K on pidev, siis koosneb $\mathcal{H}$ ka pidevatest funktsioonidest.*

Universaalne tuum. Olgu $\mathcal{X}$ *kompaktne* meetriline ruum ja tuum K pidev. Sellisel juhul nimetatakse tuuma **universaalseks**, kui vastav RKHS $\mathcal{H}$ on kõikjla tihe ruumis $C(\mathcal{X})$, st iga $\epsilon > 0$ ja $g \in \mathcal{X}$ korral leidub $h \in \mathcal{H}$ nii, et

$$\|h - g\|_\infty := \sup_{x\in\mathcal{X}} |h(x) - g(x)| \le \epsilon.$$

On kerge näha, et kui K on universaalne, siis iga kahe lõikumatu kompaktse hulga A ja B korral leidub funktsioon $f \in \mathcal{H}$ (RKHS), nii, et $h(x) > 0$ iga $x \in A$ ja $h(x) < 0$ iga $x \in B$ korral ([13], Prop 4.54). Sellel omadusel on oluline tähendus klassifitseerimisel. Tõepoolest, olgu $x_1, \dots, x_n$ mingi n-elendiline hulk ja defineerime $\mathcal{X} := \{x_1, \dots, x_n\}$. Mistahes lõplik hulk on alati kompaktne meetriline ruum diskreetse meetrika suhtes (diskreetne meetrika: $d(y,x) = 0$ parajasti siis, kui $x = y$, mujal $d(x,y) = 1$). Olgu $y_1, \dots, y_n$, $y_i \in \{-1, 1\}$ klassid. Defineerime lõikumatud hulgad $A = \{x_i \in \mathcal{X} : y_i = 1\}$ ja $B = \{x_i \in \mathcal{X} : y_i = -1\}$. Kui K on universaalne, siis leidub funktsion h nii, et $y_i h(x_i) > 0$ iga i korral. Seose (4.5.14) tõttu, mistahes Φ_o korral leidub $w \in \mathcal{H}_o$ nii, et klassifitseerija $g(x) = \operatorname{sgn}\langle w, \Phi_o(x)\rangle$ eraldab iga valimi $(x_1, y_1), \dots, (x_n, y_n)$ (eeldades, et $x_i \neq x_j$).
On lihtne veenduda, et kui $\mathcal{X}$ on lõpmatu ja sellel antud tuum universaalne, siis tal on järgmine omadus: suvalise lõpliku (kordusteta) alamhulga $\{x_1, \dots, x_n\}$ korral on Grami maatriks $K(x_i, x_j)$ täisastakuga ja sellest loomulikult järeldub, et vastav RKHS peab olema lõpmatudimensionaalne.

Kokkuvõttes: Iga tuuma K (positiivselt poolmääratud funktsiooni) korral leidub Hilberti ruum

$$\mathcal{H} = \overline{\operatorname{span}\{K(x,\cdot) : x \in \mathcal{X}\}}$$

nii, et K on selle ruumi reprodutseeriv tuum ja seega $\mathcal{H}$ on RKHS. Selle ruumi iga element $h \in \mathcal{H}$ on funktsioon ja reprodutseeriv omadus (4.5.8) on

$$\langle K(x,\cdot), h\rangle = h(x).$$

Sellisel juhul iga $x \in \mathcal{X}$ korral $K(x,\cdot) \in \mathcal{H}$ ja $K(x,y) = \langle K(x,\cdot), K(y,\cdot)\rangle$. Seega iga tuuma korral me võime alati võtta kujutiseks Φ kanoonilise kujutise $\Phi(x) = K(x,\cdot)$. Kuid mistahes teise kujutise $\Phi_o : \mathcal{X} \to \mathcal{H}_o$ korral nii, et $\langle \Phi_o(x), \Phi_o(y)\rangle = K(x,y)$, kehtib: iga $w \in \mathcal{H}_o$ korral leidub RKHS element h nii, et $\langle w, \Phi_o(x)\rangle = h(x)$ ja vastupidi – iga RKHS elemendi h korral leidub vähemalt üks $w' \in H_o$ nii, et $\langle w', \Phi_o(x)\rangle = h(x)$ $\forall x$, kusjuures elemendi h (RKHS-)norm $\|h\| = \|w'\|$, kui w' on ühene (vastasel korral $\|h\| = \min_{w'} \|w'|$).

Kirjandus: [13], Ch. 4.

4.5.3 Tuuma konstrueerimine ja omadused

Järgnevast lemmast järeldub, et teatud algebraliste tehete suhtes on tuum kinnine. See tulemus võimaldab veelgi kergemini tuuma ära tunda, samuti sobivat tuuma konstrueerida.

Lemma 4.1 *Olgu K_1 ja K_2 tuumad (positiivselt poolmääratud funktsioonid) hulgal $\mathcal{X}\times\mathcal{X}$. Siis järgmised funktsioonid on samuti tuumad*

1. $K(x,y) = K_1(x,y) + K_2(x,y)$
2. $K(x,y) = aK_1(x,y)$, $a \in \mathbb{R}^+$
3. $K(x,y) = g(x)g(y)$, $g : \mathcal{X} \to \mathbb{R}$
4. $K(x,y) = x'Ay$, *kus A on positiivselt poolmääratud maatriks*
5. $K(x,y) = K_1(x,y)K_2(x,y)$
6. $K(x,y) = p(K_1(x,y))$, *kus p on positiivsete kordajatega polünoom*
7. $K(x,y) = K_1(f(x), f(y)), \quad f : \mathcal{X} \to \mathcal{X}$
8. $K(x,y) = \exp[K_1(x,y)]$
9. $K(x,y) = \exp[-\frac{\|x-y\|^2}{2\sigma^2}]$, $x, y \in \mathbb{R}^d$.

Tõestus.

Ülesanne 4.19 *Tõestada 1, 2, 3, 4.*

Tõestame 5. Selleks näitame, et kahe positiivselt poolmääratud maatriksi komponentkaupa korrutis on positiivselt poolmääratud. Seda saab näidata otse nn Cholesky lahutust kasutades (kui A on positiivselt poolmääratud $n \times n$ maatriks, siis $A = BB'$, kus B on ka positiivselt poolmääratud $n \times n$ maatriks), kuid saab kasutada ka järgnevat arutelu: olgu $(V_1, \ldots, V_n)$ ja $(W_1, \ldots, W_n)$ kaks sõltumatut normaalse ühisjaotusega vektorit keskväärtusega 0 ja kovariatsioonimaatriksitega vastavalt K_1 ja K_2. Seega maatriks, mille komponendid on K_1 ja K_2 vastavate komponentide korrutis, on vektori $(V_1W_1, \ldots, V_nW_n)$ kovariatsioonimaatriks ning seega positiivselt poolmääratud.

Ülesanne 4.20 *Tõestada 6, 7.*

Tõestame 8. Selleks paneme tähele, et eksponentfunktsioon on mittenegatiivsete kordajatega polünoomide punktiviisiline piirväärtus (Taylori rittaarendus). Teisisõnu, iga x, y korral

$$\exp[K_1(x,y)] = \sum_{i=0}^{\infty} \frac{1}{i!} K_1(x,y)^i = \lim_n p_n(K_1(x,y)),$$

kus $p_n(x) = \sum_{i=0}^{n} \frac{1}{i!} x^i$. Teame, et $K_n(\cdot,\cdot) := p_n(K_1(\cdot,\cdot))$ on tuum (omadus 6). Seega leiduvad tuumad K_n nii, et iga x, y korral $\exp[K_1(x,y)] = \lim_n K_n(x,y)$. Väide on tõestatud, sest tuumade punktiviisiline piirväärtus on ka tuum.

Ülesanne 4.21 *Olgu K_n hulgal $\mathcal{X} \times \mathcal{X}$ antud tuumad nii, et iga (x,y) korral $K_n(x,y) \to K(x,y)$, kus K on sellel hulgal antud funktsioon. Tõesta, et K on tuum.*

Tõestame 9. Et $\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle$, siis

$$\exp[-\frac{\|x-y\|^2}{2\sigma^2}] = \exp[-\frac{\|x\|^2}{2\sigma^2}] \exp[-\frac{\|y\|^2}{2\sigma^2}] \exp[\frac{\langle x,y\rangle}{\sigma^2}] = K_1(x,y)K_2(x,y),$$

kus

$$K_1(x,y) = \exp[-\frac{\|x\|^2}{2\sigma^2}] \exp[-\frac{\|y\|^2}{2\sigma^2}], \quad K_2(x,y) = \exp[\frac{\langle x,y\rangle}{\sigma^2}].$$

Seosest 3 saame, et K_1 on tuum. Seosest 8 saame, et K_2 on tuum. Seosest 5 saame, et K on tuum. ■

Bochneri teoreem. Tuum

$$K(x,y) = \exp[-\frac{\|x-y\|^2}{2\sigma^2}] \tag{4.5.19}$$

on kujul $f(x-y)$, kus $f : \mathbb{R}^d \to \mathbb{R}$. Kas on võimalik iseloomustada funktsioone f nii, et $K(x,y) := f(x-y)$ oleks tuum? Selliseid funktsioone nimetatakse positiivselt poolmääratuteks. Järgnev teoreem annab vastuse.

Teoreem 4.4 (Bochner) *Pidev funktsioon $f : \mathbb{R}^d \to \mathbb{R}$ on positiivselt poolmääratud parajasti siis, kui ruumil $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ leidub lõplik mõõt μ nii, et*

$$f(x) = \int \exp[-ix'z]\mu(dz).$$

Seega on f positiivselt poolmääratud, kui ta on mingi lõpliku mõõdu Fourier teisendus. Paneme tähele, et kui $f(0) = 1$, siis μ on tõenäosusmõõt ja f on selle mõõdu karakteristlik funktsioon. Siit järeldub vahetult Lemma 4.1 väide 9, sest $f(x) = \exp[-\frac{\|x\|^2}{2\sigma^2}]$ on teadupoolest mitmemõõtmelise normaaljaotuse $N(0, \sigma^2 I_d)$ karakteristlik funktsioon.

Ülesanne 4.22 *Olgu $f : \mathbb{R}^d \to \mathbb{R}$ selline, et $f(0) = 0$, kuid $f \not\equiv 0$. Kas $f(x - y)$ on tuum?*

4.5.4 Näiteid tuumadest

Tuumad ruumil $\mathcal{X} = \mathbb{R}^d$

Polünoom-tuum. Olgu $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Lemmast 4.1 teame, et positiivsete kordajatega polünoomi p korral on $(x, y) \mapsto p(x'y)$ tuum. Seega iga $R \geq 0$ korral

$$K(x, y) := \big(x'y + R\big)^p = \sum_{s=0}^{p} \binom{p}{s} R^{p-s} (x'y)^s = \sum_{s=0}^{p} a_s (x'y)^s, \quad a_s := \binom{p}{s} R^{p-s} \tag{4.5.20}$$

on tuum, sest $a_s = \binom{p}{s} R^{p-s} \geq 0$. Ülaltoodud summast järeldub, et $(x'y)^s$ suhteline osakaal a_s sõltub R-st: mida suurem on R, seda väiksem osakaal on suurtel astmetel. Et

$$(x'y)^s = (x_1y_1 + \cdots + x_dy_d)^s = \sum_{(i_1,\ldots,i_d)\in\{0,1,\ldots,s\}^d : \sum_{i_j} = s} c(i_1, \ldots, i_d)(x_1y_1)^{i_1}(x_2y_2)^{i_2} \cdots (x_dy_d)^{i_d},$$

kus $c(i_1, \ldots, i_d)$ on kordajad. Seega saame,

$$\Phi(x) = \begin{pmatrix} \cdots \\ a(i_1, \ldots, i_d) x_1^{i_1} x_2^{i_2} \cdots x_d^{i_d} \\ \cdots \end{pmatrix}, \tag{4.5.21}$$

kus $a(i_1, \ldots, i_d)$ on kordajad ning

$$i_1, \ldots, i_d \in \{0, \ldots, p\} : \sum_{i=1}^{d} i_j \leq p.$$

Selliseid indekseid on $\binom{d+p}{p}$ (vt [14],Prop. 9.2), mistõttu $\mathcal{H}$ on $\binom{d+p}{p}$-dimensionaalne ruum. Klassifitseerimine nii suure dimensiooniga ruumis on keeruline, kuid reegel

$$\langle \Phi(x), \Phi(y) \rangle = (x'y + R)^p$$

aitab. Polünoom-tuuma abil saadud klassifitseerija (4.5.2) avaldub

$$\operatorname{sign}\Big(\sum_{i\in SV} \alpha_i^* y_i K(x_i, x) + w_o^*\Big) = \operatorname{sign}\Big(\sum_{i\in SV} \alpha_i^* y_i (x_i'x + R)^p + w_o^*\Big),$$

st sisuliselt klassifitseeritakse sellisel juhul järgmise p-järku joone abil:

$$\Big\{x : \sum_{i\in SV} \alpha_i^* y_i (x_i'x + R)^p = -w_o^*\Big\}.$$

Alamhulk-tuum. Olgu $A \subset \{1, \ldots, d\}$. Defineerime

$$\Phi_A(x_1, \cdots, x_d) := \prod_{j \in A} x_j, \quad \Phi_\emptyset = 1.$$

Alamhulk-tuuma korral on

$$\Phi(x) = \Big(\Phi_A(x)\Big)_{A \subseteq \{1,\ldots,d\}}.$$

Seega Φ koosneb x komponentide kõivõimalikest korrutistest, kusjuures kõik komponendid on vaid esimeses astmes. Sellisel juhul $\mathcal{H}$ on 2^d-dimensionaalne ruum. Tuum defineeritkse arusaadavalt

$$K_\subseteq(x, y) := \langle \Phi(x), \Phi(y) \rangle = \sum_{A \subseteq \{1,\ldots,d\}} \prod_{j \in A} x_j y_j = \prod_{j=1}^{d} (1 + x_j y_j).$$

ANOVA tuum. ANOVA tuum K_p ($p \leq d$) erineb alamhulk-tuumast vaid selle poolest, et Φ konstrueerimisel võetakse arvesse vaid selliseid almhulki, mille võimsus on p. Seega,

$$\Phi_p(x) = \Big(\Phi_A(x)\Big)_{A \subseteq \{1,\ldots,d\}:|A|=p},$$

millest $\mathcal{H}$ on $\binom{d}{p}$-dimensionaalne ruum. Paneme veel tähele, et Φ_p kujutab iga tunnusvektori polünoom-tuumale vastava $\mathcal{H}$ alamruumi, sest kõik korrutised $x_{j_1} \cdots x_{j_p}$, $j_1 < \cdots < j_p$ figureerivad ka vektoris (4.5.21). Veenduge, et see pole ilmtingimata nii $\Phi_\subseteq$ korral (kui $d > p$).
ANOVA tuum defineeritakse järelikult

$$K_p(x, y) := \langle \Phi_p(x), \Phi_p(y) \rangle = \sum_{A:|A|=p} \prod_{j \in A} x_j y_j.$$

Tuuma on võimalik arvutada rekursiivselt ülimalt $3(dp + \frac{p(p-1)}{2})$ tehte abil ([14], ptk. 9)

Gaussi tuum. Gaussi tuum ehk *RBF tuum* on kujul (4.5.19), st

$$K(x, y) = \exp[-\frac{\|x - y\|^2}{2\sigma^2}].$$

Klassifitseerija (4.5.2) avaldub

$$\operatorname{sign}\big(\sum_{i \in SV} \alpha_i^* y_i \exp[-\frac{\|x_i - x\|^2}{2\sigma^2}] + w_o^*\big),$$

st sisuliselt klassifitseeritakse sellisel juhul järgmise joone abil:

$$\Big\{x : \sum_{i \in SV} \alpha_i^* y_i \exp[-\frac{\|x_i - x\|^2}{2\sigma^2}] = -w_o^*\Big\}$$

Ülesanne 4.23 *Veendu, et $\Phi : \mathcal{X} \mapsto S_{\mathcal{H}}$, kus $S_{\mathcal{H}}$ on ühiksfäär ruumis $\mathcal{H}$.*

Kui $\mathcal{X} \subset \mathbb{R}^d$ on tõkestatud ja kinnine (st kompaktne), siis Gaussi tuum on universaalne ([13], Cor. 4.58), ta eraldab iga (kordusteta) valimi. Teisisõnu, Gaussi tuum tekitab esialgses ruumis nii suure klassifikaatorite klassi, et suvaline valim on eralduv (VC dimensioon on lõpmatu). See on ekvivalentne sellega, et iga $x_1, \ldots, x_n$ korral on Grami maatriks täisastakuga (vt ka [15], Thm 2.18)

Polünoom-tuum ja Gaussi tuum on vast kõige enam levinud tuumad ruumis $\mathbb{R}^d$. Polünoom-tuum (ega ka teised sarnased tuumad: ANOVA-tuum ja alamhulk tuum) pole universaalsed, nende kaudu defineeritud klassifitseerijate VC-dimensioon on lõplik. Märgime veel, et tuntud tuumadest ruumis $\mathbb{R}^d$ on universaalne *eksponentsiaalne* tuum

$$K(x, y) = \exp[x'y].$$

Tuumad hulkadel: $\mathcal{X} = 2^S$

Ühisosa tuum (*intersection kernel*). Olgu S lõplik hulk,

$$\Phi : 2^S \to \{0,1\}^{2^S}, \quad \Phi_U(A) = \begin{cases} 1, & \text{kui } U \subseteq A\ ; \\ 0, & \text{muidu.} \end{cases}$$

Seega igale hulga S alamhulgale U vastab vektori $\Phi(A)$ üks komponent, see komponent on 1 parajasti siis, kui A sisaldab hulka U.

Ülesanne 4.24 *Veendu, et*

$$K(A, B) := \langle \Phi(A), \Phi(B) \rangle_{l_2} = 2^{|A \cap B|}.$$

Teine võimalus ühisosa tuuma defineerimiseks on isegi lihtsam

$$K_\cap(A, B) = |A \cap B|, \tag{4.5.22}$$

Ülesanne 4.25 *Veendu, et sellisel juhul*

$$\Phi_\cap : 2^S \to \{0,1\}^{|S|}, \quad \Phi_\cap(A) = (I_A(s))_{s \in S}.$$

Juhul kui S on lõpmatu (kuid mitte ainult siis), üldistub tuum (4.5.22) lõpliku mõõdu kaudu. Olgu (S, Σ, μ) lõpliku mõõduga ruum. Võttes $\mathcal{X} = \Sigma$, defineerime

$$K_\cap : \Sigma \times \Sigma \to \mathbb{R}^+, \quad K_\cap(A, B) = \mu(A \cap B). \tag{4.5.23}$$

Sellele tuumale vastav Φ on järgmine

$$\Phi_\cap : \Sigma \to L_2(S, \Sigma, \mu), \quad \Phi_\cap(A) = I_A,$$

sest

$$\langle \Phi_\cap(A), \Phi_\cap(B) \rangle_{L_2} = \int_S I_A I_B(s) \mu(ds) = \int_{A \cap B} \mu(ds) = \mu(A \cap B).$$

Ülesanne 4.26 *Olgu* (S, Σ, μ) *tõenäosusruum, st* $\mu(S) = 1$.

- *Tõestada, et* union complement kernel

$$K(A, B) = 1 - \mu(A \cup B)$$

on tuum. Leida sobiv Φ.

- *Tõestada, et*

$$K(A, B) = 1 - \mu(A \Delta B)$$

on tuum.

Tekstide kategoriseerimine

Järgmiseks näited tuumadest tekstide (dokumentide) klassifitseerimisel (tekstide klassifitseerimist nimetatakse tihti ka nende **kategoriseerimiseks** (*categorization*)).

Ühisosa tuum tekstidel. Esimene näide on väga lihtne – teksti vaadeldakse kui (lõplikku) sõnade hulka ning klassifitseerimine toimub ühisosa tuuma abil. Seega mõõdetakse tekstide sarnasust kokkulangevate sõnade arvu abil. Sõnadele võib lisada ka kaalu (olulised sõnad või märksõnad suurema kaaluga, sidesõnad jne väiksema kaaluga ehk üldsegi mitte) ja nii saadud tuum on kujul (4.5.23).

Sõnadekott. Ülaltoodud näites vaadeldi teksti vaid kui sõnade hulka, arvesse ei võetud aga sõnade kordusi, järjestust, semantikat ega midagi muud. Järgnev näide on samm edasi – tekstide omavahelise sarnasuse defineerimisel võetakse arvesse ka sõnade kordusi. Sisuliselt vaadeldakse teksti nüüd kui sõnade hulka, kus mõni sõna võib olla mitmekordselt. Selline teksti esitusviis on tuntud kui **sõnadekott** (*bag of words*) või ka VSM (*vector space model*). Olgu $W = \{w_1, \ldots, w_m\}$ lõplik sõnastik. Sõnastik võib olla eelnevalt defineeritud, praktikas defineeritakse see tihti kui kõikide treening-dokumentide sõnade hulk. Olgu $T \in W^*$ tekst ja defineerime

$$\Phi(T) = \big(\Phi_{w_1}(T), \ldots, \Phi_{w_s}(T)\big),$$

kus $\Phi_w(T)$ on sõna w sagedus tekstis. Seega iga tekst esitub väga pika vektorina, kusjuures enamus selle vektori komponentidest on nullid. Tuum defineeritakse arusaadavalt

$$K(T_1, T_2) = \langle \Phi(T_1), \Phi(T_2) \rangle_{l_2} = \sum_w \Phi_w(T_1)\Phi_w(T_2) = \Phi'(T_1)\Phi(T_2).$$

Kuigi vektorid on väga pikad, on selle tuuma arvutamine mitte eriti töömahukas, sest enamus vektori komponentidest on nullid. Praktikas kasutatakse selleks nn *tokenisation*-protsessi, mille käigus tekst konverteeritakse kujule, mis koosneb vaid tekstis olevatest sõnadest ja nende sagedustest. Siis leitakse $K(T_1, T_2)$, selleks kuluv aeg on proportsionaalne mõlema dokumendi pikkuse summaga.

Loomulikult võib erinevatele sõnadele anda erineva kaalu. Näiteks sidesõnu mis niikuinii figureerivad igas dokumendis ei pruugigi arvesse võtta, nende kaal võib olla 0 (kaaluga 0 sõnu nimetatakse stop-sõnadeks). Kui sõna w kaal on $\mu(w)$, saame

$$\Phi(T) = \big(\Phi_{w_1}(T)\mu(w_1), \ldots, \Phi_{w_s}(T)\mu(w_s)\big),$$

tuum on siis

$$K(T_1, T_2) = \sum_w \mu^2(w)\Phi_w(T_1)\Phi_w(T_2) = \Phi(T_1)'\mathbf{R}\Phi(T_1),$$

kus $\mathbf{R}$ on diagonaalmaatriks, diagonaalelementidega $\mu^2(w)$. Levinud kaalud on nn idf-kaalud (*inverse document frequency*). Olgu treeningkorpuses n teksti ja olgu $df(w)$ nende tekstide arv, mis sisaldavad sõna w; idf-kaal defineeritakse

$$\mu(w) := \ln(\frac{n}{df(w)}).$$

Seega $\mu(w) = 0$ parajasti siis, kui sõna w on treeningkorpuse igas tekstis.

Idf-kaalud elimineerivad küll ebaolulised sõnad ja panevad suurema kaalu haruldastele sõnadele, kuid nad ei võta arvesse sünonüüme, sarnase tähendusega sõnu jne. Selleks defineeritakse nn sarnasusmaatriks *proximity matrix* $\mathbf{P} = \big(P(w, w')\big)_{w,w' \in W}$, kus diagonaalil on kaalud $P(w, w) = \mu(w)$ või 1, kuid $P(w, w') > 0$, kui sõnad w ja w' on tähenduselt sarnased. Vektori Φ asemel vaadeldakse nüüd vektorit $\Phi^* = \mathbf{P}\Phi$. Vektoris Φ^* on vähem nulle: tõepoolest, kui $P(w, w') > 0$, $\Phi(w) > 0$ kuid $\Phi(w') = 0$, siis vektoris Φ^* on mõlemale sõnale vastav komponent nullist erinev, st $\Phi^*(w) > 0$ ja $\Phi^*(w') > 0$. Tuum on nüüd

$$K(T_1, T_2) = \sum_w \Phi^*_w(T_1)\Phi^*_w(T_2) = \Phi'(T_1)\mathbf{P}'\mathbf{P}\Phi(T_2) = \Phi'(T_1)\mathbf{Q}\Phi(T_2),$$

kus $\mathbf{Q} = \mathbf{P}'\mathbf{P}$.

Kirjandus: Nendest ja teistest tekstide analüüsiks sobivatest tuumadest lugege raamatust [14] (ptk.10).

Tuumad sõnadel

Järgnevas toome paar näidet tuumadest sõnade hulgal (*strings*). Olgu Σ lõplik tähestik (näiteks $\{A, T, G, C\}$, 20 aminohapet või harilik tähestik) ning $\mathcal{X} = \Sigma^* = \cup_n \Sigma^n$ kõikide lõplike sõnade hulk. Sõna u on sõna v alamjada, kui leiduvad indeksid $1 \le i_1 < \cdots < i_{|u|} =: \mathbf{i}$ nii, et $u_j = v_{i_j}$ iga $j = 1, \ldots, |u|$ korral. Seda kirjutame $u = v(\mathbf{i})$. Seega u võib olla v alamjada ka siis, kui u ei ole v alamsõna, kuid u tähed on (õiges järjekorras) sõnas v olemas. Näiteks sõna *tln* on sõna *tallinn* alamjada, kuid mitte alamsõna. Sõna *tal* on aga nii alamjada kui ka -sõna.

p-spekter tuum. Selle tuuma korral defineeritakse (s on sõna)

$$\Phi^p(s) = \left(\Phi^p_u(s)\right)_{u\in\Sigma^p},$$

kus $\Phi^p_u(s)$ loendab, mitmes kohas asub alamsõna u. Seega $\Phi^p(s)$ on $|\Sigma|^p$-pikkune vektor, millest enamik komponente on nullid. See vektor on nn. p-spektrum. Tuum defineeritakse nüüd

$$K_p(s,t) = \sum_{u\in\Sigma^p} \Phi^p_u(s)\Phi^p_u(t)$$

ja see on seda suurem, mida rohkem on kokkulangevaid alamsõnu. Seda tuuma saab arvutada $O(\max\{|s|,|t|\})$ arvutusega, kus $|s|$ on sõna s pikkus.
Näiteks sõnade "bar", "baa", "car", "cat"2-spektrum on järgmine (kõik ülejäänud komponentid on nullid):

Φ^2	ar	at	ba	ca
bar	1	0	1	0
bat	0	1	1	0
car	1	0	0	1
cat	0	1	0	1

Tuum on seega

K	bar	bat	car	cat
bar	2	1	1	0
bat	1	2	0	1
car	1	0	2	1
cat	0	1	1	2

Pikkusega p alamjadade tuum. See tuum on sarnane eelmisega, kuid pikkusega p alamsõnade asemel loetakse pikkusega p alamjadu. Seega vektor Φ on sama pikk, kuid seal on palju vähem nulle. Saab arvutada $O(|s||t|)$ keerukusega.

Kõikide alamjadade tuum. Kujutis Φ defineeritakse järgmiselt

$$\Phi(s) = \left(\Phi_u(s)\right)_{u\in\Sigma^*},$$

kus $\Phi_u(s)$ näitab, mitu korda on sõna u sõna s alamjada. Näiteks sõna *tln* on sõna *tallinn* alamjada 4 korda. Näiteks sõnade "bar", "baa", "car", "cat"korral oleks vektor Φ järgmine, ülejäänud komponendid on 0-d.

ϕ	$\emptyset$	a	b	c	r	t	aa	ar	at	ba	br	bt	ca	cr	ct	bar	baa	car	cat
bar	1	1	1	0	1	0	0	1	0	1	1	0	0	0	0	1	0	0	0
baa	1	2	1	0	0	0	1	0	0	2	0	0	0	0	0	0	1	0	0
car	1	1	0	1	1	0	0	1	0	0	0	0	0	1	1	0	0	0	1
cat	1	1	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0	1

Tuum defineeritakse, nagu ikka,

$$K(s,t) = \sum_{u \in \Sigma^*} \Phi_u(s)\Phi_u(t).$$

Antud näite korral

K	bar	baa	car	cat
bar	8	6	4	2
baa	6	12	3	3
car	4	3	8	4
cat	2	3	4	8

Keerukus: $O(|s||t|)$.

Kaalutud alamjadade tuum. Pikkusega p alamjadade tuum arvestas iga alamjada ühe ja sama kaaluga, arvestamata seda, kui "väljavenitatud"ta on. Kaalutud alamjadade tuuma korral võetakse neid alamjadu, mille esimese ja viimase tähe vahe on suurem, väiksema kaaluga arvesse. Seega alamsõnad on kõige suurema kaaluga ning "väljavenitatud"sõnad on väiksema kaaluga. Sõna s alamjada

$$u = u_{i_1}, \ldots, u_{i_{|u|}} =: s(\mathbf{i}) \tag{4.5.24}$$

korral defineerime $l(\mathbf{i}) := i_{|u|} - i_1 + 1$. Näiteks sõna *antarktika* alamjada *tai* korral $i_1 = 3, i_2 = 4, i_3 = 8$, millest $\mathbf{i} = (3,4,8)$ ja $l(\mathbf{i}) = 6$. Alamjada u kaal on $\lambda^{l(\mathbf{i})}$, kus $\lambda \in (0,1)$. Kujutis Φ^p defineeritakse nüüd $\Phi^p(s) = \left(\Phi^p_u(s)\right)_{u \in \Sigma^p}$, kus

$$\Phi^p_u(s) = \sum_{\mathbf{i}: u = s(\mathbf{i})} \lambda^{l(\mathbf{i})}.$$

Tuum:

$$K(s,t) := \sum_{u \in \Sigma^p} \Phi^p_u(s)\Phi^p_u(t).$$

Juhul, kui $\lambda = 1$, on kõikide alamjadade kaal võrdne ühega, ning nii saame pikkusega p alamjadade tuuma. Teisest küljest, kui $\lambda \to 0$, siis läheneb tuum p-spekter tuumale.

ϕ	ar	at	ca	cr	ct
car	λ^2	0	λ^2	λ^3	0
cat	0	λ^2	λ^2	0	λ^3

$K(car, cat) = \lambda^4$.

Keerukus $O(|t||s|)$.

Kirjandus: Ülalkirjeldatud tuumadest täpsemalt loe raamatust [14], pt 11. Sealt leiad ka arvutusalgoritmid.

Alguskoodoni tuvastamine

Tihti konstrueeritakse sobiv tuum ülesande spetsiifikast lähtudes. Näitena sellisest lähenemisviisist vaatleme nn translatsiooni alguskoodoni tuvastamise probleemi. Väidetavalt osa DNA-järjestusest ei kodeeri valke ja nii koosneb DNA kodeerivast osast (CDS – *coding sequences*) ja mittekodeerivast osast. Et kodeerivat osa ülejäänust eristada, on oluline teada, millal täpselt kodeeriv osa algab. On teada, et enamasti algab see ATG koodoni (tripletiga) (sellisel juhul nimetatakse seda TIS – *translation initiation sites*), kuid mitte iga ATG koodon pole kodeeriva osa algus. TIS koodonite tuvastamine on bioinformaatika oluline probleem, üks võimalik lähenemisviis on seda vaadelda kui klassifitseerimisprobleemi. Sellisel juhul on tunnuseks DNA lõigud, mis koosnevad 200 nukleotiidist enne ja pärast ATG-koodonit, klasse on kaks (on /ei ole alguskoodon), treeningandmetena on sellised DNA lõigud, mille korral on teada, kas vastav koodon on TIS või ei ole. Tuum peaks mõõtma DNA-lõikude omavahelist sarnasust. Kõige lihtsam on tähekaupa võrdlemine ning sellisel juhul võiks kasutada polünoom-tuuma

$$K(t,s) = \big(\sum_i \delta(s_i,t_i)\big)^d,$$

kus $\delta(s_i,t_i) = 1$ parajasti siis, kui $s_i = t_i$, mujal 0. Et kõrvuti olevate nukleotiide vahel eeldatakse sõltuvust, pole tähekaupa võrdlemine päris paslik ning nii asendatakse $\delta(s_i,t_i)$ üldisema funktsiooniga

$$\delta_l(s_i,t_i) := \big(\sum_{j=-l}^{l} \mu_j \delta(s_{i-j},t_i)\big)^r,$$

kus l on suhteliselt väike (lokaalne aken), μ_j, $j = -l,\ldots,l$ on kaalud (üldiselt sümmeetrilised, keskel suuremad) ning r enamasti 1,2,3,4. Seega tuum (*locally improved DNA-kernel*) on

$$K(t,s) = \big(\sum_i \delta_l(s_i,t_i)\big)^d.$$

Saab näidata, et selline funktsioon on tuum.

P-tuumad

Järgnevas põgusalt P-tuumadest. P-tuumad on nn generatiivsed tuumad, milliste korral püütakse modelleerida tunnuste tõenäosuslikku käitumist.

P-tuumad. Olgu $\mathcal{X}$ ülimalt loenduv hulk. P-tuumaks nimetatakse sellist tõenäosusjaotust hulgal $\mathcal{X}\times\mathcal{X}$, mis on positiivselt poolmääratud.
Mitte iga tõenäosusjaotus pole positiivselt poolmääratud (enamik pole isegi sümmeetrilised).
Olgu P tõenäosusjaoutus hulgal $\mathcal{X}$. Siis korrutismõõt $K = P\times P$ on P-tuum, sest $K(x,y) = P(x)P(y)$. See on aga suhteliselt limiteeritud tuum, realistlikum on järgmine. Olgu iga $\theta\in\Theta$ korral P_θ jaotus hulgal $\mathcal{X}$ ja $K_\theta = P_\theta\times P_\theta$ korrutismõõt. Kui Θ

(mudelite hulk) on ülimalt loenduv ja π sellel antud tõenäosusmõõt (eelmõõt), siis

$$K = \sum_{\theta} K_\theta \pi(\theta) \tag{4.5.25}$$

on P-tuum. Funktsioon Φ on sellisel juhul

$$\Phi(x) = \left(P_\theta(x) \sqrt{\pi(\theta)} \right)_\theta .$$

Tuumad kujul (4.5.25) moodustavad tegelikult üsna laia klassi. Vaatleme näiteks seda tüüpi tuumade kasutamist sõnade klassifitseerimisel. Oletame, et sõnas $s = s_1, \ldots, s_n$ on tähed iid juhuslikud suurused, kusjuures tähe s_i jaotus on segujaotus, st

$$P(s_i) = \sum_{k=1}^{m} P_k(s_i)\pi_k,$$

kus $k = 1, \ldots, m$ on varjatud komponentide hulk, P_k on komponendile k vastav emissioonijaotus ja π_k komponendi k tõenäosus. Sellisel juhul on sõna $s_1, \ldots, s_k$ tekkemehhanism järgmine: eelkõige realiseeruvad varjatud komponendid $\theta = \theta_1, \ldots, \theta_n$, $\theta_i \in \{1, \ldots, m\}$ ja selle komponentide jada tõenäosus on

$$\pi(\theta) = \prod_{i=1}^{n} \pi_{\theta_i}.$$

Iga varjatud komponent θ_i emiteerib tähe s_i vastavalt jaotusele P_{θ_i}, sõltumata teistest komponentidest ja nende poolt emiteeritud tähtedest. Seega

$$P(s|\theta) = \prod_{i=1}^{n} P_{\theta_i}(s_i), \quad P(s) = P(s|\theta)\pi(\theta).$$

Sõnad s ja t on sarnased, kui nad on tekitatud sama varjatud komponentide jada korral, kuid tähed on emiteeritud teineteisest sõltumatult. Seega

$$\begin{aligned} P(s,t) &= \sum_{\theta} P(s,t|\theta)\pi(\theta) = \sum_{\theta} \prod_{i=1}^{n} P(s_i,t_i|\theta_i)\pi_{\theta_i} \\ &= \sum_{\theta} \prod_{i=1}^{n} P_{\theta_i}(s_i)P_{\theta_i}(t_i)\pi_{\theta_i} = \sum_{\theta} P(s|\theta)P(t|\theta)\pi(\theta). \end{aligned}$$

Seega saadud tuum on P-tuum. On selge, et liitmise ja korrutamise võib ära vahetada, st

$$P(s,t) = \prod_{i=1}^{n} P(s_i,t_i) = \prod_{i=1}^{n} \sum_{\theta} P(s_i,t_i|\theta)\pi_{\theta_i} = \prod_{i=1}^{n} \sum_{\theta} P_{\theta_i}(s_i)P_{\theta_i}(t_i)\pi_{\theta_i} = \prod_{i=1}^{n} K(s_i,t_i),$$

kus $K(s_i,t_i)$ on P-tuum tähestikul.

Oletame nüüd, et varjatud komponendid θ_i pole sõltumatud, vaid moodustavad Markovi ahela. Sellisel juhul pole sõltumatud ka tähed s_i (kuid nad pole ka üldiselt Markovi ahel). Sellist mudelit nimetatakse varjatud Markovi ahelaks (HMM). Tõenäosus $\pi(\theta)$ avaldub sellisel juhul

$$\pi(\theta) = \pi(\theta_1, \ldots, \theta_l) = \pi(\theta_1)\pi(\theta_2|\theta_1)\pi(\theta_3|\theta_2)\cdots\pi(\theta_n|\theta_{n-1}),$$

kus $\pi(\theta_i|\theta_{i-1})$ on üleminekutõenäosused ja $\pi(\theta_1)$ on algtõenäosused. Siis, võttes $\pi(\theta_1|\theta_0) := \pi(\theta_1)$,

$$P(s,t) = \sum_\theta P(s,t|\theta)\pi(\theta) = \sum_\theta \prod_{i=1}^n P(s_i,t_i|\theta_i)\pi(\theta_i|\theta_{i-1}) \tag{4.5.26}$$

$$= \sum_\theta \prod_{i=1}^n P_{\theta_i}(s_i)P_{\theta_i}(t_i)\pi(\theta_i|\theta_{i-1}) = \sum_\theta P(s|\theta)P(t|\theta)\pi(\theta). \tag{4.5.27}$$

Seega P-tuum. Sarnaseid HMM-l põhinevaid tuumi kasutatakse DNA-järjestuste võrdlemisel. Järjestused on homoloogsed *homologous*, kui nad pärinevad ühest ja samast ahelast kuid on hilisema evolutsiooni käigus (teineteisest sõltumatuna) muutunud. Homoloogsete järjestuste modelleerimiseks sobib hästi HMM – varjatud ahel on ühine algne jada, emissioonijaotused modelleerivad aga muutusi (sealhulgas ka insertsioon ja deletsioon). Seega loomulik sarnasuse mõõt on tõenäosus, et kaks järjestust on homoloogsed. Selleks leiame iga võimaliku ühise algjada korral tõenäosuse, et järjestused pärinevad sealt ning seejärel keskmistame üle kõikide võimalike algjadade. Nii saame tõenäosuse (4.5.27).

Kirjandus: Täpsemalt P-tuumadest ja teistest generatiivsetest tuumadest loe [14], ptk 12.

4.6 Esitusteoreem

Tuletame meelde SVM klassifitseerimist (4.5.16):

$$\min_{h\in\mathcal{H},w_o} \frac{1}{2}\|h\|^2 + C\sum_{i=1}^n \big(1 - (y_i h(x_i) + w_o)\big)_+^p,$$

kus $\mathcal{H}$ on RKHS ja p on 1 või 2. Eelpool mainisime, et optimaalne h^* on kujul

$$\sum_{i=1}^n (\alpha_i^* y_i) K(x_i, \cdot).$$

Alljärgnevas tõestame selle formaalselt ning veendume, et selline h^* esitus läbi tugivektorite on seaduspära.

Paneme tähele, et (4.5.15) minimiseerimise võib läbi viia kahes osas: kõigepealt fikseerime konstandi w_o ja minimiseerime üle h, seejärel üle konstandi w_o. Seega fikseerime w_o ja vaatleme ülesannet

$$\min_{h\in\mathcal{H}} \frac{1}{2}\|h\|^2 + C\sum_{i=1}^{n}\big(1 - y_i(h(x_i) + w_o)\big)_+^p,$$

mis nüüd on kujul

$$\min_{h\in\mathcal{H}} \frac{1}{2}\|h\|^2 + L(h(x_1), \ldots, h(x_n)), \tag{4.6.1}$$

kus $L : \mathbb{R}^n \to \mathbb{R}$.

Teoreem 4.5 (Esitusteoreem (Representer Theorem)) *Olgu $\mathcal{X}$ suvaline hulk, $\mathcal{H}$ RKHS ja K vastav tuum. Siis iga funktsiooni $L : \mathbb{R}^n \to \mathbb{R}$ ning iga mittekahaneva funktsiooni $\Omega : \mathbb{R} \to \mathbb{R}$ korral on minimiseerimisülesandel*

$$\min_{h\in\mathcal{H}} \Omega(\|h\|^2) + L(h(x_1), \ldots, h(x_n)), \tag{4.6.2}$$

lahend kujul $h = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$, kus $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

Seega iga fikseeritud w_o korral on (4.6) optimaalne lahend h^* kujul $\sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$, järelikult on sellisel kujul ka parimale konstandile vastav h^*.

Tõestus. Defineerime

$$\mathcal{H}_\| = \operatorname{span}\big\{K(x_i, \cdot), i = 1, \ldots, n\big\}.$$

Seega $\mathcal{H}_\|$ on alamruum, mille moodustavad elemendid kujul $\sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$. Iga $h \in \mathcal{H}$ esitub üheselt kujul

$$h = h_\| + h_\perp,$$

kus $h_\| \in \mathcal{H}_\|$ ja $h_\perp$ kuulub $\mathcal{H}_\|$ ortogonaalsesse täiendisse. Iga j korral

$$h_\perp(x_j) = \langle h_\perp, K(x_j, \cdot)\rangle = 0,$$

millest

$$h(x_j) = h_\|(x_j) + h_\perp(x_j) = h_\|(x_j)\Big(= \sum_{i=1}^{n} \alpha_i K(x_j, x_i)\Big).$$

Seega iga w korral

$$L(h(x_1), \ldots, h(x_n)) = L(h_\|(x_1), \ldots, h_\|(x_n))$$

ehk $L(h(x_1), \ldots, h(x_n))$ minimiseerimine üle $\mathcal{H}_\|$ annab sama tulemuse, mis $L(h(x_1), \ldots, h(x_n))$ minimiseerimine üle $\mathcal{H}$.
Et Ω on mittekahanev, siis

$$\Omega(\|h\|^2) = \Omega(\|h_\|\|^2 + \|h_\perp\|^2) \geq \Omega(\|h_\|\|^2).$$

Kokkuvõttes, $\Omega(\|h\|^2) + L(h(x_1), \ldots, w(h_n))$ minimiseerimine üle $\mathcal{H}$ on ekvivalentne minimiseerimisega üle $\mathcal{H}_\|$. ■

Näide: Vaatleme *1-norm soft margin* probleemi (4.5.15) ($p = 1$):

$$\min_{h\in\mathcal{H},w_o\in\mathbb{R}} \frac{1}{2}\|h\|^2 + C\sum_{i=1}^{n}\big(1 - y_i(h(x_i) + w_o)\big)_+. \tag{4.6.3}$$

Tänu esitusteoreemile otsime lahendit kujul $h = \sum_{i=1}^{n} c_i K(x_i, \cdot)$. Seega

$$\|h\|^2 = \langle\sum_{i=1}^{n} c_i K(x_i,\cdot), \sum_{i=1}^{n} c_i K(x_i,\cdot)\rangle = \sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = c'Kc,$$

kus K on Grami maatriks ja $c = (c_1, \ldots, c_n)' \in \mathbb{R}^n$ on otsitav vektor .
Et iga $i = 1, \ldots, n$ korral

$$\langle h, K(x_i,\cdot)\rangle = \langle\sum_{j=1}^{n} c_j K(x_j,\cdot), K(x_i,\cdot)\rangle = \sum_{j=1}^{n} c_j K(x_j, x_i) = \sum_{j=1}^{n} K(x_i, x_j)c_j,$$

saame optimiseerimisprobleemile alljärgneva kuju:

$$\min_{c\in\mathbb{R}^n,w_o\in\mathbb{R}} \frac{1}{2}c'Kc + C\sum_{i=1}^{n}\big(1 - y_i\sum_{j=1}^{n} K(x_i, x_j)c_j + w_o\big)_+.$$

Abimuutujate abil:

$$\min_{c\in\mathbb{R}^n,w_o\in\mathbb{R},\xi} \frac{1}{2}c'Kc + C\sum_{i=1}^{n}\xi_i$$

$$\text{nii, et } y_i\big(\sum_{j=1}^{n} K(x_i, x_j)c_j + w_o\big) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

Lagrange'i funktsionaal:

$$\begin{aligned} L(c, \xi, w_o, \alpha, \gamma) &= \frac{1}{2}c'Kc + C\sum_{i=1}^{n}\xi_i + \sum_i \alpha_i\Big(1 - \xi_i - y_i\big(\sum_{j=1}^{n} K(x_i, x_j)c_j + w_o\big)\Big) - \sum_i \gamma_i\xi_i. \\ &= \frac{1}{2}c'Kc - u'Kc + \sum_{i=1}\xi_i(C - \alpha_i - \gamma_i) + \sum_i \alpha_i + w_o\sum_i \alpha_i, \end{aligned}$$

kus $u = (u_1, \ldots, u_n)$, $u_i = y_i\alpha_i$.
Gradient ∇_c:

$$\nabla_c L(c, \xi, w_o, \alpha, \gamma) = Kc - Ku, \quad \Rightarrow \quad \nabla_c L(c, \xi, w_o, \alpha, \gamma) = 0 \quad \Leftrightarrow \quad K(c - u) = 0.$$

Seega (tuleta meelde, et K on sümmeetriline) $c'Kc = c'Ku = u'Kc = u'Ku$. Võrdsustades osatuletised ξ ja w_o järgi nulliga, saame meile juba tuttavad tingimused:

$$\sum_i y_i\alpha_i = 0, \quad \alpha_i + \gamma_i = C, \quad i = 1, \ldots, n.$$

Seega

$$\theta(\alpha,\gamma)=\frac{1}{2}u'Ku-u'Kc+\sum_i\alpha_i.$$

Võttes nüüd $Kc=Ku$, saame meile juba tuttava duaalse probleemi:

$$\max_\alpha\sum_i\alpha_i-\frac{1}{2}u'Ku=\max_\alpha\sum_i\alpha_i-\frac{1}{2}\sum_{i,j=1}^n\alpha_i\alpha_jy_iy_jK(x_i,x_j)$$
$$\text{nii, et }\sum_i y_i\alpha_i=0,\quad C\geq\alpha_i\geq 0,\quad i=1,\ldots,n.$$

Olgu α^* duaalse probleemi lahend. Meid huvitav optimaalne c^* peab rahuldama võrdust $Kc^*=Ku^*$, kus $u^*=(y_1\alpha_1^*,\ldots,y_n\alpha_n^*)$. Kui leidub pöördmaatriks K^{-1}, siis $c^*=u^*$ ja otsitav lahend on

$$h^*=\sum_i y_i\alpha_i^*K(x_i,\cdot).\tag{4.6.4}$$

Kui aga K^{-1} ei leidu, peab c^* olema selline, et $K(c^*-u^*)=0$. See tähendab, et iga $j=1,\ldots,n$ korral peab summa $\sum_i(c_i^*-u_i^*)K(x_i,\cdot)$ rahuldama tingimust

$$\sum_i K(x_j,x_i)(c_i^*-u_i^*)=\langle\sum_i(c_i^*-u_i^*)K(x_i,\cdot),K(x_j,\cdot)\rangle=0.$$

Seega peab vektor $\sum_i(c_i^*-u_i^*)K(x_i,\cdot)$ olema ortogonaalne iga vektoriga alamruumist

$$\text{span}\{K(x_j,\cdot):\ \ j=1,\ldots,n\}.$$

Et aga $\sum_i(c_i^*-u^*)K(x_i,\cdot)$ kuulub samasse ruumi, siis

$$\sum_i(c_i^*-u_i^*)K(x_i,\cdot)=0,\quad\Rightarrow\quad h^*=\sum_i c_i^*K(x_i,\cdot)=\sum_i y_i\alpha_i^*K(x_i,\cdot).$$

Seega probleemi (4.6.3) lahend on ikkagi (4.6.4) isegi kui $c^*\neq u^*$.

Konstandi saame endiselt KKT tingimusest: kui $0<\alpha_i^*<C$, siis

$$y_i\Big(\sum_{j=1}^n K(x_i,x_j)c_j^*+w_o\Big)\Big)=y_i\Big(\sum_{j=1}^n K(x_i,x_j)u_j^*+w_o\Big)\Big)=y_i\Big(\sum_{j=1}^n\alpha_j^*y_i\langle K(x_j,\cdot),K(x_i,\cdot)\rangle+w_o\Big)\Big)=1.$$

Märkus: Olgu ülesanne kujul

$$\min_{w\in\mathcal{H}}\Omega(\|w\|^2)+L(\langle w,\Phi(x_1)\rangle,\ldots\langle w,\Phi(x_1)\rangle),\tag{4.6.5}$$

kus $\mathcal{H}$ ei ole ilmtingimata RKHS ja Φ seega ei pruugi olla kanooniline kujutis. Valemist (4.5.16) teame aga, et iga sellisel kujul oleva ülesande saab esitada kujul (4.6.1). Olgu w^* ülesane (4.6.5) lahend. Sellele vektorile w^* vastab funktsioon $h^*(x):=\langle w^*,\Phi(x)\rangle$,

mis on RKHS element ja mille korral $\|w^*\| = \|h^*\|$ (normide võrdsus tuleb sellest, et w^* on lahend seega w^* on väikseima normiga kõikide selliste w-de seast, mille korral $\langle w, \Phi(\cdot)\rangle = \langle w^*, \Phi(\cdot)\rangle$). Esitusteoreemist teame aga, et iga ülesande (4.6.1) lahendi võib esitada kujul $h^* = \sum_{i=1}^n \alpha_i^* K(x_i, \cdot)$ ehk iga x korral

$$h^*(x) = \langle \sum_{i=1}^n \alpha_i^* \Phi(x_i), \Phi(x)\rangle.$$

Et aga $h^*(x) := \langle w^*, \Phi(x)\rangle$, saame, et

$$w^* = \sum_{i=1}^n \alpha_i^* \Phi(x_i).$$

Seega, sõltumata millist Hilberi ruumi $\mathcal{H}$ ja kujutist Φ kasutame, ülesande (4.6.5) lahendit võib alati otsida kujul $\sum_{i=1}^n \alpha_i \Phi(x_i)$. Eeltoodud näites võib seega $\mathcal{H}$ asendada suvalise Hilberti ruumiga, $K(x, \cdot)$ asendada $\Phi(x)$ ja h asendada w-ga.

Kirjandus: Esitusteoreemi üldistusi võib leida raamatutest ([15], Ch. 4; [13], Ch. 5)

4.7 Regressioon

4.7.1 Kantregressioon ja lassoregressioon

Tuletame meelde harilikku lineaarset regressiooni, kus andmed on $(x_1, y_1), \ldots, (x_n, y_n)$ ning $x_i \in \mathbb{R}^d$ ja $y_i \in \mathbb{R}$. Harilik lineaarne vähimruutude meetod (*ordinary least squares* (OLS)) otsib konstante $w \in \mathbb{R}^d$ ja $a \in \mathbb{R}$, mis minimiseeriksid järgmist kaofunktsiooni

$$\sum_{i=1}^n \big(y_i - (w' x_i + a)\big)^2. \tag{4.7.1}$$

Varasemast (tuleta meelde (3.6.10)) teame, et lahendid avalduvad järgmiselt:

$$\hat{w} = \hat{\Sigma}^{-1}\big(\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{y}\bar{x}\big), \quad \hat{a} = \bar{y} - \hat{w}'\bar{x}.$$

Defineerides

$$Z := \begin{pmatrix} x_1^1 - \bar{x}^1 & x_1^2 - \bar{x}^2 & \cdots & x_1^d - \bar{x}^d \\ x_2^1 - \bar{x}^1 & x_1^2 - \bar{x}^2 & \cdots & x_1^d - \bar{x}^d \\ \cdots & \cdots & \cdots & \cdots \\ x_n^1 - \bar{x}^1 & x_n^2 - \bar{x}^2 & \cdots & x_n^d - \bar{x}^d \end{pmatrix}, \quad y := \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}$$

saame (siin 1 on vaid ühtedest koosnev vektor)

$$\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{y}\bar{x} = \frac{1}{n} Z'(y - \bar{y}1), \quad \hat{\Sigma} = \frac{1}{n} Z'Z,$$

nii, et

$$\hat{w} = \hat{\Sigma}^{-1}\big(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x}\big) = (Z'Z)^{-1}Z'y^o, \quad \text{kus} \quad y_i^o := y_i - \bar{y}.$$

Kantregression. Juhul kui maatriks $\hat{\Sigma}$ pole pööratav (näiteks $d > n$), kasutatakse alternatiive. Üks võimalus on nn. **kantregressioon** (*ridge regression*). Sellisel juhul minimiseeritakse kaofunktsiooni

$$\sum_{i=1}^{n}\big(y_i - (w'x_i + a)\big)^2 + \lambda\|w\|^2, \tag{4.7.2}$$

kus $\lambda > 0$ on regulariseeriv konstant. Seega on eelistus väiksema normiga (lamedamatel) regressioonifunktsioonidel, mis praktikas tähendab absoluutselt väiksemaid (nullilähedasi) regressioonikordajaid. Teame (vt peatükk 4.2.3), et ülesande (4.7.2) võib esitada ka kujul ($s > 0$ sõltub andmetest)

$$\min_{w,a} \sum_{i=1}^{n}\big(y_i - (w'x_i + a)\big)^2,$$
$$\text{nii, et } \|w\|^2 \le s.$$

Ülesande (4.7.2) lahendid $\hat{w}$ ja $\hat{a}$ avalduvad järgmiselt

$$\hat{w} = \big(Z'Z + \lambda I_d\big)^{-1}Z'y^o, \quad \hat{a} = \bar{y} - \hat{w}'\bar{x}. \tag{4.7.3}$$

Ülesanne 4.27 *Tõesta (4.7.3). Selleks kasuta tsentreerimist ja toimi järgmiselt:*

1. *Asenda tunnusvektorid x_i tsentreeritud vektoritega $z_i := x_i - \bar{x}1$ (kus 1 on vaid ühtedest koosnev vektor) ja vaatame ülesannet:*

$$\min_{v,b} \sum_{i=1}^{n}\Big(y_i - \big(v'z_i + b\big)\Big)^2 + \lambda\|v\|^2. \tag{4.7.4}$$

   Veenduda, et (4.7.2) ja (4.7.4) on ekvivalentsed, kusjuures minimiseerivate vektorite $\hat{v}$ ja $\hat{w}$ ning minimiseerivate konstantide $\hat{a}$ ja $\hat{b}$ vahelised seosed on $\hat{w} = \hat{v}$, $\hat{a} = \hat{b} - \bar{x}'\hat{v}$.

2. *Leidmaks $\hat{b}$, tsentreeri ka vektor y, st asenda see vektoriga $y^o = y - \bar{y}1$ ja esita (4.7.4) kujul*

$$\min_{v,b} \sum_{i=1}^{n}\Big(y_i^o - \big(v'z_i + (b - \bar{y})\big)\Big)^2 + \lambda\|v\|^2. \tag{4.7.5}$$

   Seejärel näita, et ülaltoodud ülesande lahendid on sellised, et $\hat{b} = \bar{y}$.

3. *Leidmaks $\hat{v}$ piisab nüüd järgmise ülesande lahendamisest.*

$$\min_v \sum_{i=1}^{n} \big(y_i^o - v'z_i\big)^2 + \lambda\|v\|^2 = \min_{v\in\mathbb{R}^d} \langle y^o - Zv, y^o - Zv\rangle + \lambda v'v. \tag{4.7.6}$$

Näita, et selle lahend on

$$\hat{v} = \big(Z'Z + \lambda I_d\big)^{-1} Z'y^o.$$

Kantregressiooni teoreetilisi põhjendusi:

1. Kui d on suur, otsime tunnuseid, mis oleksid regressioonis olulised. Seega jätame osa tunnuseid välja (vektori w vastavad komponendid on nullid). Kantregresioonis ei seata tõke mitte tunnuste arvule vaid nende suurustele.

2. Olgu (X, Y) sõltumatu valimist ning seetõttu ka kantregressiooni abil saadud hinnanguist $\hat{w}$ ja $\hat{a}$. Olgu $x_o \in \mathbb{R}^d$ ning $f(x_o) = E[Y|X = x_o]$. Tähistame $\hat{f}(x_o) = \hat{w}'x_o + \hat{a}$. Et (X, Y) ei sõltu valimist, siis $E\hat{f}(x_o) = E[\hat{f}(x_o)|X = x_o]$. Hindame prognoosi keskmist ruutviga fikseeritud tunnuse x_o korral:

$$E\big[\big(Y - \hat{f}(x_o)\big)^2|X = x_o\big] =$$
$$E\big[(Y - f(x_o))^2|X = x_o\big] + \big(f(x_o) - E\hat{f}(x_o)\big)^2 + E\big(\hat{f}(x_o) - E\hat{f}(x_o)\big)^2.$$

   Lahutus kehtib, sest fikseeritud valimi, st fikseeritud $\hat{f}$ korral (keskväärtus üle Y)

$$E[(Y - f(x_o))(f(x_o) - \hat{f}(x_o))|X = x_o] = E[(Y - f(x_o))|X = x_o](f(x_o) - \hat{f}(x_o)) = 0.$$

   Nii saame

$$E\big[\big(Y - \hat{f}(x_o)\big)^2|X = x_o\big] = E\big[(Y - f(x_o))^2|X = x_o\big] + E\big(f(x_o) - \hat{f}(x_o)\big)^2.$$

   Kolmeks jagamiseks liidame ja lahutame viimasele liidetavale $E\hat{f}(x_o)$ ning arvestame, et $\hat{f}(x_o)$ ei sõltu vektorist (X, Y), mistõttu

$$E\big((f(x_o) - E\hat{f}(x_o))(\hat{f}(x_o) - E\hat{f}(x_o)\big) = (f(x_o) - E\hat{f}(x_o))(E\hat{f}(x_o) - E\hat{f}(x_o)) = 0.$$

   Kolmeks lahutuse liidetavatest esimene on tunnuse tinglik varieeruvus keskväärtuse ümber, mis meist ei sõltu, teine on hinnangu ruutnihe ja kolmas hinnangu tinglik dispersioon. Kantregressioon hoiab dispersiooni kontrolli all nihke võimaliku suurenemise arvelt. Tõepoolest,

$$\hat{f}(x_o) - E\hat{f}(x_o) = (\hat{w} - E\hat{w})'x_o + \hat{a} - E\hat{a} = (\hat{w} - E\hat{w})'(x_o - \bar{x}) + (\bar{y} - E\bar{y}),$$

   sest $\hat{a} = \bar{y} - \hat{w}'\bar{x}$. Seega dispersioon

$$\begin{aligned} E\big((\hat{f}(x_o) - E\hat{f}(x_o)\big)^2 &= E\big((\hat{w} - E\hat{w})'(x_o - \bar{x})\big)^2 + D\bar{y} \\ &\le E[\|\hat{w} - E\hat{w}\|^2\|x_o - \bar{x}\|^2] + D\bar{y} \le (4s)E\|x_o - \bar{x}\|^2 + D\bar{y}. \end{aligned}$$

3. Maatriks $(Z'Z + \lambda I)$ on pööratav ka siis, kui $Z'Z$ pole.

Lassoregressioon **Lassoregressioon** (Tibshirani, 1996) erineb kantregressioonist selle poolest, et karistusliikmeks on $\|w\|_1$. Seega minimiseeritakse kaofunktsiooni

$$\sum_{i=1}^{n}\big(y_i-(w'x_i+a)\big)^2+\lambda\sum_i|w_i|, \tag{4.7.7}$$

kus $\lambda>0$ on regulariseeriv konstant. Teame, et iga $\lambda>0$ korral leidub positiivne $t>0$ (mis sõltub ka andmetest) nii, et (4.7.7) lahendid on ka järgmise ülesande lahendid:

$$\min_{w,a_o}\sum_{i=1}^{n}\big(y_i-(w'x_i+a_o)\big)^2,$$
$$\text{nii, et } \|w\|_1\le t.$$

Seega ka lasso korral on eelistus väiksema normiga (1-normiga) regressioonifunktsioonidel, mis praktikas tähendab jällegi absoluutselt väiksemaid (nullilähedasi) regressioonikordajaid. Nii lasso- kui ka kantregressiooni korral parameetri λ suurendamine (seega s ja t vähendamine) vähendab regressioonikordajaid $|\hat{w}_i|$. Selgub aga, et 1-normil ja 2-normil on regressiooniülesannete korral suur vahe: kantregressiooni korral jäävad regressioonikordajad enamasti positiivseks kuid lassoregressiooni korral muutuvad nad tihti nulliks. Seega, kui λ on suur, on nullist erinevaid regressionikordajaid vähe. Seda omadust nimetatakse *hõreduseks (sparseness)* ja see omadus teeb lassoregressiooni atraktiivseks.

Tsentreeritud versioon lassoregressioonist on

$$\min_{v,b}\sum_{i=1}^{n}\Big(y_i-\big(v'(x_i-\bar{x})+b\big)\Big)^2+\lambda\|v\|_1 \tag{4.7.8}$$

ning analoogiliselt kantregressiooniga saab näidata, et minimiseerivate vektorite $\hat{v}$ ja $\hat{w}$ ning minimiseerivate konstantide $\hat{a}$ ja $\hat{b}$ vahelised seosed on $\hat{v}=\hat{w}$, $\hat{a}=\hat{b}-\bar{x}'\hat{v}$. Tsentreeritud ülesande saab, jällegi, lahutada kaheks iseseisvaks ülesandeks: konstant a avaldub $\hat{a}=\bar{y}-\hat{w}'\bar{x}$ ning vektor $\hat{w}=\hat{v}$ on järgmise ülsesande lahend:

$$\min_{v}\sum_{i=1}^{n}\big(y_i^o-v'z_i\big)^2+\lambda\|v\|_1. \tag{4.7.9}$$

Erinevalt kantregressioonist, pole lassoregressiooni kordajaid lihtne leida, sest $\|\cdot\|_1$ norm pole pidevalt diferentseeruv. T. Hastie koos kolleegidega on välja pakkunud lassoregressiooniks sobiva nn. LARS-tarkvara [17].

Kirjandus: Kant- ja lassoregressioonist loe lähemalt raamatust [7].

4.7.2 Kantrgressioon tuuma abil

Grami maatriks ja hajuvusmaatriks. Tuleta meelde maatriksit Z. Tema read on tsentreeritud tunnusvektorid ning maatriks $S:=Z'Z$ on hajuvusmaatriks (*scatter matrix*), tema dimensioon on $d\times d$. Vaatleme maatriksit $K^o:=ZZ'$. Selle maatriksi elemendid

on skalaarkorrutised:

$$K^o(i,j) = (x_i - \bar{x})'(x_j - \bar{x})$$

ja selle maatriksi dimensioon on $n \times n$. Seega, võttes $\Phi(x) = x$, saame, et K^o on Grami maatriks (ülaindeks o näitab, et tunnusvektorid on eelnevalt tsentreeritud). Klassikalises mitmemõõtmelises analüüsis d on harilikult palju väiksem kui n, mistõttu K^o pole täisastakuga ja enamasti tegeletakse hajuvusmaatriksiga S või, ekvivalentselt, kovariatsioonimaatriksiga $\hat{\Sigma} = \frac{1}{n}S$. Tuumameetodite korral aga enamasti $d > n$, mistõttu nüüd pakub huvi just maatriks K^o.

Kantregression maatriksi K^o kaudu. Tuletame meelde, et kantregressiooni ülesande (4.7.3) lahend on

$$\hat{w} = \big(Z'Z + \lambda I_d\big)^{-1} Z' y^o.$$

Järgmine ülesanne näitab, et lahendi saab esitada kujul $\sum_{i=1}^n \alpha_i z_i = Z'\alpha$, kus (tulete meelde), z_i on tsentreeritud tunnusvektorid (ja maatriksi Z' veerud) ja α on n-dimensionaalne vektor.

Ülesanne 4.28 *Tõesta, et $\hat{w} = Z'\alpha$, kus*

$$\alpha = \lambda^{-1}(y^o - Z\hat{v}) = (K^o + \lambda I_n)^{-1} y^o.$$

Tuleta ka meelde, et $\hat{w}$ on tsentreeritud ülesande (4.7.6) lahend, mistõttu esitusteoreemist saame, et selle saab alati kirjutada kujul $\sum_{i=1}^n \alpha_i z_i = Z'\alpha$. Selles valguses pole ülaltoodud ülesanne üllatav. Seega, teades esitusteoreemi, võime lahendit w otsida alati kujul $w = Z'\alpha$. Seega $w'w = \alpha' K^o \alpha$ ja ülesanne (4.7.6) on

$$\min_{\alpha} \langle y^o - K^o\alpha, y^o - K^o\alpha\rangle + \lambda\alpha' K^o \alpha. \tag{4.7.10}$$

Võttes tuletise α järgi ning võrdsustades selle nulliga, saame

$$K^o\big((K^o + \lambda I_n)\alpha - y^o\big) = 0. \tag{4.7.11}$$

Kui K^o on pööratav, siis ainus lahend on

$$\alpha^* = (K^o + \lambda I_n)^{-1} y^o.$$

Kui K^o pole pööratav, siis põhimõtteliselt võib olla erinevaud vektreid α, mis rahuldavad ülaltoodud võrdust (4.7.11), kuid igaüks neist peab olema selline, et $Z'\big((K^o+\lambda I_n)\alpha - y^o\big) = 0$. See tuleneb sellest, et võrdusest (4.7.11) järeldub

$$\big((K^o+\lambda I_n)\alpha - y^o\big)' K^o \big((K^o+\lambda I_n)\alpha - y^o\big) = \Big(Z'\big((K^o+\lambda I_n)\alpha - y^o\big)\Big)'\Big(Z'\big((K^o+\lambda I_n)\alpha - y^o\big)\Big) = 0.$$

Nüüd aga

$$Z'\big((K^o+\lambda I_n)\alpha - y^o\big) = 0 \quad \Rightarrow \quad Z'(K^o+\lambda I_n)\alpha = Z'ZZ'\alpha + \lambda Z'\alpha = (S+\lambda I_d)Z'\alpha = Z'y^o.$$

Et $(S + \lambda I_d)$ on pööratav, saame et kõik võrdust (4.7.11) rahuldavad vektorid α defineerivad ühe ja sama vektori $Z'\alpha = Z'\alpha^*$.

Prognoos. Vektori $\hat{w} = Z'\alpha^*$ kaudu saadud prognoos punktis z_i on

$$\hat{y}_i = \hat{w}'z_i = (Z'\alpha^*)'z_i = (\alpha^*)'Zz_i = \sum_{j=1}^{n} \alpha_j^* z_j' z_i = \sum_{j=1}^{n} \alpha_j^* K_{ji}^o.$$

Oletame korraks, et $d > n$ ja maatriks K^o on pööratav. Olgu $\lambda = 0$. Siis kantregressiooniülesanne on tavaline vähimruutude regressiooniülesanne ning $\alpha^* = (K^o)^{-1}y^o$. Prognoosivektor:

$$\hat{y}' = \hat{w}'Z' = \alpha^{*\prime}K^o = (y^o)'(K^o)^{-1}K^o = (y^o)'$$

ehk regressioonitasand läbib kõiki punkte – ülesobitumus. Seega regerssioon kõrge dimensiooniga ruumis nõuab regulariseerimist. Kantregressioon on üks lihtne võimalus.

Kantregression Lagrange'i meetodil. Suure n korral ei pruugi $(K^o + \lambda I_n)^{-1}$ leidmine olla kerge ning kasulikum on kantregressiooni vaadelda optimiseerimisülesandena. Abimuutujate abil on (4.7.10) järgmine:

$$\begin{aligned} &\min_{a,\xi,\alpha} \ \lambda\alpha' K^o\alpha + \|\xi\|^2 \\ &\text{nii et } K^o\alpha = y^o - \xi. \end{aligned} \tag{4.7.12}$$

Lagrange'i funktsionaal:

$$L(\alpha, \xi, \gamma) = \lambda\alpha' K^o\alpha + \|\xi\|^2 + \gamma'(y^o - K^o\alpha - \xi).$$

Gradiendid:

$$\nabla_\alpha L(\alpha, \xi, \gamma) = 2\lambda K^o\alpha - \gamma' K^o = 0 \quad \Rightarrow \quad K^o\alpha = \frac{1}{2\lambda}K^o\gamma.$$

$$\nabla_\xi L(\alpha, \xi, \gamma) = 2\xi - \gamma = 0 \quad \Rightarrow \quad \xi = \frac{1}{2}\gamma.$$

Seega

$$\alpha' K^o\alpha = \frac{1}{2\lambda}\alpha' K^o\gamma = \frac{1}{2\lambda}\gamma' K^o\alpha = \frac{1}{4\lambda^2}\gamma' K^o\gamma, \quad \|\xi\|^2 = \frac{1}{4}\gamma'\gamma$$

ja asendades saadud valemid Lagrange'i funktsionaali, saame

$$\theta(\gamma) = \gamma' y^o - \frac{1}{4\lambda}\gamma K^o\gamma - \frac{1}{4}\gamma'\gamma = \gamma' y^o - \frac{1}{4\lambda}\gamma'\big(K^o + \lambda I_n\big)\gamma.$$

Seega duaalne probleem on

$$\max_\gamma \gamma' y^o - \frac{1}{4\lambda}\gamma'\big(K^o + \lambda I_n\big)\gamma. \tag{4.7.13}$$

Jällegi, kui K^o on pööratav, siis ülesandel (4.7.12) on üks lahend $\alpha^* = \frac{1}{2\lambda}\gamma^*$, kus γ^* on duaalse ülesande lahend. kui K^o pole pööratav, võib olla mitu vektorit α, mis kõik rahuldavad võrdust

$$K^o\alpha = \frac{1}{2\lambda}K^o\gamma^*,$$

kuid kõik nad defineerivad sama vektori

$$\hat{w} = Z'\alpha = Z'\alpha^* = \frac{1}{2\lambda}(\gamma^*)'Z'.$$

KKT:

$$y_i^o - \sum_j K_{ij}^o \alpha_j^* = \xi_i^* = \frac{\gamma_i^*}{2}.$$

Seega

$$\gamma_i^* = 0 \quad \Leftrightarrow \quad y_i^o = \sum_j K_{ij}^o \alpha_j^* \tag{4.7.14}$$

ehk tugivektor $\alpha_i^* = 0$ parajasti siis, kui prognoos selles punktis on y_i^o.

Pane tähele, et antud juhul on duaalset ülesannet kerge lahendada ning lahend on

$$\gamma^* = 2\lambda(K^o + \lambda I_n)^{-1} y^o, \quad \text{millest} \quad \alpha^* = \frac{1}{2\lambda}\gamma^* = (K^o + \lambda I_n)^{-1} y^o.$$

Kantregressioon kujutise Φ abil

Olgu $\Phi : \mathcal{X} \to \mathcal{H}$. Hilberti ruumis $\mathcal{H}$ on valim järgmine:

$$(\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n).$$

Kantregressiooni probleem (*kernel ridge regression*) on

$$\min_{w \in \mathcal{H}, a \in \mathbb{R}} \sum_i^n \big(y_i - (\langle w, \Phi(x_i)\rangle + a)\big)^2 + \lambda \|w\|^2. \tag{4.7.15}$$

Keskmistamine ruumis $\mathcal{H}$. Sarnaselt ruumiga $\mathbb{R}^d$, võime kasutada keskmistamist, mis aga nüüd tuleb läbi viia Hilberti ruumis. Olgu Φ_S valimi $\Phi(x_1), \ldots, \Phi(x_n)$ keskmine, st

$$\Phi_S = \frac{1}{n}\sum_{i=1}^n \Phi(x_i).$$

Olgu kujutis

$$\Phi^o : \mathcal{X} \to \mathcal{H}, \quad \Phi^o(x) = \Phi(x) - \Phi_S$$

ja olgu K^o tsentreeritud Grami maatriks st K^o on $n \times n$ maatriks, mille elemendid on

$$\begin{aligned} K_{ij}^o :=: K^o(x_i, x_j) &= \langle \Phi^o(x_i), \Phi^o(x_j)\rangle = \langle \Phi(x_i), \Phi(x_j)\rangle - \langle \Phi(x_i), \Phi_S\rangle - \langle \Phi_S, \Phi(x_j)\rangle + \langle \Phi_S, \Phi_S\rangle \\ &= K(x_i, x_j) - \frac{1}{n}\sum_{k=1}^n K(x_i, x_k) - \frac{1}{n}\sum_{k=1}^n K(x_k, x_j) + \frac{1}{n^2}\sum_{i,j} K(x_i, x_j). \end{aligned}$$

Seega maatriksi K^o saab esialgsest Grami maatriksist K järgmise lihtsa teisenduse abil (siis 1 on ühtedest koosenv vektor) :

$$K^o = K - \frac{1}{n}K11' - \frac{1}{n}11'K + \frac{1}{n^2}1'K1. \tag{4.7.16}$$

Lahendamine. Pärast tunnusvektorite ja vektori y tsentreerimist saame (just nagu ennegi) ülesande (4.7.15) lahendid $\hat{a}$ ja $\hat{w}$, kus

$$\hat{a} = \bar{y} - \langle \hat{v}, \Phi_S \rangle, \quad \hat{w} = \hat{v}$$

ja $\hat{v}$ on järgmise probleemi lahend

$$\min_{v \in \mathcal{H}} \sum_i^n \big(y_i^o - \langle v, \Phi^o(x_i) \rangle\big)^2 + \lambda \|v\|^2.$$

Teame, et üldisudt kitsendamata võime ruumiks $\mathcal{H}$ võtta RHKS, ent kui see ka ei ole nii, siis esitusteoreemist saame (tulete meelde märkus peatükis 4.6), et

$$\hat{v} = \sum_i \alpha_i^* \Phi^o(x_i),$$

kus α^* on järgmise, meile juba tuttava ülesande (4.7.10), lahend:

$$\min_{\alpha \in \mathbb{R}^n} \big\langle y^o - K^o\alpha, y^o - K^o\alpha \big\rangle + \lambda \alpha' K^o \alpha.$$

Ülalatoodust teame, et

$$\alpha^* = (K^o + \lambda I_n)^{-1} y^o$$

ning kõik teised lahendid (kui neid on) defineerivad sama $\hat{v}$. Me teame samuti, et probleemi (4.7.10) võib vaadelda optimiseerimisprobleemina (4.7.12) ning (iga) selle probleemi lahend on $\alpha^* = \frac{1}{2\lambda}\gamma^*$, kus γ^* on duaalse probleemi (4.7.13) lahend.

Konstandi määramine. Vektori α^* kaudu saame optimaalse konstandi

$$\hat{a} = \bar{y} - \langle \hat{v}, \Phi_S \rangle = \bar{y} - \langle \sum_j \alpha_j^* \Phi^o(x_j), \Phi_S \rangle = \bar{y} - \langle \sum_j \alpha_j^* \Phi(x_j), \Phi_S \rangle - (\sum_j \alpha_j^*) \langle \Phi_S, \Phi_S \rangle$$
$$= \bar{y} - \frac{1}{n} 1' K \alpha^* - (1'\alpha^*) \langle \Phi_S, \Phi_S \rangle = \bar{y} - \frac{1}{n} 1' K \alpha^* - \frac{(1'\alpha^*)}{n^2} 1' K 1.$$

Prognoos. Iga $x \in \mathcal{X}$ korral prognoos

$$\hat{y}(x) := \langle \hat{w}, \Phi(x) \rangle + \hat{a} = \sum_i \alpha_i^* K(x, x_i) + \hat{a}.$$

Tugivektorid. Võrdus (4.7.14): kui $\alpha_i^* = 0$, siis

$$y_i^o = \sum_j K_{ij}^o \alpha_j^* = \langle \hat{w}, \Phi^o(x_i) \rangle,$$

st kordaja α_i^* on null ainult siis, kui punkt $(\Phi^o(x_i), y_i^o)$ on regressioonitasandil. Kui $\lambda > 0$, siis seda tuleb üsna harva ette, mistõttu enamus kordajatest α_i^* on nullist erinevad ehk enamus vektoritest x_i on tugivektorid.

4.7.3 ϵ-tugivektorregression

Kantregressiooni lahend $\hat{w} = \sum_i \alpha_i^* \Phi(x_i)$ ei ole enamasti hõre, st enamus valimist on tugivektorid. Hõredus on aga tihti eesmärk, mistõttu kasutatakse alternatiive. Üks neist seiseneb ruutkaofunktsiooni asendamises funktsiooniga, mis ignoreerib väikesi prognooosivigu, täpsemalt neid, mis on väiksemad kui ϵ.

Formaalselt, olgu $\epsilon > 0$ fikseeritud ja iga $y, f(x) \in \mathbb{R}$ korral defineerime **ϵ-tundetu p kao** (ϵ-*insensitive* p-*loss*):

$$|y - f(x)|_\epsilon^p, \quad \text{kus} \quad |y - f(x)|_\epsilon := \max\{0, |y - f(x)| - \epsilon\}.$$

Seega

$$\big(|y - f(x)|_\epsilon\big)^p = 0 \quad \text{iff } |y - f(x)| \le \epsilon.$$

Tüüpiliselt $p = 1$ või $p = 2$, ning mõlemad on kombineeritud kantregressiooooniga.

Ruutkadu: $p = 2$. Kui $p = 2$, saame kantregressiooni üldistuse

$$\min_{w \in \mathcal{H}, a \in \mathbb{R}} \sum_{i=1}^{n} |y_i - (\langle w, \Phi(x_i)\rangle + a)|_\epsilon^2 + \lambda \|w\|^2. \tag{4.7.17}$$

Vaadeledes seda kui optimeerimisülesannet ning otsides lahendit kujul $\sum_i \alpha_i \Phi(x_i)$ (esitusteoreem), saame maatrikskujul järgmise probleemi

$$\begin{aligned} &\min_{a,\xi,\alpha} \ \lambda \alpha' K \alpha + \|\xi^+\|^2 + \|\xi^-\|^2 \\ &\text{nii, et } K\alpha + a1 - y \le \xi^- + \epsilon 1 \\ &\qquad y - K\alpha - a1 \le \xi^+ + \epsilon 1. \end{aligned} \tag{4.7.18}$$

Et $\epsilon > 0$ korral pole kaofunktsioon enam ruutfunktsioon, siis keskmistamisest pole kasu ning me ei saa enam leida vektorit $\hat{w}$ ja konstanti $\hat{a}$ eraldi. Seetõttu ülaltoodud ülesandes on esialgne (mitte tsentreeritud) Grami maatriks ning konstant tuleb hiljem määrata KKT tingimustest.

Duaalne ülesanne (võrdle (4.7.13))

$$\begin{aligned} &\max_\gamma \ \gamma' y - \frac{1}{4\lambda} \gamma' \big(K + \lambda I_n\big)\gamma - \epsilon \|\gamma\|_1 \\ &\text{nii, et } 1'\gamma = 0. \end{aligned} \tag{4.7.19}$$

Duaalse ülesande lahendi γ^* kaudu avaldub esialgse ülesande (4.7.18) (ning seeläbi ka ülesande (4.7.17)) lahend järgmiselt:

$$\alpha^* = \frac{1}{2\lambda} \gamma^*, \quad \hat{w} = \sum_i \alpha_i^* \Phi(x_i).$$

KKT tingimustest järeldub, et kui $\gamma_i^* \neq 0$, siis

$$y_i - \sum_j K_{ij}\alpha_i^* - \hat{a} = y_i - \langle \hat{w}, \Phi(x_i)\rangle - \hat{a} = \begin{cases} \frac{\gamma_i^*}{2} + \epsilon, & \text{kui } \gamma_i^* > 0; \\ \frac{\gamma_i^*}{2} - \epsilon, & \text{kui } \gamma_i^* < 0; \end{cases} \tag{4.7.20}$$

Nendest võrdustest saab optimaalse $\hat{a}$. Võrdusest (4.7.20) järeldub samuti, et

$$|y_i - (\sum_j K_{ij}\alpha_i^* + \hat{a})|_\epsilon = |y_i - (\langle \hat{w}, \Phi(x_i)\rangle + \hat{a})|_\epsilon = \frac{|\gamma_i^*|}{2}.$$

järelikult tugivektorid (need vektorid x_i, mille korral $\alpha_i^* \neq 0$) on kõik sellised, et x_i paar $(y_i, \Phi(x_i))$ on regressioonitasandist kaugemal kui ϵ.

Kaofunktsioon $p = 1$. Probleem

$$\min_{w\in\mathcal{H}, a\in\mathbb{R}} \sum_{i=1}^{n} |y_i - (\langle w, \Phi(x_i)\rangle + a)|_\epsilon + \lambda\|w\|^2. \tag{4.7.21}$$

Vaatleme ülesannet jällegi optimiseerimisülesandena ning otsime laghendit kujul $\sum_i \alpha_i \Phi(x_i)$. Nii saame ülesande

$$\begin{aligned} \min_{a,\xi,\alpha} \ & \lambda\alpha' K\alpha + \|\xi^+\|_1 + \|\xi^-\|_1 \\ \text{nii, et } & K\alpha + a1 - y \leq \xi^- + \epsilon 1 \\ & y - K\alpha - a1 \leq \xi^+ + \epsilon 1 \\ & \xi^+ \geq 0, \quad \xi^- \geq 0. \end{aligned}$$

Selle ülesande lahend α^* on

$$\alpha^* = \frac{\gamma^*}{2\lambda},$$

kus γ^* on duaalse ülesande lahend. Duaalne ülesanne:

$$\begin{aligned} \max_\gamma \ & -\frac{1}{4\lambda}\gamma' K\gamma + \gamma' y - \epsilon\|\gamma\|_1 \\ \text{nii, et } & \gamma' 1 = 0 \\ & |\gamma_i| \leq 1, \quad \forall i. \end{aligned}$$

Et tegemist on 1-normiga, kohtame jällegi nn *box constraints*: $|\alpha_i^*| \leq \frac{1}{2\lambda}$ ning, just nagu 1-morm SVM klassifitseerimise korralgi kehtib: kui $(y_i, \Phi(x_i))$ on regressioonitasandist kaugemal kui ϵ, siis $|\alpha_i^*| = \frac{1}{2\lambda}$.

Seega *in-bound* tugivektorid on need valimi elemendid, mille korral

$$|\alpha_i^*| \in \big(0, \frac{1}{2\lambda}\big).$$

KKT tingimustest saame, et *in-bound* tugivektori x_i korral

$$y_i - \sum_j K_{ij}\alpha_i^* - \hat{a} = y_i - \langle \hat{w}, \Phi(x_i)\rangle - \hat{a} = \begin{cases} \epsilon, & \text{kui } \gamma_i^* > 0; \\ -\epsilon, & \text{kui } \gamma_i^* < 0. \end{cases} \tag{4.7.22}$$

Sellest võrdusest saame leida optimaalse $\hat{a}$ ning see võrdus näitab samuti, et x_i on *in-bound* tugivektor, kui paari $(y_i, \Phi(x_i))$ kaugus regressioonitasandist on täpselt ϵ.

Kirjandus: Regressioonist tuumadega loe [14], Ch 7.3; [15], Ch 9; [12], Ch 6; [13], Ch. 9.

4.8 SVM mõjususest

Selleks, et tugivektormasinad (nii klassifitseerimisel või regressioonil) annaksid mõjusa hinnangu, peab tuum olema selline, et tema tekitatud RKHS oleks piisavalt heade lähendamisomadustega. Eelpool nägime, et selline omadus on universaalsetel tuumadel - neile vastav RKHS on kõikjal tihe ruumis $C(\mathcal{X})$, st iga pidevat funktsiooni saab kuitahes hästi lähendada mõne RKHS elementiga supremum-normi mõttes. Kitsendav eeldus oli see, et $\mathcal{X}$ peab olema komapktne meetriline ruum, näiteks ruumi $\mathbb{R}^d$ tõkestatud kinnine alamhulk (kuid mitte ruum ise). Kõikjal tihedus supremum-normi mõttes on väga tugev nõue. Selgub, et mõjususeks piisab vähemast – tuum peab olema selline, et vastav RKHS on kõikjal tihe ruumis $L_p(\mathbb{R}^d, \mathcal{B}, P_X)$, kus P_X on tunnusvektori jaotus ja $p > 0$ on sobiv aste (sõltub kaofunktsioonist). Kui funktsioonide jada koondub supremum-normi mõttes, siis koondub ta ka ruumis L_p, mistõttu tihedus ruumis L_p on nõrgem nõue, samas aga puudub ruumi $\mathcal{X}$ kompaktsuse eeldus.

Soft-margin SVM mõjususest. Tuletame meelde, et p-norm SVM minimiseerib funktsiooni (4.5.16):

$$\min_{h\in\mathcal{H}, w_o} \frac{1}{2}\|h\|^2 + C\sum_{i=1}^n \big(1 - y_i(h(x_i) + w_o)\big)_+^p.$$

Järgnevas vaatleme 1- ja 2-norm SVM klassifitseerimisreegli mõjusust, kuid lihtsuse mõttes loobume konstandist w_o. Teame, et kui RKHS on piisavalt rikas (näiteks Gaussi tuum), siis on selline eeldus õigustatud. Samuti tähistame $\lambda = \frac{1}{2Cn}$ ning et $\mathcal{H}$ on RKHS, mille elemendid on funktsioonid h, saame p-norm SVM-klassifitseerija kui järgmise optimiseerimisülesande lahendi:

$$\min_{h\in\mathcal{H}} \lambda\|h\|^2 + \frac{1}{n}\sum_{i=1}^n \big(1 - y_i h(x_i)\big)_+^p \tag{4.8.1}$$

Teame, et lahend on kujul $h_n = \sum_{i=1}^n \alpha_n^* K(x_i, \cdot)$ ja saadud klassifitseerija $g_n = \operatorname{sgn}(h_n)$. Kui $p = 1$ või $p = 2$, siis saadud klassifitseerija riski kaugust Bayesi riskist mõõdab järgmised nn *oraaklivõrratused* ([13], Thm 8.1 ja 6.24)

Teoreem 4.6 *Vaatleme $p \in \{1, 2\}$, $\mathcal{X} = \mathbb{R}^d$. Olgu tuum K selline, et $\|K\|_\infty \le 1$ ja vastav RKHS $\mathcal{H}$ on separaabel. Olgu (X, Y) jaotus selline, et $\mathcal{H}$ on kõikjal tihe ruumis*

$L_p(\mathbb{R}^d, \mathcal{B}, P_X)$. Siis iga $\lambda > 0$, $n \geq 1$ ja $\delta \in (0,1)$ korral kehtib tõenäosusega vähemalt $1-\delta$ (üle valimite)

$$R(g_n) - R^* < A^1(\lambda) + \frac{1}{\lambda}\Big(\sqrt{\frac{8\ln\frac{1}{\delta}}{n}} + \frac{8\ln\frac{1}{\delta}}{n} + \sqrt{\frac{4}{n}}\Big), \quad \textit{kui} \quad p = 1 \tag{4.8.2}$$

$$\big(R(g_n) - R^*\big)^2 < A^2(\lambda) + \frac{1}{\lambda}(2\lambda^{-\frac{1}{2}} + 2)^2\Big(\sqrt{\frac{8\ln\frac{1}{\delta}}{n}} + \frac{8\ln\frac{1}{\delta}}{n} + \sqrt{\frac{4}{n}}\Big), \quad \textit{kui} \quad p = 2. \tag{4.8.3}$$

Siin A^1 ja A^2 on hulgal $[0,\infty)$ defineeritud mittenegatiivsed, nõgusad funktsioonid, kusjuures $A^i(0) = 0$ ning $g_n = sgn(h_n)$, kus h_n on (4.8.1) lahend.

Märkused:

1. Funktsioonid A^1 ja A^2 on sisuliselt SVM lähendamiviga, nende täpse definitsiooni ja omadused leiad ([13], Def. 5.14 ja Lemma 5.15). Mõjususe sisukohalt on oluline, et $A^i(\lambda_n) \to 0$, kui $\lambda_n \to 0$.

2. Ülaltoodud võrratused kehtivad kumbki tõenäosusega $1-\delta$ (mõlemad koos kehtivad järelikult tõenäosusega vähemalt $1-2\delta$).

Ülaltoodud võrratustest järeldub 1- ja 2-norm SVM klassifitseerimisreegli universaalne mõjusus Gaussi tuuma ja sobivalt valitud regulariseerimiskonstantide λ_n korral. Tõepoolest, Gaussi tuuma korral $\|K\|_\infty = \sup_x \sqrt{K(x,x)} = 1$. Gaussi tuum on pidev ja, et $\mathbb{R}^d$ on separaabel, on separaabel ka vastav $\mathcal{H}$. Saab näidata, et mistahes $\sigma > 0$, $p = [1,\infty)$ ja ruumil $\mathbb{R}^d$ antud tõenäosusmõõdu P korral, Gaussi tuumale vastav RKHS on kõikjal tihe ruumis $L_p(\mathbb{R}^d, \mathcal{B}, P)$ ([13], Thm 4.64). Seega Gaussi tuuma korral on ülaltoodud võrratused universaalsed, st nad kehtivad mistahes (X,Y) jaotuse korral. Selleks, et kehtiks mõjusus, peame valima jada λ_n nii, et võrratuste parem pool koonduks nulliks. Eelkõige peab kehtima $\lambda_n \to 0$, sest muidu ei koondu nulliks $A^i(\lambda_n)$, kus $i = 1,2$. Samas aga ei saa λ_n koonduda nulliks liiga kiiresti, sest vastasel juhul ei koondu nulliks teine liige. Kui $p = 1$, peab λ_n koonduma nii aeglaselt, et $n\lambda_n^2 \to \infty$ (veendu!), kui $p = 2$, peab λ_n koonduma veelgi aeglasemalt: $\lambda_n^4 n \to \infty$ (veendu!). Kui λ_n on selline, et võrratuse parem pool koondub nulliks, on mõjusust, st koondumist $R(g_n) \xrightarrow{P} R^*$ kerge näidata. Tõepoolest vastavalt tõenäosuse järgi koondumise definitsioonile piisab, kui näitame, et iga $\epsilon > 0$ ka $\delta > 0$ korral leidub n_o nii, et $\mathbf{P}(R(g_n) - R^* > \epsilon) \leq \delta$ iga $n > n_o$ korral. Näitamaks 1-norm SVM reegli mõjusust fikseeri $\epsilon > 0$ ja $\delta > 0$, ning vali n_o nii suur, et

$$A^1(\lambda_n) + \frac{1}{\lambda_n}\Big(\sqrt{\frac{8\ln\frac{1}{\delta}}{n}} + \frac{8\ln\frac{1}{\delta}}{n} + \sqrt{\frac{4}{n}}\Big) < \epsilon,$$

kui $n > n_o$. Analoogiliselt järeldub 2-norm SVM mõjusus. Et võrratused on universaalsed, on saadud reeglid universaalselt mõjusad.

Ülesanne 4.29 *Näita, et kui λ_n koondub piisavalt aeglaselt, siis 1- ja 2-norm SVM-reeglid ka tugevalt (universaalselt) mõjusad.*

Regressiooni mõjususest. Olgu $\mathcal{H}$ RKHS, $h \in \mathcal{H}$ ning vaatleme ϵ-tunndetut kaofunktsiooni $L(y, h(x)) = |y - h(x)|_\epsilon^p$, kus $\epsilon \geq 0$ ja $p \geq 1$. Seega erijuhul on tegemist hariliku p-kaofunktsiooniga. Eelmises peatükis vaadeldud regressiooniülesanded (va Lasso) võib seega esitada (ilma konstantideta) kujul

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} |y_i - h(x_i)|_\epsilon^p + \lambda \|h\|^2. \tag{4.8.4}$$

Kui $\epsilon = 0$ ja $p = 1$, saame hariliku kantregressiooni. Paneme tähele, et erinevalt eelmiset paetükist on ülaltoodud summa esimene tegur korrutatud läbi suurusega $\frac{1}{n}$. Ülaltoodust teame jällegi, et ülesande (4.8.4) lahend on kujul $h_n = \sum_{i=1}^n \alpha_n^* K(x_i, \cdot)$. Olgu (X, Y) jaotus selline, et $E|Y|^p < \infty$. Sellisel juhul on funktsiooni $h \in \mathcal{H}$ risk

$$R(h) = E|Y - h(X)|_\epsilon^p = \int_{\mathbb{R}^{d+1}} |y - h(x)|_\epsilon^p dF(y, x)$$

ja minimaalne võimalik risk R^* on, nagu ikka, $R^* = \inf_f R(f)$, kus infiimum on võetud üle kõikide mõõtuvate funktsioonide. Sarnaselt klassifitseerimisreegliga ütleme, et **SVM-regressioon on (tugevalt) mõjus**, kui (4.8.4) lahendid h_n on selliselt, et

$$R(h_n) \xrightarrow{P} R^*, \quad \text{(p.k.)}.$$

Teoreem 4.7 ([13], Thm. 9.1) *Vaatleme $p \in [1, \infty)$, $\mathcal{X} = \mathbb{R}^d$. Olgu tuum K tõkestatud, st $\|K\|_\infty < \infty$ ja pidev. Olgu (X, Y) jaotus selline, et tuumale vastav RKHS on $\mathcal{H}$ kõikjal tihe ruumis $L_p(\mathbb{R}^d, \mathcal{B}, P_X)$ ning $E|Y|^p < \infty$. Kui $\lambda_n \to 0$ on selline jada, et $\lambda_n^{p^*} n \to \infty$, kus $p^* = \max\{2p, p^2\}$, siis*

$$R(h_n) \xrightarrow{P} R^*,$$

kus h_n on (4.8.4) lahend.

Et Gaussi tuum rahuldab teoreemi eeldusi (pidev, tõkestatud, kõikjal tihe ruumis L_p), siis järeldub teoreemist SVM-regressiooni mõjusus iga (X, Y) jaotuse korral, mis rahuldab $E|Y|^p < \infty$. Pane tähele, et juhul kui $p = 1$ või $p = 2$, siis mõjusust garanteerivad koondumiskiirused on λ_n samad, mis klassifitseerimise korral.

Vähimruudud. Lõpetuseks vaatame põgusalt veel klassikalist ruutkaofunktsiooni

$$L(y, h(x)) = |y - h(x)|^2.$$

Ülaltoodud teoreemist järeldub kantregressiooni mõjusus Gaussi tuuma korral. Kasutades aga asjaolu, et iga funktsiooni h korral (veendu!)

$$R(h) = E(Y - h(X))^2 = E(Y - f^*(X))^2 + E(h(X) - f^*(X))^2 = R^* + E(h(X) - f^*(X))^2,$$

kus $f^*(x) = E[Y|X = x]$ on parim võimalik regressioonifunktsioon – tinglik keskväärtus – saame, et $R(h_n)$ koondub arvuks R^* tõenäosuse järgi (või peaaegu kindlasti), kui

$$\int_{\mathbb{R}^d} (h_n(x) - f^*(x))^2 dF(x) \tag{4.8.5}$$

koondub nulliks tõenäosuse järgi (või peaaegu kindlasti). Ruutkaofunktsiooni korral defineeritaksegi mõjusus tihti suuruse (4.8.5) koondumise kaudu (vt [**?**]). Seega võime järeldada, et Gaussi tuumaga kantregressiooni korral L_2-kaugus funktsiooni h_n ja regressioonifunktsiooni f^* vahel koondub tõenäosuse järgi nulliks.

Peatükk 5

Boosting

Eeldus: Alljärgnevas eeldame, nagu ikka, et klassid on märgistatud: +1,-1. Samuti lepime kokku, et $\operatorname{sgn}(0) = 1$.

5.1 Risk ja surrogaatrisk

Tuletame meelde SVM klassifitseerimisülesanet (4.5.16):

$$\min_{h\in\mathcal{H},w_o\in\mathbb{R}} \frac{1}{2}\|h\|^2 + C\sum_{i=1}^{n}\big(1 - y_i(h(x_i) + w_o)\big)_+^p.$$

See on tehisõppes tihti ettetuleval kujul olev optimiseerimismisülesanne ($\lambda = (2nC)^{-1}$):

$$\min_{h\in\mathcal{H},w_o\in\mathbb{R}} \lambda\|h\|^2 + \frac{1}{n}\sum_{i=1}^{n}\phi\big(y_i(h(x_i) + w_o)\big), \tag{5.1.1}$$

kus $\phi(t)$ on mingi kaofunktsioon. Antud juhul siis $\phi(t) = \max(0, 1-t)^p$. Teame (vt (4.2.20)), et ülesande (5.1.1) võib esitada kujul (siin B on konstant, mis sõltub andmetest):

$$\begin{aligned}&\min_{h\in\mathcal{H},w_o\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\phi\big(y_i(h(x_i) + w_o)\big) \\ &\text{nii, et } \|h\| \le B\end{aligned} \tag{5.1.2}$$

Tähistades iga $B > 0$ korral

$$\mathcal{F}_B := \{f : \mathcal{X} \to \mathbb{R} : f = h + w_o, \quad h \in \mathcal{H}, \quad \|h\| \le B, \quad w_o \in \mathbb{R}\}$$

saame probleemile (5.1.2) kuju:

$$\min_{f\in\mathcal{F}_B} \frac{1}{n}\sum_{i=1}^{n}\phi\big(y_i f(x_i)\big). \tag{5.1.3}$$

Seega SVM-klassifitseerimine on üks paljudest meetoodidest, kus funktsionaalset marginaali $y_i f(x_i)$ üritatakse maksimiseerida **empiirilise ϕ-riski**

$$\frac{1}{n}\sum_{i=1}^{n}\phi\big(y_i f(x_i)\big)$$

minimiseerimise läbi. Kui n on piisavalt suur, siis valimi keskmine on ligikaudu võrdne keskväärtusega ja empiiriline ϕ-risk on peaaegu võrdne **ϕ-riskiga**:

$$R_\phi(f) := E\phi(Yf(X)) = \int \phi(yf(x))dF(x,y).$$

Seetõttu võib empiirilise ϕ-riski minimiseerimist (üle klassi $\mathcal{F}$) vaadelda kui ERM-printsiibi rakendamist ϕ-riski minimiseerimiseks (üle klassi $\mathcal{F}$). Nagu ikka, otse me ϕ-riski minimiseerida ju ei saa, sest jaotus $F(x,y)$ on meile tundmata.

ϕ-risk kui surrogaatrisk. Tuletame meelde, et meie eesmärk on klassist $\mathcal{F}$ valida selline f, et klassifitseerija $g = \operatorname{sgn}(f)$ risk

$$R(g) = \mathbf{P}(Y \neq g(X)) = \mathbf{P}(Y \neq \operatorname{sgn}(f(X)) =: R(f)$$

oleks nii väike kui võimalik. Raamatus [13] nimetatakse ϕ-riski **surrogaatriskiks**: eesmärk on minimiseerida klassifitseerimisriski $R(f)$, selle asemel minimiseeritakse surrogaatriski $R_\phi(f)$. Ent kas surrogaatriski minimiseerimine on ikka õige? Teisisõnu: kuidas on seotud riskid $R_\phi(f)$ ja $R(f)$. Täpsemalt: kui R_ϕ^* on minimaalne ϕ-risk üle *kõigi (mõõtuvate) funktsioonide* ja $R_\phi(f_n) \to R_\phi^*$, kas siis kehtib koondumine $R(f_n) \to R^*$? Kui ülaltoodud koondumine kehtib, siis surrogaatriski (peaaegu) minimiseeriv f annab klassifitseerimisriski (peaaegu) minimiseeriva klassifitseerija.

Viimasele küsimusele on lihtne vastata, kui ϕ rahuldab järgmist üsna loomulikku tingimust:

$$\phi(t) \geq I_{\{t\leq 0\}}, \quad \forall t. \tag{5.1.4}$$

ning surrogaatrisk $R_\phi(f_n) \to R_\phi^*$. Tõepoolest, seosest (5.1.4) järeldub, et iga f korral $R(f) \leq R_\phi(f)$. Seega, kui $R_\phi(f_n) \to 0$, siis ka $R(f_n) \to 0$. Kuid viimane olukord (kus Bayesi risk on peaaegu null) on pigem erand kui reegel, mistõttu tingimusest (5.1.4) üksi jääb väheks. Märgime veel, et ϕ-riski minimiseerimine ei sõltu funktsiooni ϕ läbi korrutamisest skalaariga, siis mittekasvava (positiivse) ϕ korral võib üldisust kitsendamata eeldada, et (5.1.4) kehtib. Kokkuvõttes: kuigi ϕ kahanemine loomulik eeldus funktsioonile ϕ, sellest üksi ei piisa.

Kalibreeritud ϕ. Tähistame

$$\eta(x) := \mathbf{P}(Y = 1|X = x) = p(1|x).$$

Seega Bayesi classifitseerija on

$$g^*(x) = \operatorname{sgn}(\eta(x) - 0.5).$$

Iga f korral ϕ-risk (surrogaatrisk) on

$$R_\phi(f) = E\big(E[\phi(Yf(X))|X]\big) = E\big(\phi(f(X))\eta(X) + \phi(-f(X))(1-\eta(X))\big).$$

Seega (täpselt nii nagu Bayesi klassifikatori korral) ϕ-riski minimiseerimine on ekvivalentne *tingliku ϕ-riski*

$$E\big[\phi(Yf(X))|X = x\big] = \eta(x)\phi(f(x)) + \phi(-f(x))(1-\eta(x))$$

minimiseerimisega üle $f(x)$ väärtuste. Tõenäosus $\eta(x)$ on iga x korral fikseeritud ning seega iga $\eta \in [0,1]$ korral meid huvitab minimaalne tinglik ϕ-risk

$$H(\eta) := \inf_{\alpha\in\mathbb{R}} \big(\eta\phi(\alpha) + \phi(-\alpha)(1-\eta)\big).$$

Seega

$$R_\phi^* := \inf_f R_\phi(f) = E\big(H(\eta(X))\big).$$

Defineerime

$$\alpha^*(\eta) := \arg\min_{\alpha\in[-\infty,\infty]} \big(\eta\phi(\alpha) + \phi(-\alpha)(1-\eta)\big)$$

(eeldusel, et miinimium on saavutatav ja α^* on mõõtuv). Seega funktsioon

$$f^*(x) := \alpha^*(\eta(x))$$

küll minimiseerib ϕ-riski üle kõikide (mõõtuvate) funktsioonide, kuid kas ta on aga parim klassifitseerimise mõttes? Teisisõnu, kas f^* abil defineeritud klassifitseerija $g^*(x) = \operatorname{sgn}(f^*(x))$ on Bayesi klassifitseerija? Kui ϕ on selline, et iga $\eta \neq \frac{1}{2}$ korral

$$\alpha^*(\eta) > 0, \quad \text{kui } \eta > \frac{1}{2} \quad \text{ja} \quad \alpha^*(\eta) < 0 \quad \text{kui } \eta < \frac{1}{2}, \tag{5.1.5}$$

siis $g^*(x)$ on Bayesi klassifitseerija. Tingimus (5.1.5) on ekvivalentne järgmise tingimusega

$$\operatorname{sgn}(\alpha^*(\eta)) = \operatorname{sgn}(\eta - 0.5), \quad \forall \eta \neq 0.5. \tag{5.1.6}$$

Tingimus (5.1.6) eeldab α^* olemasolu, samuti sõltub ta $\operatorname{sgn}(0)$ definitsioonist ja α^* valikust (juhul kui see pole ühene). Seetõttu kasutatakse tingimuse (5.1.6) asemel üldisemat tingimust

$$H^-(\eta) > H(\eta), \quad \forall \eta \neq 0.5, \tag{5.1.7}$$

kus $H^-(\eta)$ on funktsiooni $\alpha \mapsto \eta\phi(\alpha) + \phi(-\alpha)(1-\eta)$ miinimum üle argumentide, mille märk erineb $(2\eta - 1)$ märgist. Formaalselt

$$H^-(\eta) := \inf_{\alpha\in\mathbb{R}:\alpha(2\eta-1)\leq 0} \big(\eta\phi(\alpha) + \phi(-\alpha)(1-\eta)\big).$$

Paneme tähele, et kui $\alpha^*(\eta)$ $(\eta \neq 0)$ leidub, siis tingimusest (5.1.7) järeldub (5.1.6) iga α^* korral (miks?).

Funktsioonide H ja H^- omadused.

- $H^-(\frac{1}{2}) = H(\frac{1}{2})$.
- $H : [0,1] \to \mathbb{R}^+$ on nõgus ja sümmeetriline punkti 0.5 suhtes:

$$H(\eta) = H(1-\eta).$$

  Sellest järelduv, et $H(\eta) \leq H(0.5)$;
- $H^- : [0,1] \to \mathbb{R}^+$ on sümmeetriline punkti 0.5 suhtes.
- H^- on nõgus hulkadel $[0, \frac{1}{2}]$ ja $[\frac{1}{2}, 1]$.
- Kui ϕ on kumer, siis $H(\frac{1}{2}) = \phi(0)$.

Ülesanne 5.1 *Tõesta omadused.*

Definitsioon 5.1 *Öeldakse, et ϕ on* **kalibreeritud** *(*classification-calibrated*) kui kehtib (5.1.7): iga $\eta \neq 0.5$ korral $H^-(\eta) > H(\eta)$.*

Seega, kui ϕ on kalibreeritud, ja α^* leidub, siis funktsioon g^* on Bayesi klassifitseerija iga α^* valiku korral.

Näited:

- Olgu $\phi(t) = (1-t)_+$. Siis pole reske veenduda, et iga η korral saavutab funktsioon

$$\eta(1-\alpha)_+ + (1-\eta)(1+\alpha)_+$$

  miinimumi punktides $+1$ või -1. Sellest järeldub vahetult, et

$$H(\eta) = 2\min(\eta, 1-\eta), \quad \alpha^*(\eta) = \begin{cases} 1, & \text{kui } \eta > 0.5; \\ -1, & \text{kui } \eta < 0.5. \end{cases}$$

  Seega ϕ on kalibreeritud. Veendu, et $H^-(\eta) = 1$.
- Olgu $\phi(t) = (1-t)_+^2$.

Ülesanne 5.2 *Veendu, et*

$$H(\eta) = 4\eta(1-\eta), \quad \alpha^*(\eta) = 2\eta - 1.$$

Kas ϕ on kalibreeritud? Leia $H^-(\eta)$.

- Olgu $\phi(t) = \exp[-t]$. Siis

$$H(\eta) = \inf_{\alpha}(\eta e^{-\alpha} + e^{\alpha}(1-\eta)).$$

Ülesanne 5.3 *Veendu, et*

$$\alpha^*(\eta) = \frac{1}{2}\ln\Big(\frac{\eta}{1-\eta}\Big), \quad H(\eta) = 2\sqrt{\eta(1-\eta)}, \quad H^-(\eta) = 1.$$

Kas ϕ on kalibreeritud?

- Olgu $\phi(t) = I_{(-\infty,0]}$.

Ülesanne 5.4 *Veendu, et*

$$H(\eta) = \min(\eta, 1-\eta), \quad H^-(\eta) = \max(\eta, 1-\eta).$$

Kas ϕ on kalibreeritud?

Funktsiooni H omadustest teame, et $H(\eta) \leq H(0.5)$. Järgnev lause näitab, et kui ϕ on kalibreeritud, siis see võrratus on range iga $\eta \neq 0.5$ korral.

Lause 5.1 *Kui ϕ on kalibreeritud, siis*

$$H(\eta) < H(\frac{1}{2}), \quad \textit{kui } \eta \neq \frac{1}{2}.$$

Tõestus. Oletame, et $\alpha^*(0.5)$ leidub. Oletades vastuväiteliselt, et $H(\eta) = H(\frac{1}{2})$, saame

$$\frac{1}{2}\phi(\alpha^*) + \frac{1}{2}\phi(-\alpha^*) = H(\frac{1}{2}) = H(\eta) \leq \eta\phi(\alpha^*) + (1-\eta)\phi(-\alpha^*), \tag{5.1.8}$$

millest järeldub, et

$$\big(\phi(\alpha^*) - \phi(-\alpha^*)\big)(\eta - \frac{1}{2}) \geq 0.$$

Seosest $H(\eta) = H(1-\eta)$ järeldub, et

$$\frac{1}{2}\phi(\alpha^*) + \frac{1}{2}\phi(-\alpha^*) = H(\frac{1}{2}) = H(1-\eta) \leq (1-\eta)\phi(\alpha^*) + \eta\phi(-\alpha^*),$$

millest saame, et

$$\big(\phi(-\alpha^*) - \phi(\alpha^*)\big)(\eta - \frac{1}{2}) \geq 0.$$

Kui $\eta \neq \frac{1}{2}$, saavad mõlemad võrratused kehtida parjasti siis, kui $\phi(-\alpha^*) = \phi(\alpha^*)$. Sellisel juhul saame, et võrratus (5.1.8) on

$$\phi(\alpha^*) = H(\frac{1}{2}) = H(\eta) = \phi(\alpha^*).$$

Et

$$H(\eta) = \min_{\alpha} \big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big) = \phi(\alpha^*) = \phi(-\alpha^*),$$

saame, et α^* ja $-\alpha^*$ on mõlemad miinimumkohad. Seega

$$H(\eta) = \min_{\alpha\le 0} \big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big) = \min_{\alpha\ge 0} \big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big),$$

millest omakorda järeldub, et $H(\eta) = H^-(\eta)$. Vastuolu eeldusega, et ϕ on kalibreeritud. Tõestuse üldistamine juhule kui α^* ei leidu, on lihtne. ■

Kumer ϕ. Enamasti praktikas $\phi(t)$ on kumer, näiteks $(1-t)_+^p$ (SVM), eksponentsiaalne $\exp[-t]$ (AdaBoost), logistiline $\log_2(1+\exp[-2t])$ (logitBoost). Järgnev lause nnäitab, et kumera ϕ korral on kalibreeritust lihtne kontrollida – selleks piisab (ja on tarvilik), kui funktsioonil ϕ on punktis 0 tuletis ja see on rangelt negatiivne.

Lause 5.2 *Kumer ϕ on kalibreeritud parajasti siis kui $\phi'(0) < 0$.*

Tõestus. Olgu ϕ kumer. Siis iga η korral on kumer ka funktsioon

$$C_\eta(\alpha) := \eta\phi(\alpha) + (1-\eta)\phi(-\alpha).$$

Kumeral funktsioonil on alati miinimum, seega $\alpha^*(\eta)$ eksisteerib. Olgu $\phi'(0) < 0$. Siis

$$C'_\eta(0) = \eta\phi'(0) - (1-\eta)\phi'(0) = (2\eta-1)\phi'(0).$$

Kui $\eta > \frac{1}{2}$, siis $(2\eta-1)\phi'(0) < 0$ ehk $C'_\eta(0) < 0$. Et C_η on kumer ja tema tuletis punktis 0 on negatiivne, järeldub sellest, et miinimumkoht α^* on reaaltelje positiivsemas osas ehk $\alpha(\eta)(2\eta-1) > 0$. Kui $\eta < \frac{1}{2}$, siis $C'_\eta(0) > 0$ ning kumerusest järeldub, et miinimumkoht α^* on reaaltelje endatiivsemal poolel ning jällegi $\alpha(\eta)(2\eta-1) > 0$. Seega ϕ on kalibreeritud. Teistpidi: esitame tõestuse idee: Kui $\eta = \frac{1}{2}$, on see $C_\eta(\alpha)$ sümmeetriline ja 0 on miinimumkoht. Kui ϕ on punktis 0 diferentseeruv, siis η muutmisel, liigub miinimumkoht 0-punktist eemale. Kui $\phi'(0) < 0$, liigub miinimumkoht "õiges suunas". Kui aga ϕ punktis 0 diferentseeruv, saab leida $\eta \neq 0.5$ nii, et C_η miinimumpunkt on ikka 0. See on vastuolus kalibreerituse definitsiooniga. ■

Järeldus 5.1.1 *Olgu ϕ kumer ja kalibreeritud. Siis iga $\eta \in [0,1]$ korral*

$$H^-(\eta) = \phi(0).$$

Tõestus. Funktsiooni ϕ kumerusest järeldub, et

$$\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \ge \phi(\alpha(2\eta-1))$$

Seega

$$\min_{\alpha:\alpha(2\eta-1)\le 0} \big(\eta\phi(\alpha) + \phi(-\alpha)(1-\eta)\big) \ge \min_{\alpha:\alpha(2\eta-1)\le 0} \phi(\alpha(2\eta-1)\big) = \min_{u\le 0} \phi(u) = \phi(0),$$

sest $\phi'(0) < 0$ tähendab, et ϕ on kahanev piirkonnas $(-\infty, 0]$. Siit saame, $H^- = \phi(0)$. ■

Funktsioon ψ. Defineerime $\psi : [0,1] \to \mathbb{R}^+$,

$$\psi(t) := H^-(\frac{1+t}{2}) - H(\frac{1+t}{2})$$

ψ omadused:

- mittenegatiivne (miks?);
- $\psi(0) = 0$ (miks?);
- pidev (sest nõgus funktsioon on pidev ja ψ on kahe nõgusa funktsiooni vahe);

Kui ϕ on kumer ja kalibreeritud, siis (lisaks eelmistele omadustele)

- $\psi(t) = \phi(0) - H(\frac{1+t}{2})$, sest Järelduse 5.1.1 põhjal $H^-(\eta) = \phi(0)$.
- ψ on kumer, sest $H(\frac{1+t}{2})$ nõgus;
- $\psi(t) > 0$ iff $t > 0$ (järeldub Lausest 5.1).

Ülesanne 5.5 *Olgu ϕ kumer ja kalibreeritud, olgu $\{t_n\}$ reaalarvude jada. Näita, et*

$$\psi(t_n) \to 0 \quad \Leftrightarrow \quad t_n \to 0 \tag{5.1.9}$$

Näited.

- Olgu $\phi(t) = (1-t)_+$. Siis $H(\eta) = 2\min(\eta, 1-\eta)$, millest

$$\psi(t) = \phi(0) - H(\frac{1+t}{2}) = 1 - 2\min(\frac{1+t}{2}, 1 - \frac{1+t}{2}) = 1 - (1-t) = t,$$

- Olgu $\phi(t) = (1-t)_+^2$. Siis $H(\eta) = 4\eta(1-\eta)$, millest

$$\psi(t) = \phi(0) - H(\frac{1+t}{2}) = 1 - 4(\frac{1+t}{2})(\frac{1-t}{2}) = 1 - (1-t^2) = t^2.$$

- Olgu $\phi(t) = \exp[-t]$. Siis $H(\eta) = 2\sqrt{\eta(1-\eta)}$,

$$\psi(t) = 1 - 2\sqrt{(\frac{1+t}{2})(\frac{1-t}{2})} = 1 - \sqrt{1-t^2}.$$

- Olgu $\phi(t) = I_{(-\infty,0]}$. Siis $H(\eta) = \min(\eta, 1-\eta)$, $H^-(\eta) = \max(\eta, 1-\eta)$ ja

$$\psi(t) = \frac{1+t}{2} - \frac{1-t}{2} = t.$$

Kalibreerimisvõrratus. Seega, kui ϕ on kalibreeritud, on ϕ-riski R_ϕ minimiseerimine võrdväärne riski minimiseerimisega. Oletame nüüd aga, et mingi f korral ϕ risk $R_\phi(f)$ on peaaegu võrdne miinimumiga R_ϕ^*. Kui suur on aga $R(f) - R^*$? Järgnev teoreem annab vastuse.

Teoreem 5.2 *(Bartlett, Jordan, McAuliffe, 2004) Olgu ϕ kumer ja kalibreeritud. Siis iga f korral*

$$\psi\big(R(f) - R^*\big) \le R_\phi(f) - R_\phi^*. \tag{5.1.10}$$

Seega, kui ϕ on kumer ja kalibreeritud, siis koondumisest $R_\phi(f_n) \to R_\phi(f)$ järeldub, et $\psi(R(f_n) - R^*) \to 0$. Seosest (5.1.9) saame, et koondumine $\psi(R(f_n) - R^*) \to 0$ kehtib parajasti siis, kui $R(f_n) \to R^*$. Kokkuvõttes:

$$R_\phi(f_n) \to R_\phi^* \quad \Rightarrow \quad R(f_n) \to R^*.$$

Märkused:

1. Võrratust (5.2) ei saa üldiselt parandada, sest kehtib lemma ([18, 19], Thm 3).

   Lemma 5.1 *Olgu $|\mathcal{X}| \ge 2$. Siis iga mittenegastiivse ϕ iga $\theta \in [0,1]$ ja iga $\epsilon > 0$ korral leidub tõenäosusmõõt ruumil $\mathcal{X} \times \{-1,1\}$ ja funktsioon $f : \mathcal{X} \to \mathbb{R}$ nii, et $R(f) - R^* = \theta$, kuid $\psi(\theta) \le R_\phi(f) - R_\phi^* \le \psi(\theta) + \epsilon$.*

2. Tõestasime, et võrratus (5.1.10) kehtib, kui ϕ on kumer ja kalibreeritud. Tegelikult kehtib toodud võrratus iga mittenegatiivse ϕ korral, kuid sellisel juhul on ψ defineeritud mõnevõrra üldisemalt. Ka sellisel juhul kehtib $\psi(0) = 0$, kuid

   $$\psi(t) > 0 \quad \forall t > 0 \quad \Leftrightarrow \quad \phi \text{ on kalibreeritud.}$$

   Seega, kui ϕ pole kalibreeritud, siis võrdusest $R_\phi(f) = R_\phi^*$ järeldub $\psi(R(f) - R_\phi^*) = 0$, kuid see ei tähenda, et $R(f) = R^*$ (vt [18, 19]).

Kirjandus: Kalibreerimisteooriast loe [18, 19] või [13], Ch. 3.

5.1.1 Kalibreerimisvõrratuse tõestus*

Lemma 5.2 *Olgu $g : \mathcal{X} \to \{-1,1\}$ klassifikaator. Siis*

$$R(g) - R^* = E\Big(|2\eta(X) - 1| I_{\{g(X) \ne g^*(X)\}}\Big) = \int_{\{g(x) \ne g^*(x)\}} |2\eta(x) - 1| h(x) dx, \tag{5.1.11}$$

kus g^* on Bayesi klassifikaator ja h tunnuse X tihedus (mingi mõõdu suhtes, mida tähistame dx).

Tõestus. Leiame klassifikaatori g tingliku riski tingimusel $X = x$:

$$\begin{aligned}\mathbf{P}[g(X) \neq Y|X = x] &= 1 - \mathbf{P}[g(X) = Y|X = x] \\ &= 1 - \Big(\mathbf{P}[Y = 1, g(X) = 1|X = x] + \mathbf{P}[Y = -1, g(X) = -1|X = x]\Big) \\ &= 1 - \Big(I_{\{g(x)=1\}}\mathbf{P}[Y = 1|X = x] + I_{\{g(x)=-1\}}\mathbf{P}[Y = -1|X = x]\Big) \\ &= 1 - \Big(I_{\{g(x)=1\}}\eta(x) + I_{\{g(x)=-1\}}(1 - \eta(x))\Big)\end{aligned}$$

Seega

$$\begin{aligned}&\mathbf{P}[g(X) \neq Y|X = x] - \mathbf{P}[g^*(X) \neq Y|X = x] \\ &= \eta(x)\big(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\big) + (1 - \eta(x))\big(I_{\{g^*(x)=-1\}} - I_{\{g(x)=-1\}}\big) \\ &= \big(2\eta(x) - 1\big)\big(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\big) \\ &= |2\eta(x) - 1|I_{\{g^*(x)\neq g(x)\}}.\end{aligned}$$

Keskmistades üle x-de, saame

$$\begin{aligned}R(g) - R(g^*) &= \int \Big(\mathbf{P}[g(X) \neq Y|X = x] - \mathbf{P}[g^*(X) \neq Y|X = x]\Big)h(x)dx \\ &= \int_{\{g(x)\neq g^*(x)\}} |2\eta(x) - 1|h(x)dx.\end{aligned}$$

■

Lause 5.3 *Olgu ϕ on kalibreeritud ning kumer. Kui $\alpha(\eta - \frac{1}{2}) < 0$, siis*

$$\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha) \geq H(\frac{1}{2}).$$

Tõestus. Kui ϕ kumer ja kalibreeritud, siis $\phi'(0) < 0$, millest $\phi(s) \geq \phi(0)$ iga $s < 0$ korral; ϕ kumerusest saame

$$\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha) \geq \phi(\alpha(2\eta - 1)) \geq \phi(0) = H(0.5).$$

■

Teoreemi 5.2 tõestus. Olgu

$$g(x) = \operatorname{sgn}(f(x)), \quad g^*(x) = \operatorname{sgn}(\eta(x) - 0.5).$$

Meid huvitab ülemine hinnang suurusele

$$\psi\big(R(g) - R(g^*)\big) = \psi\big(E(I_{\{g(X)\neq g^*(X)\}}|2\eta(X) - 1|)\big) = \psi\Big(\int_{\{g(x)\neq g^*(x)\}} |2\eta(x) - 1|)h(x)dx\Big).$$

Et ψ on kumer, siis Jenseni võrratusest saame

$$\psi\big(E(I_{\{g(X)\neq g^*(X)\}}|2\eta(X)-1|)\big) \leq E\psi\big(I_{\{g(X)\neq g^*(X)\}}|2\eta(X)-1|)\big) = EI_{\{g(X)\neq g^*(X)\}}\psi(|2\eta(X)-1|),$$

kus viimane võrratus tuleb sellest, et $\psi(0) = 0$. Vastavalt ψ definitsioonile ja H sümmeetrilisusele, saame

$$\psi(|2\eta(x) - 1|) = H(\frac{1}{2}) - H(\eta(x)),$$

millest

$$\begin{aligned} EI_{\{g(X)\neq g^*(X)\}}\psi(|2\eta(X) - 1|) &= \int_{\{g(x)\neq g^*(x)\}} \psi(|2\eta(x) - 1|)h(x)dx \\ &= \int_{\{g(x)\neq \mathrm{sgn}(2\eta(x)-1)\}} \big(H(\frac{1}{2}) - H(\eta(x)\big)h(x)dx. \end{aligned}$$

Nüüd paneme tähele, et kui $g(x) = \mathrm{sgn}(f(x)) \neq \mathrm{sgn}(2\eta(x) - 1)$, siis lausest 5.3 järeldub, et

$$\eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)) = E[\phi(Yf(X)|X = x] \geq H(\frac{1}{2}).$$

Samas iga x korral

$$\eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)) = E[\phi(Yf(X)|X = x] \geq H(\eta(x)),$$

millest

$$\begin{aligned} &\int_{\{g(x)\neq \mathrm{sgn}(2\eta(x)-1)\}} \big(H(\frac{1}{2}) - H(\eta(x)\big)h(x)dx \leq \\ &\int_{\{g(x)\neq \mathrm{sgn}(2\eta(x)-1)\}} \big(E[\phi(Yf(X)|X = x] - H(\eta(x)\big)h(x)dx \leq \\ &\int \big(E[\phi(Yf(X)|X = x] - H(\eta(x)\big)h(x)dx = R_\phi(f) - R^*_\phi. \end{aligned}$$

5.2 AdaBoost

5.2.1 Boosting ja AdaBoost: põhimõte

Oletame, et uurija käsutuses on suhteliselt lihtne ja kergesti rakendatav klassifitseerimisreegel (nn. rusikareegel), mis andmetel treenituna annab ehk vaid natuke parema tulemuse kui huupi (andmetest sõltumatu) klassifitseerimine. Sellise klassifitseerimiseeskirja kasutamine otsuste tegemisel tähendab enamasti suurt riski. Klassifitseerimismeetodi **boosting (võimendamine)** idee seisneb rusikareegli (*weak learner, base learner*) korduvas rakendamises esialgse valimi "modifitseeritud"(ümberkaalutud) variandile. Nii saadud rusikareeglite jada $h_1, \ldots, h_T$, $h_t : \mathcal{X} \to \{-1, 1\}$ kombineeritakse lõpp-klassifitseerijaks kujul

$$g(x) = \mathrm{sgn}\big(\sum_{t=1}^{T} \alpha_t h_t(x)\big), \tag{5.2.1}$$

kus α_t on sobivalt valitud kaalud. Reegli h_t abil defineeritakse kaalud $D_{t+1}(i), i = 1, \ldots, n$ ning nõnda kaalutud valimi abil treenitakse reegal h_{t+1}. Selgub, et kuigi omaette võetuna on iga reegel h_t keskpärane, võib nende (võimendatud) kombinatsioon (5.2.1) olla väga hea klassifitseerija.

Esimesed boosting-tüüpi algoritmid pakuti välja 1980.-ndate aastate lõpus (Kearns, Valiant, Shapire). Need esimesed algoritmid polnud kuigi praktilised, kuid pakkusid elavat teoreetilist huvi. 1995.-l aastal esitasid Y. Freud ja R. Shapire vast kõige tuntuma boosting-tüüpi algoritmi nn. AdaBoost (adaptive boosting), millel on head teoreetilised omadused ning mis väidetavalt on ka hästi rakendatav.

AdaBoost: Antud (valitud) lihtne klassifitseerimisreegel (rusikareegel) h.

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T.
2. Olgu $D_1(i) = \frac{1}{n}, \quad i = 1, \ldots, n$;
3. **Iga** $t = 1, \ldots, T$ **korral:**

   a Treeni rusikareegel kaalutud valimi (S, D_t) korral. Nii saad (lihtsa) reegli

   $$h_t : \mathcal{X} \to \{-1, 1\}.$$

   b Leia kaalutud treeningviga

   $$\epsilon_t := \sum_{i=1}^{n} D_t(i) I_{\{y_i \neq h_t(x_i)\}}.$$

   c Defineeri

   $$\alpha_t := \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}.$$

   d Defineeri uued kaalud

   $$D_{t+1}(i) := \frac{D_t(i) \exp[-\alpha_t y_i h_t(x_i)]}{Z_t},$$

   kus Z_t on normaliseeriv konstant.
4. **Lõpeta kui:** $\epsilon_t = 0$ või $\epsilon_t \geq \frac{1}{2}$; sellisel juhul $T = t - 1$.
5. **Väljund:**

   $$g(x) := \operatorname{sgn}\big(f_T(x)\big), \quad f_T = \sum_{t=1}^{T} \alpha_t h_t.$$

Seega esimesel sammul treenitakse rusikareegel h originaalvalimi põhjal, sest $D_1(i) = \frac{1}{n}$. Klassifitseerija h_t treeningvea ϵ_t põhjal leitakse reegli h_t kaal α_t (et iga t korral võime eeldada $\epsilon_t \leq 0.5$, siis $\alpha_t \geq 0$). Mida väiksem on ϵ_t, seda suurem on α_t. Pärast klassifitseerija h_t ja kaalu α_t leidmist leitakse uued valimikaalud $D_{t+1}(i)$. Seejuures toimitakse järgmiselt: iga i korral võetakse arvesse selle valimi punkti vana kaal $D_t(i)$ ning korrutatakse see läbi teguriga

$$\exp[-\alpha_t y_i h_t(x_i)] = \begin{cases} e^{\alpha_t} > 1, & \text{kui } h_t \text{ klassifitseerib punkti } x_i \text{ valesti ;} \\ e^{-\alpha_t} < 1, & \text{kui } h_t \text{ klassifitseerib punkti } x_i \text{ õieti.} \end{cases}$$

Seega valesti klassifitseeritud punktid saavad suurema kaalu, mistõttu h_{t+1} treenimisel tuleb nende punktidega rohkem arvestada. Samuti sõltub ümberkaalumine konstandist α_t – mida suurem on α_t (mida korrektsem on h_t), seda suurem on valesti klassifitseeritud elementide suhteline kaal (võrreldes õiesti klassifitseeritud elementide kaaluga).

Märkused:

1. Tihti (arvutipakettides) loobutakse nõudest $\epsilon_t < \frac{1}{2}$. Sellisel juhul võib α_t olla ka negatiivne, st $\alpha_t < 0$. Pane tähele, et siis AdaBoost sisuliselt asendab vastava klassifitseerija h_t vastasmärgiga klassifitseerijaga $-h_t$. Klassifitseerijale $-h_t$ vastav empiiriline viga on $1 - \epsilon_t$ (sest kaalude summa on 1) ning temale vasta α_t on seega esialgse negatiivse α_t absoluutväärtus $|\alpha_t| \geq 0$. Kokkuvõttes: kui $\alpha_t < 0$, siis $\alpha_t h_t = |\alpha_t|(-h_t)$, mistõttu nõudest $\epsilon_t \leq \frac{1}{2}$ võib ka loobuda. Küll on loomulik lõpetada, kui $\epsilon_t = \frac{1}{2}$, sest siis $\alpha_t = 0$ ning ümberkaalumist ei toimu ja järgmisel iteratsioonil on kõik samamoodi.

2. Paneme tähele, et
$$-y_i h_t(x_i) = 2I_{\{y_i \neq h_t(x_i)\}} - 1,$$
millest
$$\exp[-\alpha_t y_i h_t(x_i)] = \exp[2\alpha_t I_{\{y_i \neq h_t(x_i)\}}] \exp[-\alpha_t].$$
Järelikult võib uued kaalud $D_{t+1}(i)$ defineerida ka järgmiselt (vt. näiteks [7])
$$D_{t+1}(i) := \frac{D_t(i) \exp[2\alpha_t I_{\{y_i \neq h_t(x_i)\}}]}{Z_t}. \tag{5.2.2}$$

3. Ülaltoodud algoritm on nn. diskreetne AdaBoost, sest me eeldasime, et rusikarregel h on klassifitseerija, st $h : \mathcal{X} \to \{-1, 1\}$. Tihti on aga h väljund hulk $[0, 1]$ (sellisel juhul võib h-d vaadelda kui tõenäosuse hinnangut) või koguni $\mathbb{R}$ ja sellisel juhul räägitakse reaalsest AdaBoostist. Reaalse AdaBoosti korral pole aga ümberkaalumine
$$D_{t+1}(i) \propto D_t(i) \exp[-\alpha_t y_i h_t(x_i)]$$
ekvivalentne seosega (5.2.2).

5.2.2 AdaBoost ja eksponentsiaalne kadu

Veendume, et kui h_t on leitud ERM-meetodil, siis AdaBoost on algoritm, mis teatud viisil minimiseerib empiirilist ϕ-riski üle klassi $\mathcal{F}$, kus $\phi(t) = \exp[-t]$ ja

$$\mathcal{F} := \{\sum_{t=1}^{T} \alpha_t h_t : \alpha_t \in \mathbb{R},\ h_t \in \mathcal{H}\}.$$

Siin $\mathcal{H}$ on kõigi baasklassifitseerijate (rusikareeglite) hulk. Järelikult on kõnealune empiiriline ϕ-risk kujul

$$\sum_{i=1}^{n} \exp[-y_i f(x_i)] = \sum_{i=1}^{n} \exp[-y_i(\alpha_1 h_1(x_i) + \cdots + \alpha_T h_T(x_i))], \tag{5.2.3}$$

ja minimiseerimine toimub üle kõikide $\alpha_1, \ldots, \alpha_T$ ja $h_1, \ldots, h_T$. Eeldame, et $\mathcal{H}$ on kinnine märgi vahetamise suhtes: kui $h \in \mathcal{H}$, siis ka $-h \in \mathcal{H}$. Siis on iga valimi korral treeningviga maksimaalselt 0.5. Praktikas on see nõue alati täidetud.

Funktsiooni (5.2.3) minimiseerimine üle kõikvõimalike funktsioonide kujul $\alpha_1 h_1 + \cdots + \alpha_T h_T$ on keeruline ning AdaBoost toimib järgmiselt: esimesel sammul leitakse funktsiooni

$$\sum_{i=1}^{n} \exp[-y_i \alpha h(x_i)]$$

minimiseerivad $h_1 \in \mathcal{H}$ ja $\alpha_1 \in \mathbb{R}$.
Sammul t leiab AdaBoost α_t ja $h_t \in \mathcal{H}$ nii, et

$$(h_t, \alpha_t) = \arg\min_{\alpha,h} \sum_{i=1}^{n} \exp[-y_i(f_{t-1}(x_i) + \alpha h(x_i))], \tag{5.2.4}$$

kus

$$f_{t-1} = \sum_{s=1}^{t-1} \alpha_s h_s, \quad f_0 = 0.$$

Seega minimiseerib AdaBoost funktsiooni (5.2.3) järk-järgult: esimesel sammul leitakse $\alpha_1 h_1$, teisel sammul leitakse $\alpha_2 h_2$ (nii, et $\alpha_1 h_1$ jääb samaks), kolmandal sammul $\alpha_3 h_3$ (nii, et $\alpha_1 h_1 + \alpha_2 h_2$ jääb samaks) jne. Veendume selles.

Kõigepealt tuletame meelde, et

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \exp[-\alpha_t y_i h_t(x_i)], \quad Z_t := \sum_i D_t(i) \exp[-\alpha_t y_i h_t(x_i)].$$

Seega iga i korral

$$D_1(i) = \frac{1}{n},\ D_2(i) = \frac{\exp[-\alpha_1 y_i h_1(x_i)]}{nZ_1},\ D_3(i) = \frac{\exp[-y_i(\alpha_1 h_1(x_i) + \alpha_2 h_2(x_i))]}{nZ_1 Z_2},$$
$$D_t(i) = \frac{\exp[-y_i(\alpha_1 h_1(x_i) + \alpha_2 h_2(x_i) + \cdots + \alpha_{t-1} h_{t-1}(x_i))]}{Z_0 Z_1 \cdots Z_{t-1}} = \frac{\exp[-y_i f_{t-1}(x_i)]}{Z_0 Z_1 Z_2 \cdots Z_{t-1}},$$

kus $Z_0 := n$. Järelikult

$$\exp[-y_i f_{t-1}(x_i)] = \prod_{s=0}^{t-1} Z_s D_t(i) = n \prod_{s=1}^{t-1} Z_s D_t(i). \tag{5.2.5}$$

Seosest (5.2.5) saame, et

$$\begin{aligned}\sum_{i=1}^{n} \exp[-y_i(f_{t-1}(x_i) + \alpha h(x_i))] &= \sum_{i=1}^{n} \exp[-y_i(f_{t-1}(x_i))] \exp[-y_i \alpha h(x_i)] \\ &= \prod_{s=0}^{t-1} Z_s \Big(\sum_{i=1}^{n} D_t(i) \exp[-y_i \alpha h(x_i)] \Big).\end{aligned}$$

Vaatleme nüüd optimiseerimisülesannet

$$\min_{\alpha, h} \sum_{i=1}^{n} D_t(i) \exp[-y_i \alpha h(x_i)]. \tag{5.2.6}$$

Selle ülesande saab lahendada nii, et iga α korral leiame parima h (mis üldiselt sõltub α-st) ka siis minimiseerime korrutist üle α-de. Paneme tähele, et iga $\alpha > 0$ korral (5.2.6) minimiseerimine on ekvivalentne

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{n} D_t(i) I_{\{h(x_i) \neq y_i\}} \tag{5.2.7}$$

ehk (5.2.6) minimiseerimine on ekvivalentne kaalutud empiirilise riski minimiseerimisega. Seda on kerge näha, sest

$$\begin{aligned}\sum_{i=1}^{n} D_t(i) \exp[-y_i \alpha h(x_i)] &= e^{-\alpha} \sum_{i: h(x_i) = y_i} D_t(i) + e^{\alpha} \sum_{i: h(x_i) \neq y_i} D_t(i) \\ &= (e^{\alpha} - e^{-\alpha}) \sum_{i=1}^{n} D_t(i) I_{\{y_i \neq h(x_i)\}} + e^{-\alpha} \sum_{i=1}^{n} D_t(i).\end{aligned}$$

Seega, sõltumata $\alpha > 0$ väärtusest, on h_t selline, mis minimiseerib (kaalutud) empiirilist riski ehk

$$h_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} D_t(i) I_{\{h(x_i) \neq y_i\}}. \tag{5.2.8}$$

Kui $\alpha < 0$, siis parim h on $-h_t$, kus h_t minimiseerib empiirilist risk, st on defineeritud seosega (5.2.8). Järgnev ülesanne näitab, et parim α on tõepoolest selline nagu AdaBoost'is.

Ülesanne 5.6 *Olgu h_t antud seosega (5.2.8). Asetades see seosesse (5.2.6) ja minimiseerides üle α, veendu, et lahend on kujul*

$$\alpha_t := \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}. \tag{5.2.9}$$

Märkus: Seega, kui h_t on saadud (kaalutud) valimi empriirilise riski minimiseerimisel, siis AdaBoost minimiseerib ϕ-riski üle klassid $\mathcal{F}_T$ teataval (ahnel) moel. Eelmises peatükis nägime, et kaofunktsioon $\phi(t) = \exp[-t]$ on kalibreeritud. Muidugi, AdaBoost üldiselt ei nõua, et rusikareegel h_t leitakse just (kaalutud) empiirilise riski minimiseerimisel, kuid funktsiooni h_t võib käsitleda kui empiirilist riski minimiseeriva rusikareegli (teatavas mõttes) lähendit. Mõnikord siiski defineeritakse AdaBoost nii, et h_t minimiseerib empiirilist riski.

5.2.3 AdaBoosti treeningviga ja marginaalviga

Treeningvigade arvu (empiirilise riski) hinnang. Olgu f_T AdaBoosti tulemus, st $f_T = \alpha_1 h_1 + \cdots + \alpha_T h_T$. Tuleta meelde, et

$$Z_t = \sum_i D_t(i) \exp[-\alpha_t y_i h_t(x_i)]$$

ning (5.2.5):

$$D_t(i) = \frac{\exp[-y_i(\alpha_1 h_1(x_i) + \alpha_2 h_2(x_i) + \cdots + \alpha_{t-1} h_{t-1}(x_i))]}{Z_0 Z_1 \cdots Z_{t-1}} = \frac{\exp[-y_i f_{t-1}(x_i)]}{Z_0 Z_1 Z_2 \cdots Z_{t-1}},$$

millest saame, et

$$(n \prod_{t=1}^{T} Z_t) \sum_i D_{T+1}(i) = (n \prod_{t=1}^{T} Z_t) = \sum_i \exp[-y_i f_T(x_i)],$$

millest saame treeningveale ilusa hinnangu:

$$R_n(f_T) = \frac{1}{n} |\{i : \operatorname{sgn}(f_T(x_i)) \neq y_i\}| \leq \frac{1}{n} \sum_i \exp[-y_i f_T(x_i)] = \prod_{t=1}^{T} Z_t.$$

Ülaltoodud avaldisest on näha, et kui igal sammul t minimiseerime

$$Z_t = \sum_i D_t(i) \exp[-\alpha_t y_i h_t(x_i)]$$

(aga, nagu teame, just seda AdaBoost teeb), siis seeläbi minimiseerime ka treeningviga.

Ülesanne 5.7 *Tõestada, et*

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}. \tag{5.2.10}$$

Seega

$$R_n(f_T) \leq \prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} = \prod_t \sqrt{1-4u_t^2} \leq \exp[-2\sum_t u_t^2], \tag{5.2.11}$$

kus

$$u_t := \frac{1}{2} - \epsilon_t.$$

Et $\mathcal{H}$ on kinnine märgi vahetamise suhtes, siis maksimaalne treeningviga on 0.5. Arv u_t mõõdab seda, kuivõrd on rusikareegel h_t on parem halvimast võimalikust klassifitseerijast. Juhul, kui $\sum_{t=1}^{\infty} u_t^2 = \infty$, läheneb treeningvea hinnang nullile. See tähendab, et mingist kohast alates on treeningviga 0 (miks?). Piisav tingimus selleks on, et leidub $u_o > 0$ nii, et $u_t \geq u_o$ iga t korral. Sellisel juhul

$$R_n(f_T) \leq e^{-(2u_o^2)T}$$

ehk treeningviga koondub T kasvamisel nulliks eksponentsiaalse kiirusega. Kas selline u_o leidub või mitte, sõltub nii valimist kui rusikareeglist (hulgast $\mathcal{H}$). Tihti on aga nii, et suvaliselt kaalutud suvalise valimi korral leidub h mille empiiriline viga on väiksem kui $\frac{1}{2} - u_o$. Ning teinekord pole u_o garanteerivat reeglit vaja otsida empiirilise riski minimiseerimisel (mis, nagu teame, on raske ülesanne), vaid saab ka kergemini. Näiteks lineaarsete klassifitsserijate korral iga kordusteta valimi korral $u_o > \frac{c}{n}$ (c on mingi konstant) ning sellist viga garanteeriva h saab leida polünomiaalse keerukusega [20, 21].

Margnaalvigade arvu hinnang. Nii või teisiti tuleb empiirilise riski koondumist praktikas tihti ette, mistõttu tekib küsimus optimaalsest iteratsioonide arvust T – see ei tohi olla liiga väike (f_T liiga "lihtne") ega ka liiga suur (f_T liiga "keeruline"). Samas tuleb silmas pidada, et $R_n(f_t) = 0$ ei pruugi tähendada, et $\epsilon_t = 0$, sest ϵ_t on rusikareegli h_t (kaalutud) treeningviga, $R_n(f_t)$ on aga $\mathrm{sgn}(f_t)$ (esialgne, st ümberkaalumata) treeningviga. Nii võib algoritm tööd jätkata ka peale seda kui valim on eraldatud – algoritmi edasine töö $y_i f_t(x_i)$ kasvatamise suunas (tuletame meelde, et AdaBoost minimiseerib $\sum_i \exp[-y_i f_t(x_i)]$ ja see tähendab $y_i f_t(x_i)$ maksimiseerimist). Juhul, kui $y_i f_t(x_i) = y_i(\alpha_1 h_1(x_i) + \cdots + \alpha_t h_t(x_i)) > 0$, võib $y_i f_t(x_i)$ kasvatada sakaladide α_i läbikorrutamisel positiivse konstandiga ja seetõttu pakub huvi suurus

$$\rho_i(f_T) := \frac{y_i f_T(x_i)}{\sum_{t=1}^T \alpha_t} = y_i\Big(\sum_{t=1}^T \beta_t f_t(x_i)\Big), \quad \beta_t := \frac{\alpha_t}{\sum_{t=1}^T \alpha_t}, \tag{5.2.12}$$

mida nimetatakse punkti (x_i, y_i) **funktsionaalseks marginaaliks**. Tuletame meelde, et $\alpha_t > 0$ iga t korral ja nii võib suurust $f_T / \sum_t \alpha_t$ vaadelda adaBoosti väljundina ja kasutada klassifitseerimisel. Funktsiooni $f_t(x_i)$ jagamist summaga $\sum_{i=1}^t \alpha_i$ võib vaadelda kui normeerimist (l_1-mõttes).

Selgub, et isegi kui esialgse valimi treeningviga on 0, jätkab AdaBoost marginaalide (5.2.12) suurendamist. Seda näitab järgmine teoreem.

Teoreem 5.3 *Olgu $f_T = \alpha_1 h_1 + \cdots + \alpha_T h_T$ AdaBoosti väljund. Siis iga $\gamma \geq 0$ korral*

$$\frac{1}{n}|\{i : \rho_i(f_T) \leq \gamma\}| \leq \prod_{t=1}^T 2\sqrt{\epsilon_t^{1-\gamma}(1-\epsilon_t)^{1+\gamma}} = \prod_{t=1}^T (1-2u_t)^{\frac{1-\gamma}{2}}(1+2u_t)^{\frac{1+\gamma}{2}}, \tag{5.2.13}$$

kus $\rho_i(f_T)$ on defineeritud seosega (5.2.12).

Tõestus. Paneme tähele:

$$\rho_i(f_T) \le \gamma \quad \Leftrightarrow \quad y_i f_T(x_i) - \gamma \sum_i \alpha_i \le 0 \quad \Leftrightarrow \quad \exp[\gamma \sum_i \alpha_i] \exp[-y_i f_T(x_i)] \ge 1.$$

Seega

$$\exp[\gamma \sum_i \alpha_i] \exp[-y_i f_T(x_i)] \ge I_{(-\infty,\gamma]}(\rho_i(f_T)) \tag{5.2.14}$$

Summeerides (7.3.24) mõlemad pooled üle i ja jagades n-ga, saame

$$\exp[\gamma \sum_i \alpha_i] \frac{1}{n} \sum_i \exp[-y_i f_T(x_i)] \ge \frac{1}{n} |\{i : \rho_i(f_T) \le \gamma\}|$$

Arvestades, et $(n \prod_{t=1}^T Z_t) = \sum_i \exp[-y_i f_T(x_i)]$, saame

$$\frac{1}{n} |\{i : \rho_i(f_T) \le \gamma\}| \le \prod_{t=1}^T Z_t \exp[\gamma \sum_i \alpha_i].$$

Kordaja α_t definitsioonist

$$\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$$

saame

$$\exp[\gamma \sum_i \alpha_i] = \prod_{t=1}^T \big(\frac{1-\epsilon_t}{\epsilon_t}\big)^{\frac{\gamma}{2}}.$$

Arvestades, et $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$, saame nüüd

$$\prod_{t=1}^T Z_t \exp[\gamma \sum_i \alpha_i] = \prod_{t=1}^T 2\big(\frac{1-\epsilon_t}{\epsilon_t}\big)^{\frac{\gamma}{2}} \big(\epsilon_t(1-\epsilon_t)\big)^{\frac{1}{2}} = \prod_{t=1}^T 2\sqrt{\epsilon_t^{1-\gamma}(1-\epsilon_t)^{1+\gamma}}.$$

Viimane võrdus ahelas (5.2.13) tuleneb u_t definitsioonist: $u_t = \frac{1}{2} - \epsilon_t$. ■

Märkus. Võttes $\gamma = 0$, same võrratusest (5.2.13) eespooltõestatud võrratuse (5.2.11), sest

$$\{i : \mathrm{sgn}(f_T(x_i)) \ne y_i\} \subset \{i : \rho_i(f_T) \le 0\}.$$

Veendu, et kui $u_t \ge u_0 > 0$, siis (5.2.13) koondub eksponentsiaalselt nulliks iga piisavalt väikese γ korral.

5.2.4 AdaBoosti riski hinnang

Olgu $f : \mathbb{R}^d \to \mathbb{R}$ ja $\gamma < 0$ fikseeritud marginaal. Tuleta meelde suurus $A_n(f)$ alampeatükist 4.3:

$$A_n(f) = \frac{1}{n} \sum_{i=1}^n \phi_\gamma(-f(x_i)y_i), \quad \phi_\gamma(t) = \begin{cases} 1, & \text{kui } t \ge 0; \\ 1 + \frac{t}{\gamma}, & \text{kui } -\gamma \le t \le 0; \\ 0, & \text{kui } t < -\gamma \end{cases}$$

Siin $\gamma > 0$ on fikseeritud. Seega $A_n(f)$ on empiiriline ϕ-risk, kus $\phi(t) = \phi_\gamma(-t)$. Samuti tuletame meelde, et $A_n(f)$ ülemine tõke on marginaalvigade arv $R_n^\gamma(f)$:

$$A_n(f) \le \frac{1}{n}\sum_{i=1}^{n} I_{\{f(x_i)y_i \le \gamma\}} = R_n^\gamma(f).$$

Meile pakub huvi AdaBoosti väljundi f_T abil saadud klassifitseerija risk (tuletame meelde):

$$R(f_T) = \mathbf{P}\big(Y \neq \operatorname{sgn}(f_T)\big), \qquad \text{kus } f_T = \sum_{t=1}^{T} \alpha_i h_t \text{ on AdaBoosti väljund.}$$

Nagu eelmises peatükis märgitud, klassifitseerija $\operatorname{sgn}(f_T)$ ja seega $R(f_T)$ ei muutu, kui kaalud α_t normaliseerida nii, et nende summa on üks. Seega vaatleme (ühtlasi) riski hinnanguid $R(f)$, kus f kuulub klassi

$$\operatorname{co}_T(\mathcal{H}) := \Big\{ \sum_{t=1}^{T} \beta_t h_t : \beta_t \ge 0, \quad \sum_{t=1}^{T} \beta_t = 1, \quad h_t \in \mathcal{H} \Big\}$$

ja $\mathcal{H}$ on $\{1,-1\}$-väärtuseliste klassifitseerijate (rusikareeglite) hulk. Pane tähele, et tänu skaleerimisele on kõik klassi $\mathcal{F}$ elemendid $[-1,1]$-väärtuselised funktsioonid.

Lihtsaim (klassikaline) meetod saamaks riski hinnanguid oleks klassifitseerijate hulga

$$\mathcal{G}_T := \{\operatorname{sgn}(f) : f \in \operatorname{co}_T(\mathcal{H})\}.$$

VC-dimensiooni kaudu. Kahjuks viimane kasvab koos iteratsioonide arvuga T. Selgub aga, et on võimalik leida ka selliseid riski hinnanguid, mis on *sõltumatud iteratsioonide arvust* T. Näitena vaatame järgmist hinnangut ([21], Cor 1). See hinnang sõltub klassi $\mathcal{H}$ VC dimensioonist, et aga rusikareeglite klass on tüüpiliselt madala dimensiooniga, on see number enamasti väike.

Teoreem 5.4 *Olgu $\gamma > 0$ fikseeritud. Siis tõenäosusega $1-\delta$ (üle valimite), iga $f \in \operatorname{co}_T(\mathcal{H})$ korral*

$$R(f) \le A_n(f) + \frac{8}{\gamma}\sqrt{\frac{2V_{\mathcal{H}}\ln(n+1)}{n}} + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}, \tag{5.2.15}$$

kus $V_{\mathcal{H}}$ on $\mathcal{H}$ VC-dimensioon.

Pane tähele, et funktsiooni $f \in \operatorname{co}_T(\mathcal{H})$ korral $y_i f(x_i) = \rho_i(f)$ nii, et

$$A_n(f) \le R_n^\gamma(f) = \frac{1}{n}|\{i : \rho_i(f) \le \gamma\}| \le \prod_{t=1}^{T} 2\sqrt{\epsilon_t^{1-\gamma}(1-\epsilon_t)^{1+\gamma}} = \prod_{t=1}^{T}(1-2u_t)^{\frac{1-\gamma}{2}}(1+2u_t)^{\frac{1+\gamma}{2}},$$

kus viimane võrratus järeldub võrratusest (5.2.13). Seega, kui $u_t \ge u_o > 0$ ja γ on piisavalt väike, siis me saame riski hinnangu, mis *väheneb iteratsioonide arvu T kasvamisel*. See aga tähendab, et riski hinnang väheneb ka siis, kui valimi treeningviga on juba null. See asjaolu (osaliselt) seletab, miks AdaBoosti juures iteratsioonide kasv ei pruugi ilmtingimata tähendada ülesobituvust

5.2.5 Võrratusest (5.2.15)

Rademacheri kompleksus. Olgu $\mathcal{F}$ mingi funktsioonide $\mathbb{R}^d \to \mathbb{R}$ klass. Selle klassi **Rademacheri complekus (Rademacheri keskmine)** $\mathrm{Ra}(\mathcal{F})$ on a

$$\mathrm{Ra}(\mathcal{F}) := E \sup_{\mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \Big|,$$

kus $X_1, \ldots, X_n$ on iid vektorid ja σ_i, $i = 1, \ldots, n$ on iid juhuslikud suurused, kusjuures $\mathbf{P}(\sigma_i = 1) = \mathbf{P}(\sigma_i = -1) = 0.5$. Juhuslikud suurused σ_i ja juhuslikud vektorid X_i, $i = 1, \ldots, n$ on sõltumatud ning keskväärtus on võetud üle vektorite X_i ja üle märkide σ_i. Mõnikord (näiteks raamatus [14]), Rademacheri kompleksuse definitsioonis on keskväärtus korrutatud arvuga kaks.

Kui klass $\mathcal{G}$ koosneb klassifikaatoritest $\mathbb{R}^d \to \{-1, 1\}$, siis Rademacheri kompleksuse abil on võimalik saada mõnevõrra täpsemaid riski hinnanguid kui VC-dimensiooni kaudu. Nimelt kehtib järgmine teoreem ([4], Thm 5; vaata ka [14], Thm 4.9 või [3], Thm 3.2).

Teoreem 5.5 *Olgu $\mathcal{G}$ mingi klassifitseerijate hulk. Siis iga $\delta > 0$ korral tõnäosusega $1 - \delta$*

$$R(g) \le R_n(g) + \mathrm{Ra}(\mathcal{G}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}. \tag{5.2.16}$$

Kahjuks on Rademacheri kompleksust praktikas raske hinnata, pealekauba sõltub ta ju X_i jaotusest, mis meile on üldiselt tundmata. Rademacheri kompleksust on võimalik ülalt hinnata klassi $\mathcal{G}$ VC-dimansiooniga (vt [3], eq (6))

$$\mathrm{Ra}(\mathcal{G}) \le 2\sqrt{\frac{2V \ln(n+1)}{n}}. \tag{5.2.17}$$

Marginaalhinnagud. Hinnangud marginaali $\gamma > 0$ kaudu põhinevad järgmisel olulisel võrratusel ([21], Thm 3; [22]; vaata ka [3], Thm 4.1).

Teoreem 5.6 *Olgu klassi $\mathcal{F}$ elemendid funktsioonid $\mathbb{R}^d \to [-1, 1]$ ning olgu $\gamma \in (0, 1)$. Olgu $\delta > 0$. Siis tõenäosusega $1 - \delta$, iga funktsiooni $f \in \mathcal{F}$ korral*

$$R(f) \le A_n(f) + \frac{4\mathrm{Ra}(\mathcal{F})}{\gamma} + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}. \tag{5.2.18}$$

Rakendamaks võrratust (5.2.18) AdaBoost'i väljundeile, võtame $\mathcal{F} = \mathrm{co}_T(\mathcal{H})$ (pane tähele, et iga f hulgast $\mathrm{co}_T(\mathcal{H})$ on $[-1, 1]$-väärtuseline funktsioon). Rademacheri kompleksuse kasulikkus tuleneb asjaolust, et (erinevalt VC dimensioonist), klassi $\mathrm{co}_T(\mathcal{H})$ Rademacheri kompleksus ei sõltu arvust T, nimelt (vt [21, 23]):

$$\mathrm{Ra}\big(\mathrm{co}_T(\mathcal{H})\big) = \mathrm{Ra}(\mathcal{H}).$$

Seega võrratuses(5.2.18) võib $\mathrm{Ra}(\mathcal{F})$ asendada klassi $\mathcal{H}$ Rademacheri komplekusesga mida omakorda võib võrratuse (5.2.17) abilt ülalt hinnata selle klassi $\mathcal{H}$ VC-dimensiooniga:

$$\mathrm{Ra}(\mathcal{F}) = \mathrm{Ra}(\mathcal{H}) \leq 2\sqrt{\frac{2V_{\mathcal{H}} \ln(n+1)}{n}}.$$

Pannes saadud hinnangud seosesse (5.2.18) saame (5.2.15).

5.2.6 AdaBoosti mõjusus

Hiljuti tõestasid P. Bartlett ja M. Traskin [24], et sobiva rusikareeglite klassi $\mathcal{H}$ korral on AdaBoost tugevalt mõjus klassifitseerimisreegel; eeldusel, et iteratsioonide arv T sõltub valimi mahust n, läheneb n kasvamisel lõpmatusse ($T_n \to \infty$) kuid mitte kiiremini kui $T = n^{\nu}$, kus $0 < \nu < 1$. Oluline aga on, et AdaBoost igal sammul minimiseerib kaalutud valimi treeningvea. Seega iga t korral

$$h_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} D_t(i) I_{\{h(x_i) \neq y_i\}}.$$

Teame, et sellisel juhul h_t ja α_t minimiseerivad

$$\sum_{i=1}^{n} \exp[-y_i(f_{t-1}(x_i) + \alpha h(x_i))]$$

üle kõikide $\alpha \in \mathbb{R}$ ja $h \in \mathcal{H}$. Seega funktsioon $f_t = f_{t-1} + \alpha_t h_t$ rahuldab iga t korral tingimust

$$\sum_{i=1}^{n} \exp[-y_i f_t(x_i)] = \inf_{\alpha, h} \sum_{i=1}^{n} \exp[-y_i(f_{t-1}(x_i) + \alpha h(x_i))].$$

Sellisel juhul võib AdaBoosti defineerida järgmiselt.

AdaBoost: Antud (valitud) rusikareeglite hulk $\mathcal{H}$.

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T.
2. Olgu $f_1 \equiv 0$;
3. **Iga** $t = 1, \ldots, T$ **korral** defineeri

   $$f_t = f_{t-1} + \alpha_t h_t,$$

   kus

   $$\sum_{i=1}^{n} \exp[-y_i f_t(x_i)] = \inf_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{n} \exp[-y_i(f_{t-1}(x_i) + \alpha h(x_i))].$$

4. **Väljund:**

$$g(x) := \operatorname{sgn}\Big(\sum_{t=1}^{T} \alpha_t h_t(x)\Big).$$

Olgu iga $\lambda > 0$ korral

$$\mathcal{F}_\lambda := \Big\{ \sum_{i=1}^{m} \lambda_i h_i, \quad m = 0, 1, 2, \ldots, \quad \sum_{i=1}^{m} \lambda_i = \lambda, \quad h_i \in \mathcal{H}, \quad \lambda_i \geq 0 \Big\}, \quad \mathcal{G}_\lambda := \{\operatorname{sgn} f : f \in \mathcal{F}_\lambda\}.$$

On selge, et AdaBoosti väljund f_T kuulub mingi piisavalt suure λ korral alati hulka $\mathcal{G}_\lambda$. Mõjusus tõestatakse võrratuse (5.1.10) abil

$$\psi\big(R(f_{T_n}) - R^*\big) \leq R_\phi(f_{T_n}) - R^*_\phi,$$

kus $\phi(t) = \exp[-t]$, sellele vastav $\psi(t) = 1 - \sqrt{1 - t^2}$ ning jada T_n on sobivalt valitud (st iteratsioonide arv kasvab koos valimimahuga, kuid mitte liiga kiiresti). Seega eesmärk on surrogaatriski (ϕ-riski) mõjusus:

$$R_\phi(f_{T_n}) \to R^*_\phi.$$

Et viimane koondumine kehtiks peab lähendamisviga koonduma nulliks ehk järgmine tingimus peab kehtima:

$$\lim_{\lambda \to \infty} \inf_{f \in \mathcal{F}_\lambda} R_\phi(f) = R^*_\phi. \tag{5.2.19}$$

Omadus (5.2.19) sõltub nii jaotusest P kui ka rusikareeglite hulgast $\mathcal{H}$. Samas leidub klassifitseerijaid $\mathcal{H}$, mis rahuldavad (5.2.19) iga jaotuse P korral. Sellised klassid on näiteks lineaarsete klassifitseerijate hulk, teatavad kahendpuud jne.

Teoreem 5.7 (Bartlett, Trashkin, 2007). *Olgu klassi $\mathcal{H}$ VC dimensioon lõplik, $R^* > 0$ ja kehtigu (5.2.19). Olgu g_n AdaBoosti kaudu leitud klassifitseerija iteratsioonide arvuga T_n. Kui $T_n \to \infty$ ja $T_n = O(n^\nu)$, kus $\nu < 1$, siis kehtib koondumine*

$$R(g_n) \to R^*, \quad \text{p.k.},$$

st reegel on tugevalt mõjus.

Rusikareeglite hulga lõplik VC-dimensioon on loomulik eeldus (meenuta riski hinnanguid), on ju rusikareeglite klass võimalikult lihtne. Samuti, nagu nägime, on loomulik eeldus (5.2.19). Kui $\mathcal{H}$ on selline, et (5.2.19) kehtib iga P korral, ei sea see eeldus kitsendusi jaotusele. Küll aga on jaotusi, mille korral $R^* = 0$. Seega, isegi kui (5.2.19) kehtib kõikide jaotuste korral, ei järeldu ülaltoodud teoreemist universaalne mõjusus.

5.3 Boosting kui gradientmeetod

Nägime, et AdaBoost minimiseerib (teatud viisil) empiirilise ϕ-riski

$$R_\phi^n(f) = \sum_i \phi(y_i f(x_i))$$

üle $\mathcal{F}$. Siin $\phi(t) = \exp[-t]$. Veendume, et AdaBoost minimiseerib seda funktsiooni teataval gradientmeetodil (*steepest gradient descent*), kus igal iteratsioonisammul t otsitakse funktsiooni $h_t \in \mathcal{H}$ nii, et funktsiooni

$$\alpha \mapsto R_\phi^n\Big(\sum_{s=1}^{t-1} \alpha_s h_s + \alpha h_t\Big)$$

kahanemine punktis 0 on maksimaalne (*coordinate-descent*). Seega otsime funktsiooni h_t nii, et

$$\begin{aligned}
h_t &= \arg\min_{h\in\mathcal{H}} \frac{\partial R_\phi^n\big(f_{t-1} + \alpha h\big)}{\partial \alpha}\Big|_{\alpha=0} \\
&= \arg\min_{h\in\mathcal{H}} \sum_i \frac{\partial \phi\Big(y_i\big(f_{t-1}(x_i) + \alpha h(x_i)\big)\Big)}{\partial \alpha}\Big|_{\alpha=0} \\
&= \arg\min_{h\in\mathcal{H}} \sum_i \phi'\big(y_i f_{t-1}(x_i)\big) y_i h(x_i) \\
&= \arg\min_{h\in\mathcal{H}} -\sum_i \big(2I_{\{y_i \neq h(x_i)\}} - 1\big)\phi'(y_i f_{t-1}(x_i)).
\end{aligned}$$

Juhul, kui ϕ on kahanev (mida ta enamikel juhtudel ka on), siis $-\phi'\big(y_i f_{t-1}(x_i)\big) \geq 0$ iga i korral ning h_t on seega selline, mis minimiseerib kaalutud empiirilise riski

$$h_t = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^n D_\phi^t(i) I_{\{y_i \neq h(x_i)\}},$$

kus

$$D_\phi^t(i) \propto -\phi'\big(y_i f_{t-1}(x_i)\big). \tag{5.3.1}$$

Kui h_t on leitud, siis leiame kaalu α_t, mis minimiseerib $\alpha \mapsto R_\phi^n\big(f_{t-1} + \alpha h_t\big)$ ehk kumera ϕ korral

$$\frac{\partial R_\phi^n\big(f_{t-1} + \alpha h_t\big)}{\partial \alpha}\Big|_{\alpha_t} = \sum_i \phi'\Big(y_i\big(f_{t-1}(x_i) + \alpha_t h_t(x_i)\big)\Big) y_i h_t(x_i) = 0.$$

Nii saame üldise boosting-tüüpi algoritmi:

GradientBoost: Antud (valitud) lihtne klassifitseerimisreegel (rusikareegel) h.

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T.
2. Olgu $D_\phi^1(i) = \frac{1}{n}, \quad i = 1, \ldots, n$;
3. **Iga** $t = 1, \ldots, T$ **korral:**

   a Treeni rusikareegel kaalutud valimi (S, D_ϕ^t) korral. Nii saad (lihtsa) reegli

   $$h_t : \mathcal{X} \to \{-1, 1\}.$$

   b Defineeri

   $$\alpha_t := \arg\min_{\alpha \in \mathbb{R}} R_\phi^n\big(f_{t-1} + \alpha h_t\big)$$

   c Defineeri uued kaalud

   $$D_\phi^{t+1}(i) \propto -\phi'\big(y_i(f_{t-1}(x_i) + \alpha_t h_t(x_i))\big) = -\phi'\big(y_i(f_t(x_i))\big).$$

4. **Väljund:**

   $$g(x) := \operatorname{sgn}\Big(\sum_{t=1}^{T} \alpha_t h_t(x)\Big).$$

AdaBoost on GradientBoost. Kui $\phi(t) = e^{-t}$, siis ülaltoodud algoritm on AdaBoost. Tõepoolest, seosest (5.3.1) saame

$$D_\phi^t(i) \propto \exp[-\big(y_i f_{t-1}(x_i)\big)]$$

st $D_\phi^t(i) = D_t(i)$ ehk kaalud on samad kui AdaBoostis. Kordaja α_t leidmiseks lahendame võttandi

$$\begin{aligned}\sum_i \phi'\Big(y_i\big(f_{t-1}(x_i) + \alpha_t h_t(x_i)\big)\Big) y_i h_t(x_i) &= -\sum_i \exp[-y_i\big(f_{t-1}(x_i) + \alpha_t h_t(x_i)\big)] y_i h_t(x_i) \\ &= -\sum_i \exp[-y_i f_{t-1}(x_i)] \exp[-y_i \alpha_t h_t(x_i)] y_i h_t(x_i) = 0\end{aligned}$$

Seega sellisel juhul α_t saame võrrandist

$$\begin{aligned}\sum_i D_t(i) \exp[-y_i \alpha_t h_t(x_i)] y_i h_t(x_i) &= \sum_{i: h_t(x_i) \neq y_i} D_t(i) e^{\alpha_t} + \sum_{i: h_t(x_i) = y_i} D_t(i) e^{-\alpha_t} \\ &= \epsilon_t e^{\alpha_t} + (1 - \epsilon_t) e^{-\alpha_t} = 0,\end{aligned}$$

mille lahend, nagu teame, on $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$.

Märkus. Oletame korraks, et eesmärk on funktsiooni $R^n_\phi(f) = \sum_i \phi(y_i f(x_i))$ minimiseerimine üle kõikvõimalike n-dimensionaalsete vektorite $(f(x_1), \ldots, f(x_n))$. Klassikalisel gradientmeetodil käib see iteratiivselt nii: sammul t leiame gradiendi (osatuletistega vektori)

$$\nabla R_\phi(f_{t-1}) = \begin{pmatrix} \frac{\partial R_\phi(f)}{\partial f(x_1)} \\ \cdots \\ \frac{\partial R_\phi(f)}{\partial f(x_n)} \end{pmatrix}_{f=f_{t-1}}$$

ja siis leia järgmine lähend

$$f_t = f_{t-1} - \alpha_t \nabla R_\phi(f_{t-1}).$$

Pane tähele, et vektori $\nabla R_\phi(f_{t-1})$ i-s element on $\phi'(y_i f_{t-1}(x_i))y_i$. Meil on vaja aga igal sammul leida uus funktsioon klassist $\mathcal{H}$. Gradientvektori $-\nabla R_\phi(f_{t-1})$ i-s komponent annab meile sisuliselt otsitava funktsiooni h_t märgi kohal x_i aga klassis $\mathcal{H}$ ei pruugi olla ühtegi funktsiooni, mille korral kõikide valimipunktised märgid kattuksid gradiendi omadega. Seetõttu otsime teatavas mõttes gradiendile lähimat funktsiooni h_t. Kirjanduses kasutatekse selleks keskmist ruutkao mõttes. Veendume, et klassifitseerimise korral annab see juba meile tuttava GradientBoost'i. Olgu

$$h_t = \arg\inf_{h\in\mathcal{H}} \sum_i \big(- \frac{\partial R_\phi(f)}{\partial f(x_i)} - h(x_i)\big)^2.$$

Et

$$\begin{aligned}\sum_i \big(- \frac{\partial R_\phi(f)}{\partial f(x_i)} - h(x_i)\big)^2 &= \sum_i \big(- \phi'(y_i f_{t-1}(x_i))y_i - h(x_i)\big)^2 \\ &= \sum_i \big(\phi'(y_i f_{t-1}(x_i))\big)^2 + 2\sum_i \phi'(y_i f_{t-1}(x_i))y_i h(x_i) + \sum_i h^2(x_i)\end{aligned}$$

ja $h^2(x_i) = 1$, näeme: ülaltoodud summa miniimeerimine on ekvivalentne summa

$$\sum_i \phi'(y_i f_{t-1}(x_i))y_i h(x_i)$$

minimeerimisega ja just seda GradientBoost teeb:

$$h_t = \arg\min_{h\in\mathcal{H}} \sum_i \phi'\big(y_i f_{t-1}(x_i)\big)y_i h(x_i).$$

Seega kasutatakse tihti kirjanduses (näiteks raamatutes [8, 10]) GradientBoosti definitsioonis h_t leidmiseks ruutkao mõttes gradiendile lähimat funktsiooni.

5.3.1 LogitBoost

Juhul, kui $\phi(t) = \ln(1 + e^{-t})$, siis saame samuti populaarse algoritmi – **LogitBoost**. Sellisel juhul

$$R^n_\phi(f) = \sum_i \ln\big(1 + \exp[-y_i f(x_i)]\big), \tag{5.3.2}$$

mis harilikust eksponentsiaalsest kaost erineb selle poolest, et suurte negatiivsete marginaalide korral on kadu väiksem. LogitBoosti korral

$$\phi'(t) = \frac{-e^{-t}}{1+e^{-t}} = \frac{-1}{1+e^{t}},$$

millest kaalud (5.3.1):

$$D_\phi^t(i) \propto \frac{1}{1+e^{y_i f_{t-1}(x_i)}}.$$

Kahjuks pole LogitBoosti korral α_t leidmine analüütiliselt lihtne. Kirjanduses: $\alpha_t \equiv \frac{1}{2}$ või nagu AdaBoostis: $\alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$.

LogitBoost ja logistiline regressioon. Logistilises regressioonis otsitakse hinnangut logistilisele tõepärasuhtele

$$f(x) = \ln\frac{\eta(x)}{1-\eta(x)}$$

mingist ettantud klassist $\mathcal{F}$. Peatükis 3.6.2 oli selleks klassiks lineaarsete funktsioonide hulk. Kui $f \in \mathcal{F}$, siis vastav η hinnang on

$$\eta_f(x) = \frac{e^{f(x)}}{1+e^{f(x)}} = \frac{1}{1+e^{-f(x)}}, \quad 1-\eta_f(x) = \frac{1}{1+e^{f(x)}}. \tag{5.3.3}$$

Logistiline regressioon suurima tõepära meetodil maksimiseerib tinglikku tõenäosust

$$\begin{aligned}\sum_{i:y_i=1}\ln\eta_f(x_i) + \sum_{i:y_i=-1}\ln(1-\eta_f(x_i)) &= -\sum_{i:y_i=1}\ln(1+e^{-f(x_i)}) - \sum_{i:y_i=-1}\ln(1+e^{f(x_i)}) \\ &= -\sum_i \ln\bigl(1+\exp[-y_i f(x_i)]\bigr).\end{aligned}$$

Teisisõnu, minimiseeritakse funktsiooni (5.3.2). Seega LogitBoost pole midagi muud, kui logistilise regressiooni tingliku tõepärafunktsiooni teataval viisil maksimiseerimine üle klassi $\mathcal{F} = \operatorname{span}\mathcal{H}$. Logistilise regressiooni ja LogitBoosti vahe seisneb meetodis, millega funktsiooni (5.3.2) minimiseeritakse. Logistilises regressioonis minimiseeritakse seda funktsiooni Newton-Rapshoni meetodil ning funktsioonide klass peab selleks olema lihtne (lineaarsed funktsioonid); GradientBoost minimiseerib seda funktsiooni teataval lihtsal gradientmeetodil, kuid seda on võimalik teha üle laiema klassi $\mathcal{F} = \operatorname{span}\mathcal{H}$.

Ülesanne 5.8 *Olgu* $\phi(t) = \ln(1+\exp[-t])$ *ja* $\phi(t) = \ln(1+\exp[-2t])$. *Veendu, et funktsioonid* α^* *on vastavalt*

$$\alpha^*(\eta) = \ln\frac{\eta}{1-\eta}, \quad \alpha^*(\eta) = \frac{1}{2}\ln\frac{\eta}{1-\eta}.$$

Teinekord kasutatakse ka kaofunktsiooni $\phi(t) = \ln(1 + \exp[-2t])$. Sellisel juhul minimiseerib LogitBoost funktsiooni

$$\sum_i \ln\bigl(1 + \exp[-2y_i f(x_i)]\bigr) \tag{5.3.4}$$

ja ülaltoodud ülesande põhjal tähendab see suuruse

$$\frac{1}{2} \ln \frac{\eta(x)}{1 - \eta(x)}$$

hindamist suurima tõepära meetodil klassist $\mathcal{F}$.

Tingliku tõenäosuse η hindamine. Logistilise regressiooni üldisem eesmärk on tingliku tõenäosuse $\eta(x)$ hindamine. LogitBoosti (5.3.2) (st $\phi(t) = \ln(1 + \exp[-t])$) kaudu leitud funktsiooni f_T abil saame tingliku tõenäosuse hinnanguks

$$\hat{\eta}(x) = \frac{e^{f_T(x)}}{1 + e^{f_T(x)}} = \frac{1}{1 + e^{-f_T(x)}}, \quad 1 - \hat{\eta} = \frac{1}{1 + e^{f_T(x)}}.$$

Kui f_T on leitud LogitBoosti (5.3.4) kaudu (st $\phi(t) = \ln(1 + \exp[-2t])$), siis η hinnang on

$$\hat{\eta}(x) = \frac{1}{1 + e^{-2f_T(x)}}. \tag{5.3.5}$$

Ülesanne 5.9 *Näita, et $\hat{\eta}(x) \geq \frac{1}{2}$ parajasti siis, kui $sgn f_T(x) = 1$.*

Kirjanduses soovitatakse ka valemis (5.3.5) kasutada ka AdaBoosti abil leitud f. Selle põhjendus on funktsioonide $\exp[-t]$ ja $\ln(1 + \exp[-2t])$ sarnane käitumine nulli ümbuses ning asjaolu, et neil on ühine $\alpha^*(\eta) = \frac{1}{2} \ln \frac{\eta}{1-\eta}$.

5.4 Regressioon

Vaatleme regressiooniülesannet, kus $y_i \in \mathbb{R}$ iga i korral ning eesmärk on minimiseerida kaofunktsiooni

$$\sum_{i=1}^{n} L(y_i, f(x_i))$$

üle klassi $\mathcal{F}$. Nagu ikka, enamasti $L(y, x) = (y - x)^2$, boosting-algoritmide korral aga

$$\mathcal{F} = \{\sum_{t=1}^{T} \alpha_t h_t : h_t \in \mathcal{H}\},$$

kus hulk $\mathcal{H}$ koosneb (mingis mõttes) lihtsatest regressioonifunktsioonidest. Selliseid mudeleid nimetatakse aditiivseteks. Tihti kasutatakse nendeks mitmesuguseid puid. Kui klassi $\mathcal{H}$ elemendid on invariantsed skalaariga korrutamise suhtes, st $\alpha h \in \mathcal{H}$ iga h korral (näiteks

puud), siis iga hulga f element on lihtsalt summa $h_1 + \cdots + h_T$. Boosting-tüüpi algoritmides käib aditiivse funktsiooni sobitamine liidetavate kaupa: pärast $f_t := \alpha_1 f_1 + \cdots + \alpha_t f_t$ sobitumist minimiseeritakse summat

$$\sum_{i=1}^{n} L(y_i, f_t(x_i) + \alpha_{t+1} h_{t+1})$$

üle skalaaride α_{t+1} ja funktsioonide h_{t+1} jne. Nii saame üldise skeemi regressiooniks.

Forward stagewise additive modelling:

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T.
2. Olgu $f_0(x) = 0$;
3. **Iga** $t = 1, \ldots, T$ **korral leia**
   - $$(\alpha_t, h_t) := \arg \min_{\alpha_t \in \mathbb{R}, h_t \in \mathcal{H}} \sum_{i=1}^{n} L\big(y_i, f_{t-1}(x_i) + \alpha_t h_t(x_i)\big). \tag{5.4.1}$$
   - $f_t = f_{t-1} + \alpha_t h_t$.
4. **Väljund:** $\hat{f} = f_T$

L_2boosting: $L(y, x) = (y - x)^2$ (ruutkaofunktsioon, regressiooni korral enamlevinud). Siis

$$\sum_i L(y_i, f_{t-1}(x_i) + \alpha_t h_t) = \sum_i \big(y_i - f_{t-1}(x_i) - \alpha_t h_t(x_i)\big)^2,$$

millest saame, et iteratsiooni igal sammul leitakse vähimruutude mõttes parima funktsioon kujul αh eelmisel sammul saadud prognoosi jääkidele: $r_{t-1}(x_i) = y_i - f_{t-1}(x_i)$. Paneme veel tähele, et

$$\arg\min_{\alpha} \sum_i \big(r_{t-1}(x_i) - \alpha h(x_i)\big)^2 = \frac{\sum_{i=1}^{n} r_{t-1}(x_i) h(x_i)}{\sum_i h^2(x_i)}$$

ning (5.4.1) minimiseerimine on sisuliselt vaid üle klassi $\mathcal{H}$.

MART. Tükeldusfunktsioonide korral on funktsioonid h kujul

$$h(x) = \sum_{j=1}^{k} \gamma_j I_{S_j}(x),$$

kus $\{S_1, \ldots, S_k\}$ moodustab ruumi $\mathbb{R}^d$ mingi struktuuriga tükeldus, näiteks puud ja $\gamma_i \in \mathbb{R}$. Sellisel juhul minimiseerimine (minimiseerimist üle α-de pole enam vaja) on

$$\min_h \sum_i L(y_i, f_{t-1}(x_i) + h(x_i)) = \min_{S_1, \ldots, S_k} \sum_{j=1}^{k} \sum_{x_i \in S_j} L(y_i, f_{t-1}(x_i) + \gamma_j),$$

kus

$$\gamma_j = \arg\min_{\gamma\in\mathbb{R}} \sum_{i\in S_j} L(y_i, f_{t-1}(x_i) + \gamma).$$

Juhul, kui $L(y,x) = (y-x)^2$, siis γ_j on jääkide keskmine üle tüki S_j:

$$\gamma_j = \frac{1}{n_j} \sum_{x_i\in S_j} r_{t-1}(x_i),$$

kus n_j on tükki S_j kuuluvate x_i-de arv.
Seega minimiseerimine taandub optimaalsete tükkide S_j leidmisele. Juhul, kui $\mathcal{H}$ koosneb puudest, nimetatakse saadud algoritmi MART (*multiple additive regression trees*) algoritmiks, mis ruutkaofunktsiooni $L(y,x) = (y-x)^2$ korral on seega järgmine:

MART (Friedman 2001):

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T, puu lehtede arv $k_1, \ldots, k_T$, $\nu \in (0, 1]$.
2. Olgu $f_0(x) = \arg\min_\gamma \sum_{i=1}^n (y_i - \gamma)^2$;
3. **Iga** $t = 1, \ldots, T$ **korral leia**
   - $r_{t-1}(x_i) = y_i - f_{t-1}(x_i)$;
   - puu $S_1^t, \ldots, S_{k_t}^t$ ning $\gamma_1^t, \ldots, \gamma_{k_t}^t$ nii, et

     $$\sum_{j=1}^{k_t} \sum_{x_i\in S_j^t} (r_{t-1}(x_i) - \gamma_j^t)^2;$$

     oleks minimaalne (seega γ_j^t on jääkide keskmine üle tüki S_j^t);
   - defineeri

     $$f_t = f_{t-1} + \nu \sum_{j=1}^{k_t} \gamma_j^t I_{S_j^t}.$$
4. **Väljund:** $\hat{f} = f_T$.

Arv $\nu \in (0, 1]$ (*shrinkage parameter*) näitab, millise kaaluga uus liidetav mudelisse võetakse. Simulatsioonid on näidanud, et mida väiksem ν, seda väiksem testviga, samas väiksem ν nõuab suuremat iteratsioonide arvu T.

Splainid. Ühedimensionaalsete splainide korral on funktsioonid h ühedimensionaalsed kuupsplainid, st kolmandat järku ja kaks korda pidevalt diferentseeruvad splainid. Igal sammul lähendatakse nendega vaid üht tunnust (d-st tunnusest). Formaalselt: jääkide $r(x_1), \ldots, r(x_n)$ ning tunnuse $j \in \{1, \ldots, d\}$ korral otsime parimat lähendatavat kuupsplaini h^j järgmises mõttes:

$$h^j := \arg\inf_h \sum_{i=1}^{n} \big(r(x_i) - h(x_i^j)\big)^2 + \lambda \int \big(h''(x)\big)^2 dx, \tag{5.4.2}$$

kus $\lambda > 0$ on regulariseeriv konstant ja x_i^j on i-nda tunnusvektori j-s komponent. Nii saame iteratsiooni igal sammul d erinevat splaini $h^1, \ldots, h^d$ ja igal neist valitakse selline, mis ruutkao mõttes lähendab jääke paremini:

$$j^* := \arg\min_{j=1,\ldots,d} \sum_{i=1}^{n} \big(r(x_i) - h^j(x_i^j)\big)^2. \tag{5.4.3}$$

Nii saame järgmise L_2-boosting algoritmi.

L_2-Boosting with componentwise smoothing splines (Bühlmann, Yu, 2003):

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T, $\lambda > 0$, $\nu \in (0, 1]$.
2. Olgu $f_0(x) = 0$;
3. **Iga** $t = 1, \ldots, T$ **korral leia**
   - $r_{t-1}(x_i) = y_i - f_{t-1}(x_i)$;
   - Leia parim parim ühedimensionaalne splain jääkidele $r_{t-1}(x_i)$ seoste (5.4.2) ja (5.4.3) mõttes. Olgu see $h_t^{j^*}$.
   - defineeri
   $$f_t = f_{t-1} + \nu h_t^{j^*},$$
4. **Väljund:** $\hat{f} = f_T$.

Boosting ja Lasso. Olgu $\mathcal{H} = \{h_1, \ldots, h_K\}$ ning vaatleme lineaarset mudelit kujul

$$f = \sum_{k=1}^{K} \alpha_k h_k.$$

Regressiooniülesanne on nüüd

$$\min_{\alpha} \Big[\sum_{i=1}^{n} \big(y_i - \sum_{k=1}^{K} \alpha_k h_k(x_i)\big)^2 + \lambda J(\alpha) \Big],$$

kus $\alpha := (\alpha_1, \ldots, \alpha_K)$ ning $J(\alpha)$ on mingi karistusfunktsioon, näiteks $J(\alpha) = \sum_k \alpha_k^2$ kantregressiooni ja $J(\alpha) = \sum_k |\alpha_k|$ Lasso korral. Tihti on K väga suur ning reguraliseerimine on seega vajalik.

Teame, et Lasso korral on optimaalse vektori α leidmine mittetriviaalne. Simulatisoonid on näidanud, et järgmise algoritmi väljundid lähendavad väga hästi Lassot.

Foreward stagewise linear regression

1. **Sisend:** Valim $S = (x_1, y_1), \ldots, (x_n, y_n)$; iteratsioonide arv T; $\epsilon > 0$.
2. Olgu $\alpha_k^0 = 0$ iga $k = 1, \ldots, K$ korral , $f_0 = 0$;
3. **Iga** $t = 1, \ldots, T$ **korral leia**
   - $r_{t-1}(x_i) = y_i - f_{t-1}(x_i)$;
   - $(\beta, l) = \arg\min_{\beta,l} \sum_{i=1}^n \left(r_{t-1}(x_i) - \beta h_l(x_i)\right)^2$;
   - $\alpha_l^t = \alpha_l^{t-1} + \epsilon \cdot \mathrm{sgn}(\beta), \quad \alpha_k^t = \alpha_k^{t-1}$, kui $t \neq l$;
   - $f_t = \sum_{k=1}^K \alpha_k^t h_k$
4. **Väljund:** $\hat{f} = f_T$.

Siin iteratsioonide arv T asendab reguraliseerimisparameetrit λ – mida rohkem iteratsioone, seda suurem $\sum_k |\alpha_k|$.

Kirjandus: boostingust loe [7], Ch 10, [10], Ch 16 või ülevaateartikleid [21, 26, 25, 27].

Peatükk 6

Ülevaade teistest meetodidest

Eeldus: Olgu klasside märgistused 0 ja 1.

6.1 Plug-in reeglid

Tuletame meelde: $\eta(x) = \mathbf{P}(Y = 1|X = x)$, Bayesi reegel (olgu käesolevas peatükis viigi korral eelistus klassile 0)

$$g^*(x) = \begin{cases} 1, & \text{kui } \eta(x) > 0.5; \\ 0, & \text{kui } \eta(x) \leq 0.5. \end{cases}$$

Lemma 5.2: Olgu g klassifikaator (see kas g väärtused on 0 ja 1 või -1 ja 1 ei muuda midagi). Siis

$$\begin{aligned} R(g) - R^* &= \mathbf{P}(g(X) \neq Y) - R^* \\ &= 2E\Big(|\eta(X) - \frac{1}{2}|I_{\{g(X)\neq g^*(X)\}}\Big) = 2\int_{\{g(x)\neq g^*(x)\}} |\eta(x) - \frac{1}{2}|F(dx), \end{aligned}$$

kus g^* on Bayesi klassifikaator, F on tunnuse X jaotus.

Plug-in klassifitseerija defineeritakse läbi funktsiooni η hinnangu η_n järgiselt

$$g_n(x) = \begin{cases} 1, & \text{kui } \eta_n(x) > 0.5; \\ 0, & \text{kui } \eta_n(x) \leq 0.5. \end{cases} \tag{6.1.1}$$

Paneme tähele: kui g_n on plug-in klassifitseerija, siis

$$g_n(x) \neq g^*(x) \Rightarrow |\eta(x) - \frac{1}{2}| \leq |\eta(x) - \eta_n(x)|. \tag{6.1.2}$$

Asendades seose (7.2.6) lemmasse 5.2, saame järgmise olulise järelduse.

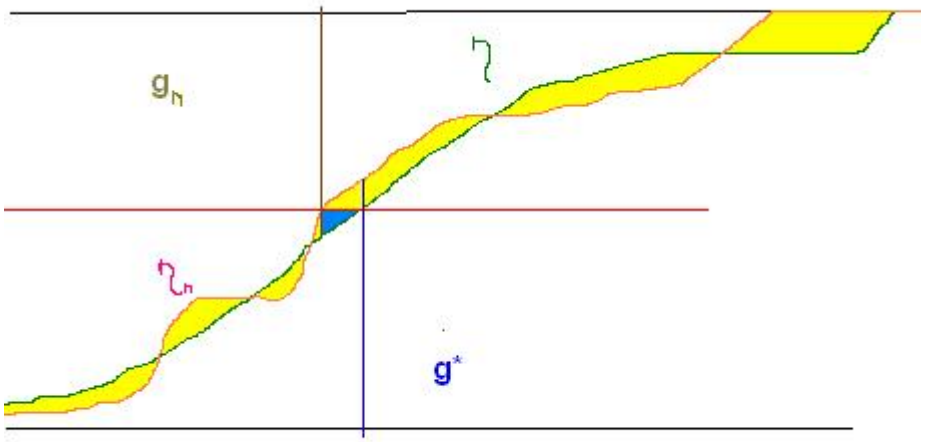
Järeldus 6.1.1 *Olgu η_n funktsiooni η hinnang ning g_n selle põhjal saadud plug-in klassifitseerija. Siis*

$$R(g_n) - R^* = 2\int_{\{g_n(x)\neq g^*(x)\}} \left|\eta(x) - \frac{1}{2}\right| F(dx) \leq 2\int |\eta(x) - \eta_n(x)| F(dx). \tag{6.1.3}$$

Ljapunovi (või ka Cauchy-bunjakowski-Schwartzi) võrratusest saame

$$R(g_n) - R^* \leq 2\int |\eta(x) - \eta_n(x)| F(dx) \leq 2\sqrt{\int \big(\eta(x) - \eta_n(x)\big)^2 F(dx)}. \tag{6.1.4}$$

Seega, kui hinnang η_n on mõjus L_2-mõttes (või L_1-mõttes), st (6.1.4) parem pool läheneb nullile, siis on seda mõjus ka plug-in klassifitseerija g_n. On aga selge, et η_n mõjusus on on vaid piisav kuid mitte tarvilik g_n mõjususeks. Tõepoolest, nagu järgmisest joonisest näha, piisab g_n mõjususeks palju vähemast.



Teinekord hinnatakse tõenäosusi $\eta(x)$ ja $1-\eta(x)$ eraldi. Olgu $\hat{\eta}_1$ tõenäosuse $\eta(x)$ hinnang ning $\hat{\eta}_0$ olgu tõenäosuse $1-\eta(x)$ hinnang, kusjuures nende summa ei pruugi ilmtingimata olla 1. Plug-in klassifitseerija sellisel juhul oleks

$$g_n(x) = \begin{cases} 1, & \text{kui } \hat{\eta}_1(x) > \hat{\eta}_0(x); \\ 0, & \text{kui } \hat{\eta}_1(x) \leq \hat{\eta}_0(x). \end{cases} \tag{6.1.5}$$

Ülesanne 6.1 *Tõesta, et*

$$R(g_n) - R^* \leq \int |\eta(x) - \hat{\eta}_1(x)| F(dx) + \int |(1-\eta(x)) - \hat{\eta}_{-1}(x)| F(dx). \tag{6.1.6}$$

Ülesanne 6.2 *Olgu f_1 ja f_0 klasside-sisesed tihedused, $p = \mathbf{P}(Y=1)$. Olgu $\hat{p}_1$ tõenäosuse p ning $\hat{p}_0$ tõenäosuse $1-p$ hinnang; olgu $\hat{f}_1$ ja $\hat{f}_0$ tiheduste f_1 ja f_0 hinnangud.*

Defineerime klassifitseerija

$$g_n(x) = \begin{cases} 1, & \textit{kui } \hat{p}_1\hat{f}_1(x) \geq \hat{p}_0\hat{f}_0(x); \\ 0, & \textit{kui } \hat{p}_1\hat{f}_1(x) < \hat{p}_0\hat{f}_0(x). \end{cases} \tag{6.1.7}$$

Veenduda, et (6.1.1) on plug-in klassifitseerija ning tõestada, et

$$R(g_n) - R^* \leq \int |pf_1(x) - \hat{p}_1\hat{f}_1(x)|dx + \int |(1-p)f_0(x) - \hat{p}_0\hat{f}_0(x)|dx. \tag{6.1.8}$$

6.1.1 Standardne plug-in: parametriseerimine

Olgu tihedused f_1 ja f_0 teada parameetrilisel kujul $f_1(\theta_1, x)$ ja $f_0(\theta_0, x)$. **Standardne plug-in reegel** seisneb jaotuse hindamises parameetrite kaudu, st tundmatute parameetrite θ_1, θ_0 ja $\pi = \mathbf{P}(Y = 1)$ asendatakse hinnangutega $\hat{\theta_1}$, $\hat{\theta_0}$ ja $\hat{\pi}$. Saadud klassifitseerija:

$$g_n(x) = \begin{cases} 1, & \text{if } \hat{\pi}f_1(\hat{\theta}_1, x) \geq (1-\hat{\pi})f_0(\hat{\theta}_0, x); \\ 0, & \text{if } \hat{\pi}f_1(\hat{\theta}_1, x) < (1-\hat{\pi})f_0(\hat{\theta}_0, x). \end{cases}$$

Kui mudel on korrektne, siis leiduvad "õiged parameetrid"θ_i^* nii, et funktsioonid $f_i(\theta_i^*, x)$, $i = 0, 1$ on tunnusvektori X tinglikud tihedused. Parameetrite hinnagud $\hat{\theta}_i$ on *(tugevalt) mõjusad* kui $\hat{\theta}_i \to \theta_i^*$ iga $i = 0, 1$ korral tõenäosuse järgi (p.k..). Kui $\hat{\pi}$ on ühtede proportsioon valimis, siis SLLN põhjal $\hat{\pi} \to \pi$, p.k.. Kas parameetrite (tugev) mõjusus garanteerib plug-in reegli (tugeva) mõjususe? Üldiselt: ei.

Näide. Olgu $f(\theta, x)$ ühtlase jaotuse $U[-\theta, 0]$ tihedus kui $\theta \neq 1$ ning ühtlase jaotuse $U[0, 1]$ tihedus, kui $\theta = 1$, $\pi = 0.5$. Olgu $\theta_1^* = 2$ ja $\theta_0^* = 1$. Vaatleme prameetri hinnangut

$$\hat{\theta}_i := \max_{j:Y_j=i} |X_i|.$$

On selge, et $\hat{\theta}_i \to \theta_i^*$ a.s.. Teisalt iga n korral $\hat{\theta}_i \neq 1$ p.k.. Seega, kui $x > 0$, siis $f(\hat{\theta}_1, x) = f(\hat{\theta}_2, x) = 0$ ning vastavalt standardsele plug-in reeglile, $g_n(x) = 1$. Seega $R(g_n) \geq \mathbf{P}(Y = 0) = 0.5$, kuid $R^* = 0$.

Pidevus. Ülaltoodud kontranäites polnud funktsioonid $\theta \to f(\theta, x)$ pidevad ning seetõttu oli ka selline näide võimalik. (Sobiva) pidevuse korral on aga ka standardne plug-in mõjus reegel Defineerime

$$\eta_\theta(x) := \frac{\pi f_1(\theta_1, x)}{\pi f_1(\theta_1, x) + (1-\pi)f_0(\theta_0, x)}.$$

Funktsioon $\theta \to \eta_\theta$ on *pidev ruumis* $L_1(\mathbb{R}^d, F)$, kui koondumisest $\theta_n \to \theta$ järeldub

$$\int |\eta_{\theta_n}(x) - \eta_\theta(x)|F(dx) \to 0.$$

Teoreem 6.1 *Olgu $\theta \mapsto \eta_\theta$ pidev ruumis $L_1(\mathbb{R}^d, F)$. Kui hinnangud $\hat{\theta}_i$ on (tugevalt) mõjusad, siis standardne plug-in reegel $\{g_n\}$ on (tugevalt) mõjus.*

Ülesanne 6.3 *Tõesta teoreem.*

Ülesanne 6.4 *Näita, et ülaltoodud kontranäites funktsioonid η_θ pole pidevad ruumis $L_1(\mathbb{R}^d, F)$.*

Piisav tingimus funktsiooni η_θ pidevuseks ruumis $L_1(\mathbb{R}^d, F)$: iga x ja $i = 0, 1$ korral on kujutis $\theta_i \mapsto f_i(\theta_i, x)$ pidev. Tõepoolest, siis iga x korral on ka $\theta \mapsto \eta_\theta(x)$ pidev ja domineeritud koondumise teoreemist järeldub pidevus ruumis $L_1(\mathbb{R}^d, F)$.

6.2 Tükeldusreeglid

Olgu $\mathcal{S} := \{S_1, S_2, \ldots\}$ ruumi $\mathbb{R}^d$ **tükeldus** (*partition*): $\mathbb{R}^d = \cup_{i \geq 1} S_i$, $S_i \cap S_j = \emptyset$, kui $i \neq j$. Iga $x \in \mathbb{R}^d$ korral olgu $S(x)$ vektorit x sisaldav tükk. Vaatleme klassifikaatorit g_n, mis igale vektorile x seab vastavusse klassi 1 parajasti siis, kui tükki $S(x)$ kuulub rohkem klassi 1 kuuluvaid treeningvalimi elemente. Seega

$$g_n(x) = \begin{cases} 0, & \text{kui } \sum_{i=1}^n I_{\{y_i=1\}} I_{\{x_i \in S(x)\}} \leq \sum_{i=1}^n I_{\{y_i=0\}} I_{\{x_i \in S(x)\}}; \\ 1, & \text{mujal.} \end{cases} \tag{6.2.1}$$

Seega klassifitseerija g_n minimiseerib empiirilise riski üle kõikide nende klassifitseerijate, mis on defineeritud tükeldusega $\mathcal{S}$.

Definitsioon 6.2 **Tükeldusreegel** *(ik* partition rule*) koosneb klassifitseerijatest $\{g_n\}$, kusjuures g_n on defineeritud kui (6.2.1) ja talle vastav tükeldus võib sõltuda nii n-st kui ka valimist $x_1, \ldots, x_n$.*

Iga klassifitseerija defineerib tükelduse (näiteks lineaarsete klassifitseerijate korral on neid tükke kaks), tükeldusreegleid aga eristab 2 tunnust:

1. tükeldus ei sõltu märkidest $y_1, \ldots, y_n$
2. iga tüki märk (g_n väärtus) määratakse häälteenamusega.

Teist tingimust rahuldavaid reegleid nimetatakse tihti ka **loomulikeks** (*natural*). Kitsendustele vaatamata on tükeldusreeglid lai klass (naabrireeglid, histogrammid, puud jne).

6.2.1 Tükeldusreeglite universaalne mõjusus

Millal on tükeldusreegel mõjus? Olgu $\{S_1^n, S_2^n, \ldots\}$ valimist sõltuv tükeldus. Et tükeldusreegel n kasvades lähendaks Bayesi klassifitseerijat punktis x, peavad (peaaegu iga) x korral tükid $S^n(x)$ n-i kasvades kindlasti "vähenema". Teisest küljest aga määratekse tüki $S^n(x)$ klass lõplike hulga (tükki kuuluvate) treeninvalimi elementide põhjal. Et seejuures

tehtav viga väheneks, peab kindlasti tükkidesse $S^n(x)$ kuuluvate kuuluvate valimi elementide arv $N(x)$ kasvama. Järgnev teoreem näitab, et need kaks tingimust on sisuliselt piisavad. Olgu diam$S$ hulga S diameeter, st

$$\text{diam}S = \sup_{a,b\in S} \|a-b\|.$$

Teoreem 6.3 *Olgu $\{S_1^n, S_2^n, \ldots\}$ valimist sõltuv tükelduste jada. Selle põhjal defineeritud tükeldusreegel on universaalselt mõjus, st $ER(g_n) \to R^*$, kui kehtivad järgmised tingimused*

1. *diam$S_n(X) \to 0$ tõenäosuse järgi*
2. *$N(X) \to \infty$ tõenäosuse järgi,*

kus X on jaotusega $F(x)$ valimist sõltumatu juhuslik suurus.

Järgmine lemma aitab kontrollida kas $N(X) \to \infty$ tõenäosuse järgi eeldusel, et tükeldus ei sõltu valimist.

Lemma 6.1 *Olgu $p = (p_1, \ldots, p_k)$ tõenäosusvektor ja $N = (N_1, \ldots, N_k)$ multinomiaalse jaotusega vektor, parameetritega n ja p. Olgu X väärtusi $1, \ldots, k$ võttev juhuslik suurus, mis on sõltumatu vektorist N ja mille jaotus on p. Siis iga M korral*

$$\mathbf{P}(N_X \le M) \le \frac{(2M+4)k}{n}.$$

Tõestus. Olgu

$$J := \{i : p_i > 2\frac{M}{n}\}, \quad J^c := \{i : p_i \le 2\frac{M}{n}\}.$$

Seega J koosneb suurtest aatomitest. Olgu $Z_i \sim B(p_i, n)$

$$\begin{aligned}
\mathbf{P}(N_X \le M) &= \sum_{i\in J} \mathbf{P}(N_i \le M, X = i) + \sum_{i\in J^c} \mathbf{P}(N_i \le M, X = i) \\
&\le \sum_{i\in J^c} p_i + \sum_{i\in J} \mathbf{P}(N_i \le M | X = i) p_i \\
&\le \frac{2Mk}{n} + \sum_{i\in J} \mathbf{P}(N_i \le M) p_i \\
&\le \frac{(2M+4)k}{n}.
\end{aligned}$$

■

Ülesanne 6.5 *Tõesta viimane võrratus, kasutades Tšebõševi võrratust.*

Toodud lemmast järeldub, et kui tükeldus koosneb $k(n)$ tükist ja ei sõltu valimist, siis tingimus 2) on täidetud, kui $\frac{k}{n} \to 0$.

Regressioon ja üldine kaofunktsioon tükeldusreeglite abil. Tükeldusreegleid on kerge üldistada rohkem kui kahele klassile ja ka üldisele kaofunktsioonile. Tõepoolest, häälteenamus tükil pole muud kui ERM-printsiip sellel tükil sümmeetrilise kao korral. Kui kaofunktsioon on midagi muud kui sümmeetriline, tuleb tükeldusreegel $g_n(x)$ defineerida nii, et $g_n(x)$ minimiseerib empiirilise riski tükil $S(x)$ (seega sellel tükil konstantne). Formaalselt siis (tuleta meelde, et $\mathcal{Y}$ on klasside hulk)

$$g_n(x) = \arg\min_{j\in\mathcal{Y}} \sum_{i=1}^{n} L(y_i, j) I_{\{x_i \in S(x)\}}.$$

Nüüd on selge, et tükeldusreegleid saab väga hästi kasutada ka regressiooniks – regressioonifunktsiooni väärtus tükil on minimiseerib empiirilist kadu:

$$g_n(x) = \arg\min_{r\in\mathbb{R}} \sum_{i=1}^{n} L(y_i, r) I_{\{x_i \in S(x)\}}.$$

Kui $L(y, r) = (r - y)^2$, siis $g_n(x)$ on Y-tunnuse keskmine üle tüki (veendu selles!):

$$g_n(x) = \frac{1}{n(x)} \sum_{i=1}^{n} y_i I_{\{x_i \in S(x)\}}, \quad \text{kus} \quad n(x) := \sum_{i} I_{\{x_i \in S(x)\}}.$$

Teoreemi 6.3 tõestus*

Olgu

$$\eta_n(x) := \frac{1}{N(x)} \sum_{i: x_i \in S(x)} y_i,$$

kusjuures $\frac{0}{0} = 0$. Seega $\eta_n(x)$ võib vaadelda kui $\eta(x)$ hinnangut. Järeldusest 6.1.1 teame, et

$$R(g_n) - R^* \le 2E[|\eta(X) - \eta_n(X)||D_n],$$

millest

$$ER(g_n) - R^* \le 2E\big(E[|\eta(X) - \eta_n(X)||D_n]\big) = 2E|\eta(X) - \eta_n(X)|, \tag{6.2.2}$$

kus keskväärtus on võetud üle iid valimi $(X_1, Y_1), \ldots, (X_n, Y_n)$ ja sellest sõltumatu juhusliku suuruse X. Teoreem on tõestatud, kui näitame, et (6.2.2) parem pool koondub nulliks.
Olgu $Z \sim F$ ja defineerime

$$\bar{\eta}(x) := E[\eta(Z)|Z \in S(x)].$$

Et tükeldus sõltub üldiselt valimist, sõltub sellest ka $\bar{\eta}$. Fikseeritud tükelduse korral on $x \mapsto \bar{\eta}(x)$ konstantne tükelduse igal tükil. Kui $S(x)$ on fikseeritud, siis

$$\bar{\eta}(x) = \mathbf{P}(Y = 1|X \in S(x)). \tag{6.2.3}$$

Ülesanne 6.6 *Tõesta (6.2.3).*

Seega, kui $S(x)$ on fikseeritud ja $N(x)$ teada, siis

$$Z(x) := N(x)\eta_n(x)$$

on binoomjaotusega juhuslik suurus, kusjuures jaotuse parameetrid on $N(x)$ ja $\bar{\eta}(x)$. Seega,

$$E|Z(x) - \bar{\eta}(x)N(x)| \le \sqrt{E\big(Z(x) - \bar{\eta}(x)N(x)\big)^2} = \sqrt{N(x)\bar{\eta}(x)(1-\bar{\eta}(x))}.$$

Kui $N(x) > 0$, siis

$$E|\eta_n(x) - \bar{\eta}(x)| = E\Big|\frac{Z(x)}{N(x)} - \bar{\eta}(x)\Big| \le \sqrt{\frac{\bar{\eta}(x)(1-\bar{\eta}(x))}{N(x)}} \le \frac{1}{2\sqrt{N(x)}}. \tag{6.2.4}$$

Paneme tähele, et võrratuse (6.2.4) parem pool sõltub valimist $X_1, \dots, X_n$ ainult läbi $N(x)$. Kui $N(x) = 0$, siis

$$E|\eta_n(x) - \bar{\eta}(x)| \le 1.$$

Seega

$$\begin{aligned}
E_{D_n}|\eta_n(x) - \bar{\eta}(x)| &= E_{D_n}|\eta_n(x) - \bar{\eta}(x)|I_{\{N(x)>0\}} + E_{D_n}|\eta_n(x) - \bar{\eta}(x)|I_{\{N(x)=0\}} \\
&\le \sum_{i=1}^{n} \frac{1}{2\sqrt{i}}\mathbf{P}(N(x) = i) + \mathbf{P}(N(x) = 0) \\
&\le \frac{1}{2}\sum_{i=1}^{k} \mathbf{P}(N(x) = i) + \sum_{i=k+1}^{k} \frac{1}{2\sqrt{i}}\mathbf{P}(N(x) = i) + \mathbf{P}(N(x) = 0) \\
&\le \mathbf{P}(N(x) \le k) + \frac{1}{2\sqrt{k}}.
\end{aligned}$$

Toodud võrratused kehtivad iga x korral, tõenäosus on üle kõikide valimite. Keskmistades üle X, saame

$$E|\eta_n(X) - \bar{\eta}(X)| = E\big(E[|\eta_n(X) - \bar{\eta}(X)|X]\big) \le E(\mathbf{P}[N(X) \le k|X]) + \frac{1}{2\sqrt{k}} = \mathbf{P}(N(X) \le k) + \frac{1}{2\sqrt{k}}.$$

Kolmnurga võrratusest saame

$$E|\eta_n(X) - \eta(X)| \le E|\eta_n(X) - \bar{\eta}(X)| + E|\bar{\eta}(X) - \eta(X)| \le \mathbf{P}(N(X) \le k) + \frac{1}{2\sqrt{k}} + E|\bar{\eta}(X) - \eta(X)|.$$

Hindame $E|\bar{\eta}(X) - \eta(X)|$. Iga $\epsilon > 0$ korral leidub kompaktne hulk C ja sellel antud pidev funktsioon η_ϵ nii, et

$$\int |\eta(x) - \eta_\epsilon(x)|F(dx) = E|\eta(X) - \eta_\epsilon(X)| \le \epsilon.$$

Kolmnurga võrratus:

$$E|\bar{\eta}(X)-\eta(X)| \le E|\bar{\eta}(X)-\bar{\eta}_\epsilon(X)| + E|\bar{\eta}_\epsilon(X)-\eta_\epsilon(X)| + E|\eta_\epsilon(X)-\eta(X)|,$$

kus

$$\bar{\eta}_\epsilon(x) = E[\eta_\epsilon(Z)|Z \in S(x)].$$

Ülesanne 6.7 *Tõesta, et*

$$E|\bar{\eta}(X)-\bar{\eta}_\epsilon(X)| \le E|\eta_\epsilon(X)-\eta(X)|.$$

Seega, vastavalt η_ϵ definitsioonile, esimese ja kolmanda liidetava summa väiksem kui 2ϵ. Keskmise liidetava hindamiseks kasutame asjaolu, et iga kompaktsel hulgal pidev funktsioon on ühtlaselt pidev. Seega, iga $\epsilon > 0$ korral $\exists\delta > 0$ nii, et kui $\|x-y\| < \delta$, siis $|\eta_\epsilon(x)-\bar{\eta}_\epsilon(y)| \le \epsilon$. Sellest järeldib, et kui $\mathrm{dim}S(x) \le \delta$, siis iga $y \in S(x)$ korral (miks?)

$$|\eta_\epsilon(y)-\bar{\eta}_\epsilon(y)| \le \epsilon.$$

Seega

$$\begin{aligned}\int |\eta_\epsilon(x)-\bar{\eta}_\epsilon(x)|F(dx) &= \sum_i \int_{S_i} |\eta_\epsilon(x)-\bar{\eta}_\epsilon(x)|F(dx)\\ &= \sum_{i:\ \mathrm{dim}S_i<\delta} \int_{S_i} |\eta_\epsilon(x)-\bar{\eta}_\epsilon(x)|F(dx) + \sum_{i:\ \mathrm{dim}S_i\ge\delta} F(dx)\\ &\le \epsilon + \sum_{i:\ \mathrm{dim}S_i\ge\delta} F(dx).\end{aligned}$$

sest $|\eta_\epsilon(x)-\bar{\eta}_\epsilon(x)| \le 1$. Seega

$$E|\eta_\epsilon(X)-\bar{\eta}_\epsilon(X)| = E\big(|\eta_\epsilon(X)-\bar{\eta}_\epsilon(X)|I_{\{\mathrm{diam}S(X)\le\delta\}}\big)+\mathbf{P}(\mathrm{diam}S(X)\ge\delta) \le \epsilon+\mathbf{P}(\mathrm{diam}S(X)\ge\delta).$$

Kokkuvõttes

$$\begin{aligned}E|\eta_n(X)-\eta(X)| &\le E|\eta_n(X)-\bar{\eta}(X)| + E|\bar{\eta}(X)-\eta(X)|\\ &\le \mathbf{P}(N(X)\le k) + \frac{1}{2\sqrt{k}} + \mathbf{P}(\mathrm{diam}S(X)\ge\delta) + 3\epsilon.\end{aligned}$$

Valides piisavalt k suure, saame

$$E|\eta_n(X)-\eta(X)| \le 4\epsilon + \mathbf{P}(\mathrm{diam}S(X)\ge\delta) + \mathbf{P}(N(X)\le k).$$

Vastavalt eeldustele, iga δ ja k korral

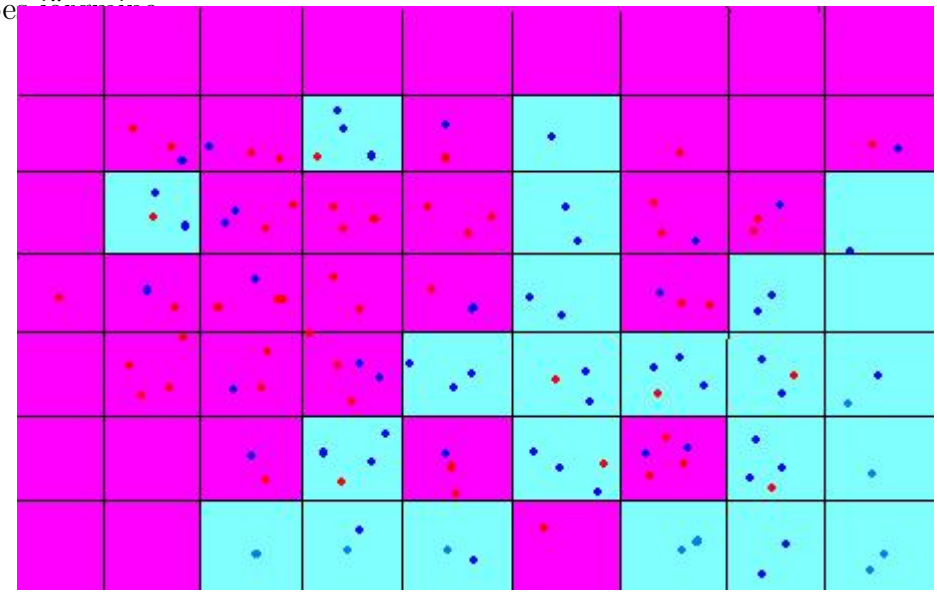
$$\mathbf{P}(\mathrm{diam}S(X)\ge\delta) + \mathbf{P}(N(X)\le k) \to 0.$$

6.2.2 Kuuphistogrammid

Kuuphistogrammid on laialt levinud tükeldusreeglid, kus ruum $\mathbb{R}^d$ tükeldatakse võrdseteks kuupideks küljepikkusega h_n. Seega

$$S_i^n = \prod_{i=1}^{d} [k_i h_n, k_{i+1} h_n),$$

k_i on täisarvud. Tükeldus ei sõltu valimist. Kui $d = 2$, siis kuuphistogramm näeb välja umbes järgmine



Küljepikkusega h_n kuubi diameeter $\sqrt{d}h_n$ ehk teoreemi 6.3 esimene tingimus on täidetud parajasti siis, kui $h_n \to 0$. Teise tingimuse kehtimiseks ei tohi koondumine $h_n \to 0$ olla liiga kiire. Selgub, et piisab, kui $nh_n^d \to \infty$:

Teoreem 6.4 *Kui $h_n \to 0$ ja $nh_n^d \to \infty$, siis kuuphistogramm on universaalselt mõjus.*

Tõestus. Tõestuseks veendume, et kehtivad teoreemi 6.3 eeldused. Et $h_n \to 0$, siis esimene tingimus on täidetud. Veendume, et iga $M < \infty$ korral $\mathbf{P}(N(X) \leq M) \to 0$. Et tükke on lõpmata palju, siis lemmat 6.1 otse kasutada ei saa. Küll aga eraldame kõigist võimalikest tükkidest välja lõpliku arvu nn "põhilisi"tükke ja nende korral kasutame sisuliselt lemmat 6.1. See käib järgnevalt
Olgu M suvaline lõplik konstant. Tähistame $S_n(i)$ histogrammi tükeldused. Tuletame meelde, et X jaotus on F. Tähistame $P(S) := \mathbf{P}(X \in S)$. Vastava empiirilise mõõdu

$$P_n(S) := \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in S\}}.$$

Seega

$$N(x) = nP_n(S(x)) = \sum_{i=1}^{n} I_{\{X_i \in S(x)\}}.$$

Olgu B kuup, mille keskpunkt on 0. Olgu iga n korral

$$I_n = \big\{i : S_n(i) \cap B \neq \emptyset\big\}.$$

Sisuliselt vaatleme edaspidi vaid neid tükke, nende hulk on lõplik. Iga n korral jagame I_n kaheks:

$$J_n := \Big\{ i \in I_n : F(S_n(i)) > \frac{2M}{n} \Big\}, \quad J_n^c := I_n - J_n.$$

Seega J_n koosneb "suurtest tükkidest"ja J_n^c koosneb "väikestest tükkidest"(hulgast I_n).

$$\begin{aligned}
\mathbf{P}(N(X) \le M) &= \sum_{i \in I_n} \mathbf{P}(X \in S_n(i), N(X) \le M) + \mathbf{P}(X \in B^c) \\
&\le \sum_{i \in J_n} \mathbf{P}(X \in S_n(i), N(X) \le M) + \sum_{i \in J_n^c} \mathbf{P}(X \in S_n(i), N(X) \le M) + P(B^c) \\
&\le \sum_{i \in J_n} \mathbf{P}(X \in S_n(i)) \mathbf{P}(N(X) \le M | X \in S_n(i)) + \sum_{i \in J_n^c} \mathbf{P}(X \in S_n(i)) + P(B^c) \\
&= \sum_{i \in J_n} P(S_n(i)) \mathbf{P}(n P_n(S_n(i)) \le M) + \sum_{i \in J_n^c} P(S_n(i)) + P(B^c) \\
&\le \sum_{i \in J_n} P(S_n(i)) \mathbf{P}\Big(P_n(S_n(i)) \le \frac{M}{n}\Big) + \frac{|J_n^c| 2M}{n} + P(B^c) \\
&= \sum_{i \in J_n} P(S_n(i)) \mathbf{P}\Big(P_n(S_n(i)) - P(S_n(i)) \le \frac{M}{n} - P(S_n(i))\Big) + \frac{|J_n^c| 2M}{n} + P(B^c).
\end{aligned}$$

Et iga $i \in J_n$ korral $P(S_n(i)) > \frac{2M}{n}$, millest

$$\frac{M}{n} - P(S_n(i)) \le -\frac{P(S_n(i))}{2}.$$

Et $E\big(P_n(S_n(i))\big) = P(S_n(i))$, siis Tšebõševi võrratusest (tähistame $S = S_n(i)$)

$$\begin{aligned}
\mathbf{P}\Big(P_n(S) - P(S) \le \frac{M}{n} - P(S)\Big) &\le \mathbf{P}\Big(P_n(S) - P(S) \le -\frac{P(S)}{2}\Big) \\
&\le 4 \frac{D\big(P_n(S)\big)}{P^2(S)} \le \frac{4}{nP(S)}.
\end{aligned}$$

Seega

$$\begin{aligned}
\sum_{i \in J_n} P(S_n(i)) \mathbf{P}\Big(P_n(S_n(i)) - P(S_n(i)) \le \frac{M}{n} - P(S_n(i))\Big) &\le 4 \sum_{i \in J_n} P(S_n(i)) \frac{1}{nP(S_n(i))} \\
&\le \frac{4|J_n|}{n} + \mathbf{P}(X \in B^c),
\end{aligned}$$

millest

$$\mathbf{P}(N(X) \le M) \le |I_n|\big(\frac{2M+4}{n}\big) + \mathbf{P}(X \in B^c).$$

Kui kuubi B küljepikkus on B, siis $|I_n| \le \big(\frac{B}{h_n}\big)^d$ ning seega

$$\frac{|I_n|}{n} \le \frac{1}{n}\big(\frac{B}{h_n}\big)^d \to 0,$$

sest $nh_n^d \to \infty$. Et B oli suvaline, teoreem on tõestatud. ■

Selgub, et nimetatud tingimused garanteerivad ka tugeva mõjususe.

Teoreem 6.5 *Kehtigu teoreemi 6.4 eeldused. Siis iga (X, Y) jaotuse ja $\epsilon > 0$ korral leidub n_0 nii, et kuuphistogrammi risk rahuldab võrratust*

$$\mathbf{P}\big(R_n - R^* > \epsilon\big) \le 2e^{\frac{-n\epsilon^2}{32}}.$$

Seega on kuuphistogramm universaalselt tugevalt mõjus.

Boreli-Cantelli lemma kaudu järeldub tõestatud lemmast tugev mõjusus.

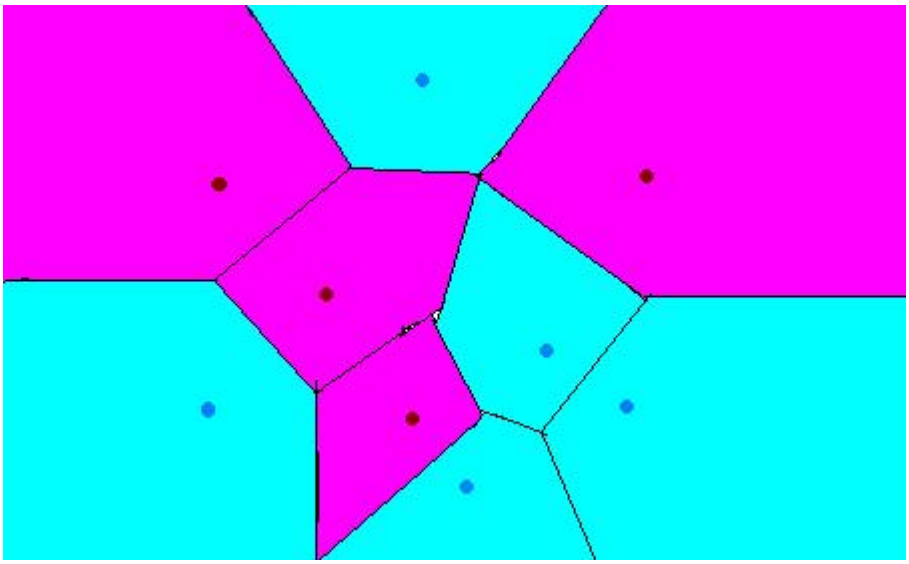
Arv n_o sõltub jaotusest. Seega on tugev mõjusus küll universaalne (s.t. n_o leidub iga jaotuse korral), kuid universaalset hinnangut ei leidu. Nagu teoreemist 2.3 ja võrratusest (2.6.6) teame, universaalset n_o ei saagi leiduda.

6.2.3 Naabrireeglid

Naabrireegli (ik *nearest neighbor rule*) olemus on lihtne ja loomult eestlaslik: tee nii nagu teevad sinu naabrid! Selgub aga, et selline lihtsakoeline lähenemine on klassifitseerimisteooria üks edukamaid meetode. Olgu $x \in \mathbb{R}^d$ klassifitseeritav tunnusvektor. Olgu

$$x_{(1)}, \ldots, x_{(k)} \tag{6.2.5}$$

treeningvalimisse kuuluvad k vektorile x lähimat tunnusvektorit (eukleidilise distantsi mõttes kusjuures võrdsete kauguste korral on lähim vähima indeksiga element) – *naabrid*. Naabrireegel klassifitseerib vektori x klassi 1 parajasti siis, kui naabrite (6.2.5) seas on rohkem klassi 1 kuuluvaid vektoreid (k on paaritu). Vastavalt otsuse tegemiseks vaadeldavate naabrite arvule räägitakse 1 naabri, 3 naabri, 5 naabri reeglist jne. Ühe naabri klassifitseerija võib seega defineerida järgmiselt: treeningvalimisse kuuluvad tunnusvektorid $x_1, \ldots, x_n$ defineerivad ruumil $\mathbb{R}^d$ n-elemendilise tükelduse $\{S_1, \ldots, S_n\}$. Regiooni (hulka) S_i kuuluvad just need $\mathbb{R}^d$ elemendid, millistele x_i on lähim. Tükeldust $\{S_1, \ldots, S_n\}$ nimetatakse **Voronoi tükelduseks** (*Voronoi partition, Voronoi tesselation*).



Igasse Voronoi hulka S_i kuulub vaid üks treeningvalimi element, x_i, ning ning 1 naabri klassifikaatori korral annab see element oma klassi kogu hulgale S_i. Paneme tähele, et 1-naabri reegli korral on treeningviga alati 0.

Naabrite asümptootika fikseeritud k korral. Fikseerime k ning uurime k naabri reegli asümptootilist käitumist. Tuletame meelde, et tõenäosusjaotuse P **kandja** (*support*) on vähim kinnine hulk mõõduga 1. Kui x kuulub jaotuse P kandjasse, siis ja X on jaotusega P juhuslik suurus, siis $\mathbf{P}(X \in B(x,r)) > 0$ iga raadiusega $r > 0$ kera $B(x,r)$ korral. Olgu $X_1, \dots X_n$ jaotusega F juhuslikud vektorid ning tähistagu $X_{(k)}(x)$ punkti x k-ndat naabrit. Seega $X_{(1)}(x)$ on punkti x lähim naaber.

Lause 6.1 *Olgu $X_1, \dots, X_n$ iid juhuslikud vektorid jaotusega P. Kuulugu x jaotuse P kandjasse ning olgu $X_{(k)}(x)$ punkti x k-s naaber. Siis n kasvades*

$$\|X_{(k)}(x) - x\| \to 0 \quad a.s. \tag{6.2.6}$$

Ülesanne 6.8 *Tõestada lause. selleks toimi järgmiselt: olgu B kera, mille keskpunkt on x ja raadius $\delta > 0$. Et x kuulub P kandjasse, siis $\mathbf{P}(X \in B) =: \mu > 0$. Näita, et*

$$\{\|X_{(k)} - x\| > \delta\} = \Big\{\frac{1}{n}\sum_{i=1}^{n} I_B(X_i) < \frac{k}{n}\Big\}.$$

Nüüd, kasutades asjaolu, et iga $0 < \epsilon < \mu$ korral leidub n_o nii, et iga $n > n_o$ korral

$$\Big\{\frac{1}{n}\sum_{i=1}^{n} I_B(X_i) - \mu < \frac{k}{n} - \mu\Big\} \subset \Big\{\frac{1}{n}\sum_{i=1}^{n} I_B(X_i) - \mu < -\epsilon\Big\}$$

tõesta koondumine tõenäosuse järgi:

$$\mathbf{P}(\|X_{(k)}(x) - x\| > \delta) \to 0.$$

järelda p.k. koondumine (6.2.6) kasutades Borteli-Cantelli lemma või kriteeriumi

$$X_n \to X \quad \text{a.s.} \quad \Leftrightarrow \quad \lim_n \mathbf{P}(\sup_{m\geq n} |X_m - X| > \epsilon) = 0, \quad \forall \epsilon > 0. \tag{6.2.7}$$

Järgmine ülesanne üldistab lauset 6.1, sest k võib sõltuda valimi mahust n.

Ülesanne 6.9 *Kuulugu x jaotuse P kandjasse ja $\frac{k(n)}{n} \to 0$. Tõestada, et*

$$\mathbf{P}(\sup_{m\geq n} \|X_{(k(m))}(x) - x\| > \epsilon) \to 0.$$

Järeldada sellest, et

$$X_{(k)}(x) \to x \quad p.k. \tag{6.2.8}$$

Näpunäide: Näita

$$\Big\{ \sup_{m\geq n} \|X_{(k(m))}(x) - x\| > \epsilon \Big\} = \bigcup_{m\geq n} \Big\{ \frac{1}{m} \sum_{i=1}^{n} I_B(X_i) - \mu < \frac{k(m)}{m} - \mu \Big\}.$$

Siis kasuta suurte arvude seadust ja (6.2.7)

Olgu X jaotusega P juhuslik suurus, sõltumatu jadast $X_1, X_2, \ldots$. Seosest (6.2.8) järeldub, et

$$X_{(k)}(X) \to X \quad \text{p.k.} \tag{6.2.9}$$

Kui $x \mapsto \eta(x)$ on pidev, siis piisavalt suure n korral $\eta(x_{(k)}) \approx \eta(x)$ ehk k-s naaber kuulub klassi 1 (peaaegu) sama tõenäosusega, mis x ja see tõenäosus on $\eta(x)$. See tähendab , et $Y_{(k)}$ ($x_{(k)}$ klass) on Bernoulli $\eta(x)$-jaotusega.

Paaritu k

Koondumisest (6.2.8) järeldub, et kui η on pidev, siis iga piisavalt suure n korral

$$\eta(x_{(k)}) \approx \eta(x)$$

nii, et $Y_{(k)}$ – k-nda naabri klass – on ligikaudu Bernoulli $B(1, \eta(x))$ jaotusega. Seetõttu k lähima naabri summa on ligikaudu $B(k, \eta(x))$ haotusega. Järelikult on tõenäosus, et k lähima naabri seas enamuse märk on 0, ligikaudu võrdne arvuga $\mathbf{P}(B < \frac{k}{2})$, kus $B \sim B(k, \eta(x))$ (k on paaritu). Seega paaritu k ja suure n korral on naabrireegli risk ligikaudu

$$\mathbf{P}\Big(B > \frac{k}{2}, Y = 0\Big) + \mathbf{P}\Big(B < \frac{k}{2}, Y = 1\Big), \tag{6.2.10}$$

kus Y on x märk, st $Y \sim B(1, \eta(x))$ ning Y ja B (Bernoulli jaotusega juhuslik suurus) on sõltumatud. Et

$$\mathbf{P}\Big(B > \frac{k}{2}\Big) = \sum_{j=\frac{k}{2}+1}^{k} \binom{k}{j} \eta(x)^j (1-\eta(x))^{k-j}, \quad \mathbf{P}\Big(B < \frac{k}{2}\Big) = \sum_{j=0}^{\frac{k}{2}-1} \binom{k}{j} \eta(x)^j (1-\eta(x))^{k-j},$$

tõenäosus (6.2.10) on

$$\sum_{j=0}^{k}\binom{k}{j}\eta(x)^j(1-\eta(x))^{k-j}\big(\eta(x)I_{\{j<\frac{k}{2}\}}+(1-\eta(x))I_{\{j>\frac{k}{2}\}}\big).$$

Keskmistades üle tunnusvektori X, saame pideva η, paaritu k ja suure n korral, et naabrireegli risk $R(g)$ on ligikaudu

$$\sum_{j=0}^{k}\binom{k}{j}E\Big[\eta(X)^j(1-\eta(X))^{k-j}\big(\eta(X)I_{\{j<\frac{k}{2}\}}+(1-\eta(X))I_{\{j>\frac{k}{2}\}}\big)\Big]=:R_k$$

Selgub, et ülaltoodud idee kehtib isegi siis, kui η pole pidev. Nimelt kehtib järgmine teoreem ([1], Thm 5.2)

Teoreem 6.6 *Olgu k paaritu ja $\{g_n\}$ k naabri reegel. Siis iga (X,Y) jaotuse korral*

$$ER(g_n)\to R_k. \tag{6.2.11}$$

On võimalik näidata, et

$$R^*\le\cdots\le R_{2k+1}\le R_{2k-1}\le\cdots\le R_5\le R_3\le R_1,$$

kusjuures need võrratused on kõik ranged, kui järgmine tingimus kehtib

$$\mathbf{P}(\eta(X)\notin\{0,1,0.5\})>0. \tag{6.2.12}$$

Seega enamike huvipakkuvate jaotuste korral on ülaltoodud võrratused ranged. Võrratustest järeldub, et asümptootiliselt on k naabri reegel seda tõhusam, mida suurem on k. On aga suhteliselt lihtne konstrueerida kontranaäiteid, kus iga n korral 1-naabri reegel on keskmiselt parema riskiga kui k naabri reegel.

Näide. Olgu S_0 ja S_1 kerad raadiusega 1, nende keskpunktide vahe olgu suurem kui 4. Olgu $\mathbf{P}(Y=1)=\mathbf{P}(Y=0)=\frac{1}{2}$ ja tunnuse X tinglik jaotus tingimusel $Y=i$ on ühtlane üle kera S_i. Olgu g_n k-naabri reegel. Kui $k=1$, siis

$$ER(g_n)=\mathbf{P}(Y=0,Y_1=\cdots=Y_n=1)+\mathbf{P}(Y=1,Y_1=\cdots=Y_n=0)=2^{-n}.$$

Kui $k\ge 3$ (paaritu arv), siis

$$\begin{aligned}ER(g_n)&=\mathbf{P}\Big(Y=0,\sum_{i=1}^{n}(1-Y_i)<\frac{k}{2}\Big)+\mathbf{P}\Big(Y=1,\sum_{i=1}^{n}Y_i<\frac{k}{2}\Big)=\mathbf{P}(B<\frac{k}{2})\\&=2^{-n}\sum_{j=1}^{\lfloor\frac{k}{2}\rfloor}\binom{n}{j}>2^{-n},\end{aligned}$$

kus $B\sim B(n,\frac{1}{2})$. Veendu, et $R_k=0$ ja tingimus (6.2.12) ei kehti.

Ühe naabri reegel: Cover-Hart võrratused. Kui $k = 1$, siis riski piirväärtus R_1 on järgmine:

$$R_1 = 2E\big[\eta(X)(1-\eta(X))\big].$$

Tähistame $A(X) := \eta(X) \wedge (1-\eta(X))$ja tuletame meelde, et $R^* = EA(X)$. Jenseni võrratusest saame

$$E[\eta(X)(1-\eta(X))] = E[A(X)(1-A(X))] \le EA(X) - (EA(X))^2 = R^*(1-R^*) \le R^*.$$

Võrratused

$$R_1 \le 2R^*(1-R^*) \le 2R^*$$

on tuntud kui **Cover-Hart võrratused**. Nendest järeldub, et kui Bayesi risk R^* on väike, on seda ka R_1, mis tähendab, et ühe naabri reegel töötab asümptootiliselt üsna hästi. Kui $R^* = 0$, siis $R_1 = R^*$ ehk 1 naabri reegel on mõjus. Teisalt aga kui $0 < R^* < \frac{1}{2}$ ja (6.2.12) kehtib, siis (et iga $\eta \notin \{0, 0.5, 1\}$ korral $2\eta(1-\eta) > \eta \wedge (1-\eta)$) saame

$$\mathbf{P}\big(2\eta(X)(1-\eta(X)) > \eta(X) \wedge (1-\eta(X))\big) > 0.$$

Seega sellisel juhul $R_1 > R^*$ ja ühe naabri reegel pole mõjus.

Näide. Olgu $d = 1$ ja konstrueerime paari (X, Y) järgmiselt: $\mathbf{P}(Y = 1) = \mathbf{P}(Y = 0) = \frac{1}{2}$. Tingimusel $Y = 0$, $X \sim U[0, 1]$. Tingimusel $Y = 1$ on X juhuslik suurus, mille väärtused on ratsionaalarvud hulgal $[0, 1]$ ja iga ratsionaalarv on aatom, st $\mathbf{P}(X = r|Y = 1) > 0$ iga $r \in \mathbb{Q}$ korral. (Sellise jaotusega on näiteks

$$\frac{\min\{Z_1, Z_2\}}{\max\{Z_1, Z_2\}},$$

kus Z_1, Z_2 on sõltumatud geomeetrilise jaotusega juhuslikud suurused.) Sellise (X, Y) korral,

$$\eta(x) = \begin{cases} 1 & \text{kui } x \text{ on ratsionaalarv,} \\ 0 & \text{kui } x \text{ on irratsionaalarv.} \end{cases}$$

Formaalne tõestus seisneb selles, et

$$\int_A \eta(x)F(dx) = \frac{1}{2}\int_A \eta(x)dx + \frac{1}{2}\int_A \eta(x)F_1(dx) =$$
$$\frac{1}{2}\int_A F_1(dx) = \frac{1}{2}\mathbf{P}(X \in A|Y = 1) = \mathbf{P}(X \in A, Y = 1).$$

Seega $R^* = 0$. Paneme tähele, et nii jaotuse F_1 kui ka F_0 kandja on $[0, 1]$ (klasside eristuvus ei tähenda nende kandjate lõikumatust!).
Olgu g_n 1-naabri reegel. Et $R^* = 0$, siis Cover-Harti võrratusest järeldub, et $ER(g_n) \to 0$ ehk

$$E\mathbf{P}(X_{(1)}(X) \text{ on ratsionaalarv }, X \text{ on irratsionaalarv } |D_n) =$$
$$\mathbf{P}(X_{(1)}(X) \text{ on ratsionaalarv }, X \text{ on irratsionaalarv }) \to 0$$
$$E\mathbf{P}(X_{(1)}(X) \text{ on ratsionaalarv }, X \text{ on irratsionaalarv } |D_n) =$$
$$\mathbf{P}(X_{(1)}(X) \text{ on ratsionaalarv }, X \text{ on irratsionaalarv }) \to 0.$$

Kaaludega naabrireegel

Naabrireegli korral igal naabril on võrdne kaal. Samas on loomulik eeldada, et lähimete naabrite kaal peaks olema suurem (tuleta meelde Gaussi tuuma). Nii jõuame **kaaludega naabrireeglini**, kus iga n korral i-nda naabri kaal on w_{ni}. Tüüpiliselt kaal väheneb i kasvades. Reegel on

$$g_n(x) = \begin{cases} 0, & \text{kui } \sum_{i=1}^n w_{ni} I_{\{y_{(i)}=1\}} \leq \sum_{i=1}^n w_{ni} I_{\{y_{(i)}=0\}}; \\ 1, & \text{mujal.} \end{cases}$$

Seega k naabri reegel vastab olukorrale, kus $w_{ni} = \frac{1}{k}$ for $i = 1, \ldots, k$ ja $w_{in} = 0$ mujal. Oletame, et kaalud $w_i = w_{ni}$ on sõltumatud valimi suurusest n ning $w_i = 0$ kui $i > k$, siis (justnagu k naabri reegli korral), leidub konstant $R(w_1, \ldots, w_k)$ so that

$$ER(g_n) \to R(w_1, \ldots, w_k).$$

On võimalik näidata ([1], Thm 5.3), et kui k paaritu, siis $R(w_1, \ldots, w_k) \geq R_k$ ja enamikel juhtudel on see võrratus range. Seega, nende klassifitseerijate seas on standarndne k naabri reegel asümptootiliselt parim.

Paaris k. Paaritu k korral on k naabri reeglid mugavad, sest hääletamisel ei teki viiki. Kui k on paarisarv, tuleb kuidagiviisi hääletustulemus defineerida ka viigi korral. Üks võimalus selleks on järgmine: viigi korral otsusta märk lähima naabri märgi põhjal. Seega

$$g_n(x) = \begin{cases} 1 & \text{kui } \sum_{i=1}^k Y_{(i)} > \frac{k}{2}, \\ 0 & \text{kui } \sum_{i=1}^k Y_{(i)} < \frac{k}{2}, \\ Y_{(1)} & \text{kui } \sum_{i=1}^k Y_{(i)} = \frac{k}{2}. \end{cases}$$

Pane tähele, et see reegel on sama, mis kaaludega naabrireegel kaaludega $(3, 2, \ldots, 2)$. Seega ka paarisarvulise k korral leidub asümptootiline risk R_k. Selgub aga, et $R_k = R_{k-1}$ ([1], Thm 5.5) ehk asümptootiliselt on naabrireegel paarisarvulise k korral on asümptootiliselt sama hea kui $k-1$ korral. Seetõttu praktikas eelistatakse paaritut k.

Mõjusus

Nägime, et k naabri reegel töötab seda paremini, mida suurem on k. Kuid isegi väga suure k korral ei ole k naabri reegel mõjus (välja arvatud juhul, kui $R^* = 0$).

Selgub, et naabrireegel on (universaalselt) mõjus kui k kasvab koos valimi mahuga n, kuid $\frac{k}{n} \to 0$. Kehtib teoreem

Teoreem 6.7 (Stone, 1977) *Kui $k \to \infty$ ja $\frac{k}{n} \to 0$, siis naabrireegel on universaalselt mõjus, s.t. iga (X, Y) jaotuse korral $ER_n \to R^*$.*

Nimetatud teoreem oli ajalooliselt esimene unversaalse mõjususe tõestus.

Selgub aga, et naabrireegel on ka tugevalt mõjus. Tõsi küll, hoolikamalt tuleb tegeleda võrdsete kauguste probleemiga. Seda probleemi ei teki siis, kui X on absoluutselt pideva jaotusega, s.t. tal on tihedus. Sellisel juhul on iga n korral treeningvalimi elemendid paigutatud (peaaegu kindlast) nii, et võrdseid kaugusi nende vahel pole. Saab näidata, et sellisel juhul kehtib teoreemi 6.7 tugevam versioon

Teoreem 6.8 *Olgu jaotus $F(x)$ absoluutselt pidev. Kui $k \to \infty$ ja $\frac{k}{n} \to 0$, siis leidub n_o nii, et iga $n > n_o$ korral*

$$\mathbf{P}(R_n - R^* > \epsilon) \leq 2e^{\frac{-n\epsilon^2}{c}}, \tag{6.2.13}$$

kus c sõltub dimensioonist d kuid mitte jaotusest. Seega on naabrireegel tugevalt mõjus, s.t. $R_n \to R^*$ p.k.

Teoreem 6.8 garanteerib tugeva mõjususe universaalselt üle kõikide selliste (X, Y) jaotuste nii, et X on pidev. Samas ei anna (ega saagi anda, tuleta meeelde teoreemi 2.3 ja võrratust (2.6.6)) võrratus (6.2.13) (kitsendatud) universaalset hinnangut Bayesi riskile R^*, sest n_o sõltub jaotusest (ka jadast k_n).

Juhul, kui X jaotus pole pidev, tuleb tugeva mõjususe saamiseks kasutada senisest rafineeritumaid meetode reegli defineerimiseks võrdsete distantside korral. Senini vaadeldud väiksema indeksi meetod oli piisav teoreemi 6.7 kehtimiseks, kuid ei taga võrratust (6.2.13) (seda saab näidata). Üks alternatiiv on reegel, mille korral võetakse arvesse kõiki neid valimi punkte, mis on punktist x nii kaugel kui k-s naaber. Ainult, et nende punktide märkidest võetakse keskmine, et vähendada nende liigset mõju. Selline reegel on universaalselt mõjus kuid mitte tugevalt mõjus.

Tugeva mõjususe saame, kui võrdsete kauguste probleemi lahendamiseks kasutame lisajuhuslikkust. See tähendab, et valimile $x_1, \ldots, x_n$ genereeritakse juhuslik lisakomponent ning esialgne valim asendatakse valimiga $(x_1, u_1), \ldots, (x_n, u_n)$, kus $u_1, \ldots, u_n$ on i.i.d. pideva jaotusega juhuslik valim. Punktide $(x_1, u_1), \ldots, (x_n, u_n)$ seas võrdseid kaugusi pole (p.k.). Et lisakomponent on genereeritud sõltumatuna, ei muuda selle lisamine Bayesi viga (veendu selles). Saab näidata, et valimile $(x_1, u_1), \ldots, (x_n, u_n)$ rakendatuna kehtib teoreem 6.8. Seega saame sellisel moel universaalse tugevalt mõjusa naabrireegli. See reegel võib oluliselt erineda vaid esialgse valimi põhjal saadud naabrireeglist ning seda isegi siis, kui esialgses valimis pole võrdseid kaugusi.
Teine võimalus on kasutada komponente u_i vaid siis, kui vektorid x_i on võrdsed. Ka sellisel juhul kehtib teoreem 6.8 mis tähendab, et saadud reegel on universaalselt tugevalt mõjus.

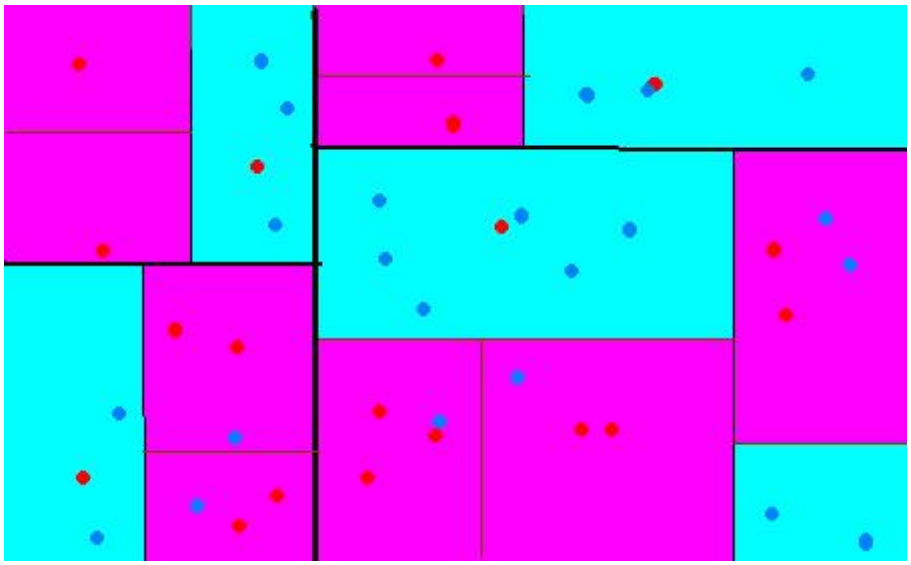
Kirjandus: Naabrireeglitest loe [1], Ch 5,6,11; [7], Ch 3, 13.

6.3 Puud

Klassifitseerimispuu igale lehele vastab ruumi $\mathbb{R}^d$ tükk, need tükid kokku moodustavad tükelduse. Puu struktuur võimaldab enamasti tundmatut x klassifitseerida (st vastava tüki leida) efektiivselt ja seetõttu on klassifitseerimispuud populaarsed.

Osa kasutatavatest klassifitseerimispuudest konstrueeritakse vaid tunnuste $x_1, \ldots, x_n$ abil ja saadud tükkide märk leitakse häälteenamusega. Sellised puud (teinekord nimetatakse neid ka x-omadusega puudeks) on seega tükeldusreeglid. Puude korral on tükeldus hierarhiline – andmete lisamisel võib seda lihtsalt peenendada juba olemasolevate tükkide poolitamisega.

Puude korral on oluline, et liikumine mööda puud oleks võimalikult odav, st punkti $x = (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{R}^d$ klassifitseeimine toimuks "lihtsate"küsimuste abil. *Kahendpuud* on puud, kus igal sõlmel on täpselt kaks järglast. Sellist puud mööda on võimalik liikuda lihtsate jah/ei küsimuste abil. Kõige levinum puu on nn **harilik kahendpuu** (*ordinary binary classification tree*), kus (enamasti kordamööda) vaadeldakse üksikuid koordinaate ja klassifitseerimine toimub küsimuste: Kas $x^{(i)} \leq a$? abil. Sellisele puule vastav tükelduse tekitavadb koordinaattelgedega parallelsed hüpertasandid.



BSP puu (*binary space partition tree*) korral klassifitseeritakse küsimuste $w'x \leq a$? abil. Seega moodustavad tükelduse hüpertasandid.

Sfääripuu (*sphere tree*) korral klassifitseeritakse küsimuste $\|x - z\| \leq a$? abil. Selline puu korral moodustavad tükelduse sfäärid.

Vaatleme mõningaid harilikke kahendpuid.

6.3.1 Mediaanipuu

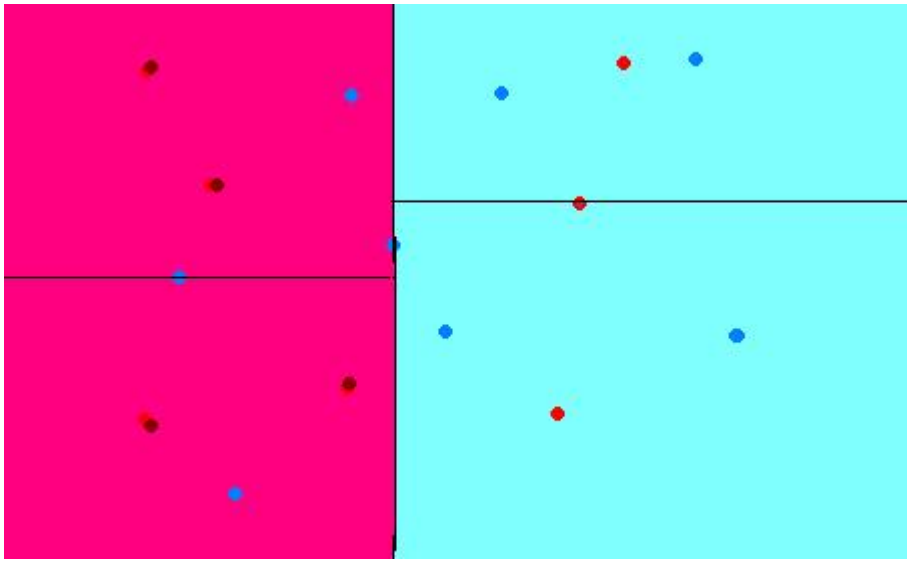
Mediaanipuu garanteerib nn. täispuu (*balanced tree*). Selleks valitakse ühe koordinaadi järgi mediaan (keskmine element). Kui mediaan leidub (paaritu arv), siis ülejäänud $n-1$ elementi jagatakse pooleks (mediaan ise jääb klassifitseerimisest välja), mõlemast poolest leitakse uuesti mediaan teise koordinaadi järgi ja nii edasi. Nii tehakse puule k kihti, kokku 2^k tükki. Igas tükis on vähemalt $\frac{n}{2^k} - k$ punkti, mida kasutatakse hääletamisel. Seega mediaanipuu garanteerib, et igas tükis enam vähem võrdne arv elemente hääletamiseks.

Mediaanipuu on tükeldusreegel. Mõjususe tõestamiseks piisab seega teoreemi 6.3 eelduste kontrollimiseks. Eeldus 2) on triviaalselt täidetud, kui $\frac{n}{2^k} - k \to \infty$ (tükeldust määravad valimi elemendid ja hääletamiseks kasutavad valimi elemendid pole valitud sõltumatult; ettevaatust lemmaga 6.1!). Eeldusega 1) on rohkem tegu, kuid selgub, et $k \to \infty$ piisavalt aeglaselt, siis kehtib ka 1).

Teoreem 6.9 *Mediaanipuu on mõjus , st* $ER(g_n) \to R^*$ *kui*

$$k \to \infty, \quad \frac{n}{k2^k} \to \infty.$$

Tõestus vaata [1], Thm 20.2.



6.3.2 Binary search trees: kronoloogiline k-puu ja k-sügav puu

Binary search trees konstrueeritakse järgmiselt $x_1, \ldots, x_n$ põhjal järgmiselt: x_1 on juur. Seejärel jagatakse ülejäänud valimi elemendid kaheks: ühele poole jäävad need, mille esi-

mene koordinaat on väiksem (suurem) kui x_1 esimene koordinaat. Kummaski grupis järjestatakse originaaljärjestuse järgi, valitakse neist väikseima indeksiga element ja võrreldakse tema teise koordinaadiga ülejäänud grupi elementide teisi koordinaate jne. Seega koordinaate valitakse kordamööda.

Talitades nii lõpuni, saame suure puu, kuid kõik tükid on tühjad. Erinevad võimalused.

Kronoloogiline puu (*chronological k-tree*) konstrueeritakse valimi esimese k elemendi põhjal: seega elemendid $x_1, \ldots, x_k$ kasutatakse puu konstrueerimiseks ja $x_{k+1}, \ldots, x_n$ kasutatakse hääletamiseks.

Teoreem 6.10 *Kronoloogiline puu on mõjus , st $ER(g_n) \to R^*$ kui*

$$k \to \infty, \quad \frac{n}{k} \to \infty.$$

Kronoloogilise puu korral on tükeldust määravad ja hääletamist teostavad valimi elemendid sõltumatud. Seetõttu saab kasutada lemmat 6.1. Et tükelduses on $k+1$ tükki, siis lemma 6.1 tõttu tingimusest $\frac{n}{k} \to \infty$ järeldub teoreemi 6.3 tingimus 2). Tingimus 1) on tõestatud raamatus [1], Thm 20.3.

k-sügav puu (*k-depth tree*) defineeritakse kui kronoloogiline puu, kuid fikseeritud k korral tehakse puu kuni sügavuseni k. Kõik sellel sügavusel olevad sõlmed kuulutatakse lehtedeks

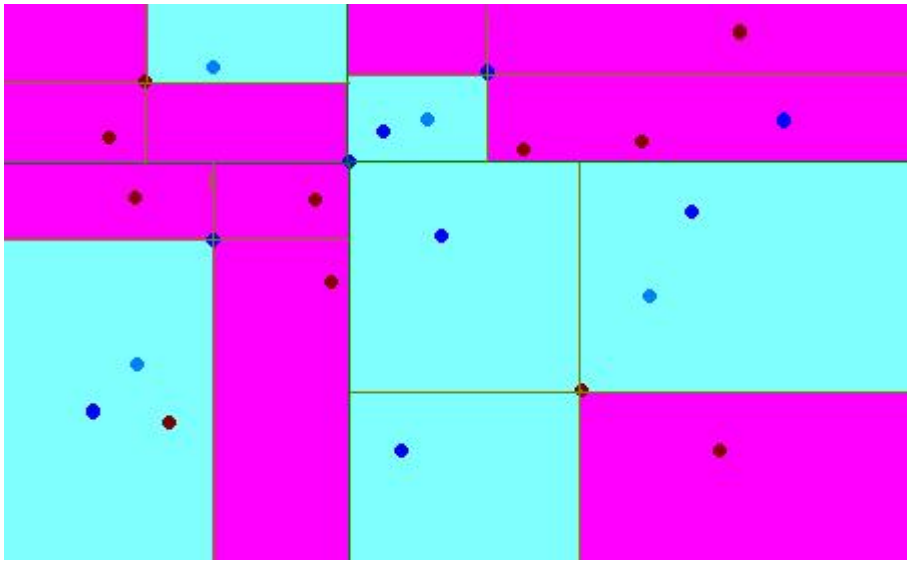
Teoreem 6.11 *k-sügav puu on mõjus , st $ER(g_n) \to R^*$ kui*

$$k \to \infty, \quad \limsup_n \frac{k}{\log n} \leq 2.$$

Tõestus on [1], Thm 20.4.

6.3.3 Quad-puud (Quadtrees)

Need puud pole kahendpuud vaid 2^d-puud. Põhimõte sama: x_1 on juur, temast tõmmatakse läbi koordinaattelgedaga paraleelsed hüpertasandid, mille abil jagatakse ülejäänud valim 2^d tükiks jne. Kasutatakse kronoloogilisi k-puid, st puu tehakse $x_1, \ldots x_k$ põhjal või k-sügavaid puid, kus sügavus on fikseeritud. Veendu, et kronoloogilisel k-puul on $k(2^d - 1) + 1$ tükki.



Teoreem 6.12 *Kronoloogiline quad-puu on mõjus , st* $ER(g_n) \to R^*$ *kui*

$$k \to \infty, \quad \frac{n}{k} \to \infty.$$

Et $\frac{n}{k(2^d-1)+1} \to \infty$, kui $\frac{n}{k} \to \infty$, siis lemmast 6.1 saame tingimuse 1). Tingimuse 2) tõestust vaata [1], Thm. 20.5.

6.3.4 CART-puud

Regressioon. Puid (tükeldusi) võib kasutada ka regressiooniks: antud tükelduse korral on regressioonifunktsioon igal tükil konstantne. Kui kaofunktsioon on ruutfunktsioon (vähimruutude kriteerium), siis konstant on tunnuste keskmine üle tüki. Seega antud tükelduse korral on regressioonifunktsiooni lihtne konstrueerida, küsimus on (optimaalses) puu konstrueerimises.

Tavaliselt kasutatakse ka regressiooniks harilikke kahendpuid (st tükeldus koorinaattelgedega paalleelne), puu püütakse samm sammult konstrueerida nii, et vähimruutude summa kahaneks maksimaalselt. Täpsemalt: olgu iga $j = 1, \ldots, d$ ja $t \in \mathbb{R}$ korral

$$S_1(j,t) := \{x \in \mathbb{R}^d : x^j \le t\}, \quad S_2(j,t) := \{x \in \mathbb{R}^d : x^j > t\}$$

kandidaadid esimesele jagamisele. Otsime sellist jagamist: koordinaati j ja lävendit t, mis garanteerib väiksema vähimruutude summa:

$$\min_{t,j}\Big[\min_{c_1} \sum_{x_i \in S_1(t,j)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in S_2(t,j)} (y_i - c_2)^2\Big]. \tag{6.3.1}$$

On kerge näha, et (6.3.1) on ekvivalentne

$$\min_{t,j}\Big[n_1 \cdot \frac{1}{n_1} \sum_{x_i \in S_1(t,j)} (y_i - \hat{c}_1)^2 + n_2 \cdot \frac{1}{n_2} \sum_{x_i \in S_2(t,j)} (y_i - \hat{c}_2)^2\Big], \tag{6.3.2}$$

kus n_l, $l = 1, 2$ on valimi elementide arv hulgas S_l ja $\hat{c}_l$ on tinglik keskmine üle hulga S_l, s.o.

$$\hat{c}_l := \frac{1}{n_l} \sum_{x_i \in S_l} y_i.$$

"For each splitting variable, the determination of the split point can be done very quickly ..."([7], p. 307.)

Pärast parima j, t leidmist jagatakse valimi elemendid kaheks ja kummaski alamvalimis leitakse omakorda parim j ja t jne. Paneme tähele, et nii saadud puu on küll harilik kahendpuu, kuid koordinaate ei valita ilmtingimata kordamööda. Kui suureks selline puu kasvatada? Esmapilgul võib tunduda, et puud tuleks kasvatada senikaua kui vähimruutude summa kahanemine on väiksem ningist etteantud piirist. Selline meetod on aga lühinägelik, sest pärast suhteliselt väikese kasuteguriga jagamist võib tulla järgmisel tasemel kõrge kasuteguriga jagamine. Nii konstrueeritakse alguses võimalikult suur puu T_0 (näiteks lõpetatakse siis, kui ühte tükki jääb vähem kui 5 elementi., On aga selge, et selline liiga suur puu võib põhjustada ülesobituvust, mistõttu toodud protseduuri on vaja regulariseerida. Selleks vaadeldakse puid T, mis on saadud suurest puust T_0 mingi arvu alampuude **pügamisel** (*pruning*). Sellise puu T korral olgu $|T|$ puu lehtede (tükkide) arv. Üks võimalusi regulariseerida on vähimruutude summale karistusliikme $\lambda|T|$ lisamine. Sellisel juhul on minimiseeritav funktsioon

$$R_\lambda(T) = \sum_{i=1}^{|T|} \sum_{x_j \in S_i} (y_j - \hat{c}_i)^2 + \lambda|T|, \tag{6.3.3}$$

kus $S_1, \ldots, S_{|T|}$ on puule vastav tükeldus. Seega eesmärk on T_o alampuude seast leida (6.3.3) minimiseerivat puud. Kui $\lambda = 0$, on lahend maksimaalne puu T_0.

Weakest link pruning. Alampuu T_λ leidmiseks kasutatakse nn *weakest link pruning* – järk-järgult kustutame T_o alampuid nii, et vähimruutude summa kustutatud lehe kohta kasvaks minimaalselt. Kirjeldame seda protseduuri täpsemalt.
Esimesel sammul otsitakse sellist T_0 alampuud T nii, et suhe

$$\frac{\sum_{i=1}^{|T|} \sum_{x_j \in S_i} (y_j - \hat{c}_i)^2 - \sum_{i=1}^{|T_0|} \sum_{x_j \in S_i^0} (y_j - \hat{c}_i)^2}{|T_0| - |T|} \tag{6.3.4}$$

oleks minimaalne. Ülaltoodud avaldise lugejas on puu pügamisel saadud vähimruutude juurdekasv. Et T on T_0 alampuu, on see vahe alati positiivne. Nimetajas on kustutatud

lehtede arv. T on saadud esialgsest puust T_0 mingist sõlmest t algava alampuu kustutamise kaudu. Olgu T_t sõlmest t algav alampuu ning $S^t_{i_k}$, $k = 1, \ldots, |T_t|$ sellele t-st algavale puule vastavad tükeldused. Pügamise käigus need tükid ühendatakse. Olgu ühendatud tükk S_t, st

$$S_t := \cup_k^{|T_t|} S^t_{i_k}.$$

Olgu c_t tüki tunnuste tinglik keskmine üle tüki S_t. Seega

$$\sum_{i=1}^{|T|} \sum_{x_j \in S_i} (y_j - \hat{c}_i)^2 - \sum_{i=1}^{|T_0|} \sum_{x_j \in S^0_i} (y_j - \hat{c}_i)^2 = \sum_{x_j \in S^t} (y_j - c_t)^2 - \sum_{k=1}^{|T_t|} \sum_{x_j \in S^t_{i_k}} (y_j - \hat{c}_{i_k})^2.$$

Samuti on selge, et $|T_0| - |T| = |T_t| - 1$. Seega, kui T on saadud esialgsest puust puu T_t pügamisel, siis (6.3.4) on

$$g_1(t) := \frac{\sum_{x_j \in S^t} (y_j - c_t)^2 - \sum_{k=1}^{|T_t|} \sum_{x_j \in S^t_{i_k}} (y_j - \hat{c}_{i_k})^2}{|T_t| - 1}. \tag{6.3.5}$$

Esimesel sammul otsitakse sõlm (mitte leht) t_1 nii, et $g_1(t_1)$ oleks minimaalne üle kõikide T_0 sõlmede. Tähistame

$$\lambda_1 := g_1(t_1).$$

Sellest sõlmest algav alampuu kustutatakse, t_1 muutub uue puu leheks. Olgu uus (pügatud) puu T_1. Puu T_1 juur on sama, mis T_0, puu T_1 on T_0 alampuu (kui sõlmi t, mille korral $g(t)$ on minimaalne, on mitu, siis pügatakse kõik vastavad alampuud).
Järgmisel sammul kärbitakse puud T_1 samal meetodil. Selleks defineeritakse iga puu T_1 sõlme, (mitte lehe) korral $g_2(t)$. Funktsioon $g_2(t)$ erineb funktsioonist $g_1(t)$ kõigi t_1 eellaste korral. Nüüd leiame t_2 nii, et $g_2(t_2)$ oleks minimaalne üle kõigi T_1 sõlmede. Sõlmest t_2 algav puu kärbitakse. Nii saame uue puu, tähistame selle T_2; t_2 on T_2 leht. Samuti tähistame

$$\lambda_2 := g_2(t_2).$$

Seejärel kärbitakse puud T_2 jne. Lõpetame puuga mis koosneb vaid juurest. Nii saame üksteisesse sisestatud alampuude jada (ei sõltu λ-st!)

$$T_0 \supset T_1 \supset \cdots \supset T_m = \{\text{juur}\}.$$

Samuti saame jada λ_i, $i = 0, \ldots, m, m+1$, kus $\lambda_0 := 0, \lambda_{m+1} = \infty$ ja $\lambda_i = g_i(t_i)$, $i = 1, \ldots, m$. Selgub, et iga λ korral on kriteeriumi (6.3.3) minimiseeriv puu $T(\lambda)$ jadas $T_0, T_1, \ldots, T_m$. Nimelt kehtib järgmine teoreem (Breiman *et al.*):

Teoreem 6.13 *Konstandid λ_i on kasvavad: $\lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$. Kui $\lambda \in [\lambda_k, \lambda_{k+1})$, siis $T(\lambda) = T(\lambda_k) = T_k$, iga $k = 0, 1, \ldots, m$ korral.*

Tihti jagatakse treeningvalim kaheks: esimese osa põhjal kasvatatakse puu ja teist osa kasutatakse kärpimisel. Sobiva λ leidmine toimub analogiliselt teiste regulariseerimismeetoditega.

Klassifitseerimine. Klassifitseerimine toimub põhimõtteliselt samamoodi kui regressioon: teatava kriteeriumi põhjal teostatakse jagamisi ja konstrueeritakse suur puu T_0. Teine samm on suure puu kahandamine lõpliku suuruseni.

Erinevus on funktsioonis (6.3.1). Ruutkaofunktsiooni asemel kasutatakse funktsiooni $\phi : [0,1] \to \mathbb{R}^+$, mille argument iga S_i korral on

$$\hat{p}(S_i) :=: \hat{p}_i := \frac{1}{n_i} \sum_{x_j \in S_i} y_j.$$

Seega (6.3.2) on üldiselt

$$\min_{t,j} \Big[n_1 \phi\Big(\hat{p}(S_1(t,j))\Big) + n_2 \phi\Big(\hat{p}(S_2(t,j))\Big)\Big]. \tag{6.3.6}$$

Loomulikult sõltuvad n_1 ja n_2 argumentidest t, j. Seoses (6.3.7) olevat funktsiooni nimetatakse *impurity function*. Enamasti üks järgnevatest:

- $\phi(p) = \min(p, 1-p) =$ (empiiriline) risk;
- $\phi(p) = 2p(1-p)$ Gini indeks;
- $\phi(p) = -p \log p - (1-p)\log(1-p)$ binaarne entroopiafunktsioon.

Neist viimased 2 on siledad, esimene mitte.

Näide. Olgu valimis ühtede (mustad) ja nullide (valged) arv vastavalt 400 ja 400 (kirjutame $(400, 400)$). Vaatame kahte konkureerivat jagamist:

- üks jagamine jagab valimi kaheks nii, et mustade ja valgete arv on vastavalt $(300, 100)$ ja $(100, 300)$. Sellisel juhul $\hat{p}_1 = 0.75, n_1 = 400$ ja $\hat{p}_2 = 0.25, n_2 = 400$. Seega empiirilise riski korral
$$n_1 \phi\Big(\hat{p}(S_1)\Big) + n_2 \phi\Big(\hat{p}(S_2)\Big) = 200$$
ja Gini indeks on
$$n_1 2 \frac{1}{4}\frac{3}{4} + n_2 2 \frac{1}{4}\frac{3}{4} = 400 \frac{6}{8}.$$
- Olgu teine jagamine selline, et mustade ja valgete arv on vastavalt $(200, 400)$ ja $(200, 0)$. Sellisel juhul $\hat{p}_1 = \frac{2}{6}, n_1 = 600$ ja $\hat{p}_2 = 1, n_2 = 200$. Seega empiirilise riski korral
$$n_1 \phi\Big(\hat{p}(S_1)\Big) + n_2 \phi\Big(\hat{p}(S_2)\Big) = 200$$
st treeningviga on sama, kuid Gini indeks on nüüd väiksem:
$$n_1 \frac{1}{3}\frac{2}{3} = 600 \frac{2}{9} = \frac{400}{3}.$$
Seega Gini indeks (ja ka entroopiafunktsioon) annab eelistuse teisele jagamisele, sest seeläbi saadi "puhas"tükk.

Nendest argumentidest lähtuvalt soovitavad raamatu [7] autorid T_0 konstrueerimisel kasutada Gini indeksit või entroopiafunktsiooni.

Ka pügamine toimub regressiooniga analoogiliselt. Funktsioon (6.3.3) on nüüd

$$R_\lambda(T) = \sum_{i=1}^{|T|} n_i \phi\big(\hat{p}(S_i)\big) + \lambda|T|. \tag{6.3.7}$$

Seos (6.3.4) on nüüd

$$\frac{\sum_{i=1}^{|T|} n_i \phi(\hat{p}(S_i)) - \sum_{i=1}^{|T_0|} n_i^o \phi(\hat{p}_i(S_i^o))}{|T_0| - |T|} \tag{6.3.8}$$

Funktsioon $g(t)$ avaldub

$$g(t) = \frac{n^t \phi(\hat{p}(S^t)) - \sum_{k=1}^{|T_t|} n_{i_k} \phi(\hat{p}(S_{i_k}))}{|T_t| - 1}, \tag{6.3.9}$$

kus n^t ja n_{i_k} on vastavalt tükeldustesse S^t ja S_{i_k} kuuluvate valimi elementide arv. järgmisest ülesandest järeldub, et kui ϕ on üks ülaltoodud kolmest funktsioonist, on (6.3.9) on alati positiivne.

Ülesanne 6.10 *Olgu ϕ üks ülaltoodud kolmest funktsioonist. Tõestada, et T kärpimisel summa $\sum_{i=1}^{|T|} n_i \phi\big(\hat{p}(S_i)\big)$ ei saa kahaneda, st $g_1(t) \geq 0$.*

Samuti kehtib teoreem 6.13, mistõttu *weakest link pruning* garanteerib alati optimaalse alampuu.

Raamatu [7] autorid soovitavad pügamisel kasutada empiirilist riski. Seega soovitavad nad puu ehitamisel ja pügamisel kasutada erinevaid kriteeriume.

Paneme tähele, et saadud reegel pole tükeldusreegel. Samuti saab näidata, et kui tükeldused on koordinaattelgedega paralleelsed, siis pole CART-puu universaalselt mõjus. Kontranäidet vaata ([1], 20.8).

Rohkem kui kaks klassi. CART-meetod üldistub loomulikult enam kui kahe klassle. Tõepoolest, iga tüki ja iga klassi $m = 0, \ldots, k-1$ korral defineeri klassi m kuuluvate tunnusvektorite proportsioon tükil i järgmiselt:

$$\hat{p}_{i,m} := \frac{1}{n_i} \sum_{x_j \in S_i} I_{\{y_j = m\}}.$$

Ellepool vaadeldud funktsioonid ϕ üldistuvad enam kui k klassile järgmiselt:

- empiiriline risk: $\phi(p_0, \ldots, p_{k-1}) = 1 - \max_{i=0,\ldots,k-1} p_i$;
- Gini indeks: $\phi(p_0, \ldots, p_{k-1}) = \sum_{i=0}^{k-1} p_i(1 - p_i)$;
- entroopia: $\phi(p_0, \ldots, p_{k-1}) = -\sum_{i=0}^{k-1} p_i \ln p_i$.

6.3.5 Bagging ja juhuslik mets

Wisdom of crowd. Olgu meil $h_1, \ldots, h_m$ *sõltumatut* sama jaotusega juhuslikku klassihitseerijat (rusikareeglit), kusjuures mingi x korral $\mathbf{P}(h_i(x) = g^*(x)) = \frac{1}{2} + \epsilon$. Seega kui $g^*(x) = 1$, siis iga juhuslik klassifitseerija h_i teeb õige otsuse $h_i(x) = 1$ tõenäosusega, mis on natuke kõrgem kui $\frac{1}{2}$. Defineerime klassifitseerija g_m rusikareeglite häälteenamuse põhjal, st saame

$$g_m(x) := I_{[\frac{1}{2},1]}\Big(\frac{1}{m}\sum_{i=1}^{m} h_i(x)\Big) = \begin{cases} 1, & \text{kui } \sum_{i=1}^m h_i(x) \geq \frac{m}{2}; \\ 0, & \text{kui } \sum_{i=1}^m h_i(x) < \frac{m}{2}. \end{cases} .$$

Suurte arvude seadusest saame, et hääletajate arvu m kasvamisel tõenäosus, et häälteenamusel võetakse vastu otsus 0 väheneb:

$$P(g_m(x) = 0) = P\Big(\frac{1}{m}\sum_{I=1}^{m} h_i(x) < \frac{1}{2}\Big) \to 0.$$

Seega, kui igaüks otsustuskogust võtab vastu õige otsuse suurema tõenäosusega kui 0.5 ja kui otsustajad on sõltumatud, siis võetakse häälteenamusega vastu õige otsus üsna suure tõenäosusega (ligikaudu 1), see tõenäosus on seda suurem, mida suurem on otsustajate arv.

Kui nüüd $h_i : \mathbb{R}^d \to \mathbb{R}$ on sõltumatud ja sama jaotusega juhuslikud regerssioonifunktsioonid ja iga x korral $h(x) := Eh_i(x) = Eg_m(s)$ siis iga y korral

$$E(y - g_m(x))^2 = E(y - h(x) + h(x) - g_m(x))^2 = (y - h(x))^2 + E(h(x) - g_m(x))^2.$$

SAS põhjal m kasvades $g_m(x) \to h(x)$ ja nii näeme jalle, et m kasvades prognoositäpsus suureneb.

Nägime, et iid juhuslikke klassifitseerijaid kombineerides (või regressioonifunktsioone keskmistades) võime saada üsna hea tulemuse ka siis, kui iga funktsioon omaette võetuna eriti hea pole. Kuidas aga saada piisavalt palju sõltumatuid klassifitseerijaid? Kui meie käsutuses oleks m sõltumatut valimit, oleks asi lihtne – iga valimi põhjal saaksime (mingil meetodil) klassifitseerija ja isegi kui ta pole suurem asi, neid üle kõikide valimite kombineerides saaksime ikkagi hea tulemuse. Enamasti on meil aga üks valim. **Bagging** (*bootstrap aggregation*) püüab lahendada seda probleemi järgmiselt: vaatle antud valimit empiirilise mõõduna P_n ja genereeri sealt m sõltumatut *bootstrap*-valimit mahuga (igaüks mahuga n, tagasipanekuga). Kui n on suur, siis $P_n \approx P$ ja nii võib neid *bootstrap*-valimeid vaadelda kui valimeid jaotusest P. Pane tähele, et kuigi iga *bootstrap*-valimi maht on samuti n, pole nad üldiselt esialgse valimi koopiad, st neis olevate erinevate elementide arv on väiksem. Seega on erinevad ka *bootstrap*-valimite põhjal konstrueeritud klassifitseerijad h_i. Lõplik klassifitseerija g_m tehakse ikka häälteenamuse põhjal (regressiooni korral keskmistades). Enamasti kasutataksegi baggingus CART puid (tõsi, tihti pügamata). Boosting on teatavas mõttes baggingu edasiarendus ning simulatsioonid on näidanud, et enamasti annab boosting parema tulemuse kui bagging (vt [8], ptk 15).

Juhuslik mets. Et *bootstrap*-valimid on genereeritud empiirilises mõõdust P_n, mitte (meile tundmatust) jaotusest P, on tegelikult saadud puud korreleeritud, st $h_1, \ldots, h_m$ pole sõltumatud. Täpsemalt, ühest valimist genereeritud puud on tinglikult (antud originaalvalimi korral) sõltumatud, kuid mitte tingimatult sõltumatud – nad kõik sõltuvad originaalvalimist. **Juhuslik mets** (*random forest*) lisab puude konstrueerimisele juhuslikkust ja nii loodetakse suurendada puude omavahelist sõltumatust ja vähendada puude kombineerimisel saadud klassifitseerija sõltuvust originaalvalimist. Selleks modifitseeritakse CART-puu kasvamist järgmisel moel: igal sammul valitakse d tunnusest (koordinaadist) juhuslikult r-elemendiline ($r < d$) alamhulk ja seejärel leitakse parim jagamine valitud tunnuste seast. Seega, kui $r = d$, on juhuslik mets ja bagging üks ja seesama. Järgmisel jagamisel leitakse uus almhulk jne. Nii kasvatatakse puu teatud suuruseni, pügamist pole. Kui r on oluliselt väiksem kui d, siis sõltub igal jagamisel kasutatavate tunnuste hulk palju juhusest ja nii võivad olla saadud puud väga erinevad ka siis, kui nende konstrueerimiseks kasutatavad *bootstrap*-valimid on väga sarnased või isegi võrdsed. Raamatus [8] soovitatakse klassifitseerimise korral võtta $r = \lfloor \sqrt{d} \rfloor$ (või isegi 1), regressiooni korral $r = \lfloor \frac{d}{3} \rfloor$.

Olgu

$$\hat{g}(x) = \lim_m g_m(x),$$

st $\hat{g}$ on vaid originaalvalimist kuid mitte *bootstrap*-valimitest sõltuv klassifitseerija, mida juhuslik mets g_m lähendab. Klassifitseerija $\hat{g}$ pole (üldiselt) puu (ka siis kui $r = d$) ja seega erineb $\hat{g}$ oluliselt vaid originaalvalimi põhjal tehtud puust, olgu see g_n. See erinevus on seda suurem, mida juhuslikum on juhuslik mets (mida väiksem on r). Järelikult juhuslikkuse suurenemisel (r vähenemisel) kaugeneb (teatavas mõttes) $\hat{g}$ originaalpuust g_n ja sõltub vähem originaalvalimist. Teisisõnu – erinevatele valimitele vastavad $\hat{g}$ võivad olla sarnasemad kui erinevatele valimitele vastavad originaalpuud. Ja see tähendab, et juhusliku suuruse $\hat{g}(x)$ dispersioon võib olla väiksem kui $g_n(x)$ dispersioon. Väidetavalt on CART puud tuntud kui suure hajuvusega klassifitseerijad ja regressioonifunktsioonid ning selle hajuvuse vähenamiseks tihti juhuslikku metsa kasutataksegi.

Kirjandus: CART-puudest loe [8], Ch 9; [9], Ch 7. [10], 16. Teistest puudest loe[1], Ch 20. Baggingust loe [8], Ch 8, juhuslikust metsast loe [8], Ch 15, samuti [9] 8.4 ja [10], 16.

6.4 Närvivõrgud klassifitseerimises

6.4.1 Närvivõrk

Varjatud kihita närvivõrk on sisuliselt lineaarne klassifikaator

$$\begin{cases} 0 & \text{kui } \sum_{i=1}^d c_i x^i + c_o = c'x + c_o \leq 0.5, \\ 1 & \text{mujal.} \end{cases} \tag{6.4.1}$$

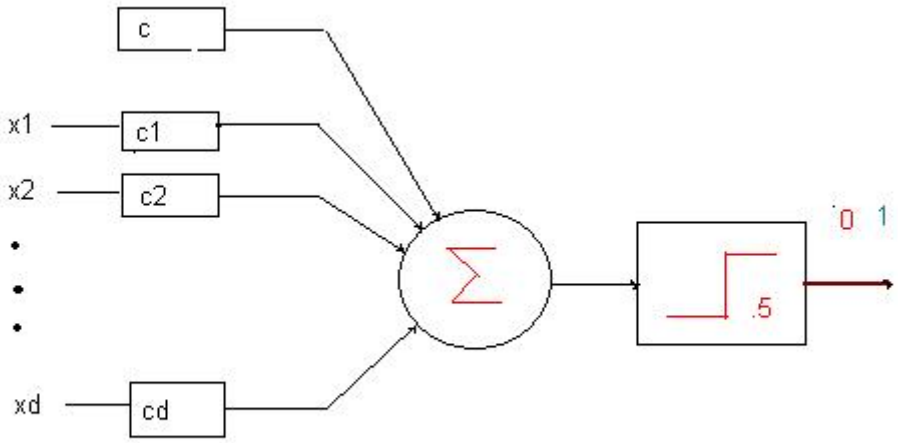
Tihti tähistatakse

$$\psi(x) = \sum_{i=1}^{d} c_i x^i + c_0 \tag{6.4.2}$$

ja nii on (6.4.1)

$$g(x) = \begin{cases} 0 & \text{kui } \psi(x) \leq 0.5, \\ 1 & \text{mujal.} \end{cases} \tag{6.4.3}$$

Närvivõrkude kontekstis esitatakse toodud klassifitseerijat enamasti kujul



Ühe varjatud kihiga närvivõrk on klassifitseerija kuhul (6.4.7), kus ψ defineeritakse

$$\psi(x) = \sum_{i=1}^{k} w_i \sigma(\psi_i(x)) + w_0. \tag{6.4.4}$$

Siin iga i korral ψ_i on lineaarne funktsioon kujul (6.4.2), kus vektor c sõltub i-st. **Sõlmfunktsioon** (*sigmoid*) σ on üldiselt mittekahanev funktsioon nii, et

$$\lim_{x\to\infty} \sigma(x) = 1, \quad \lim_{x\to-\infty} \sigma(x) = -1.$$

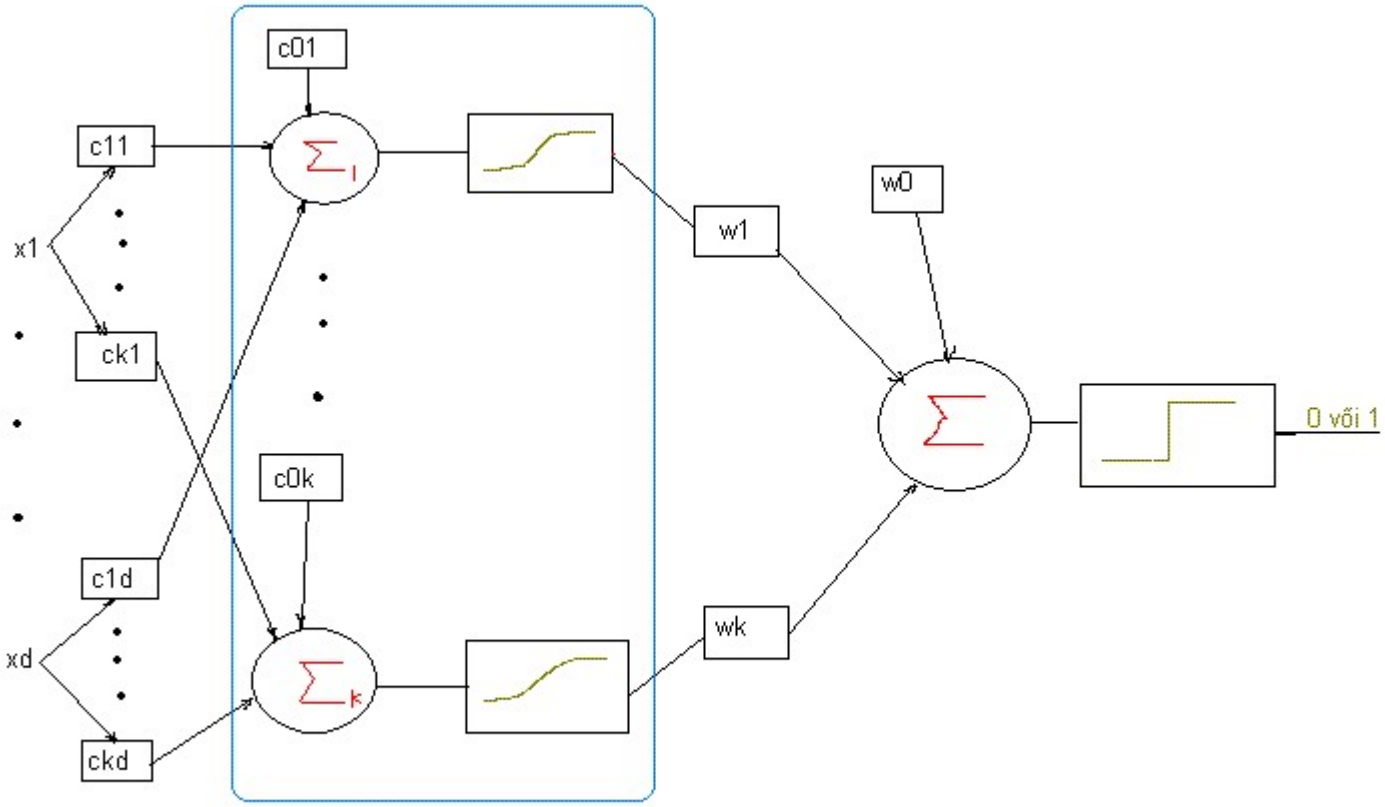
Tüüpiliselt on σ:

- lävefunktsioon (*treshold*)

$$\sigma(x) = \begin{cases} -1, & \text{kui } x \leq 0; \\ 1, & \text{kui } x > 0. \end{cases}$$

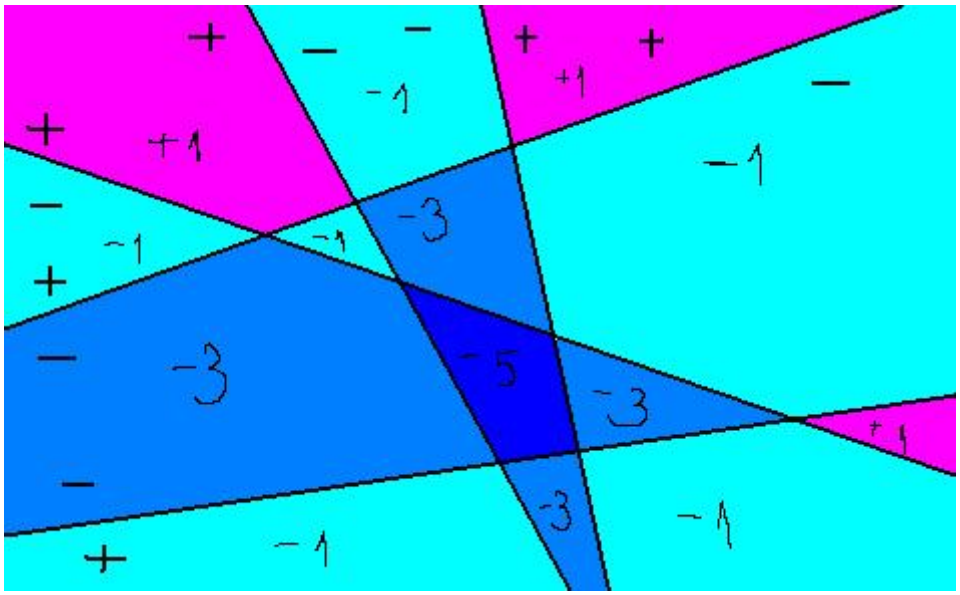
- logistiline sõlmfunktsioon
$$\sigma(x) = \frac{1 - e^{-x}}{1 + e^{-x}}$$
- arctan sõlmfunktsioon
$$\sigma(x) = \frac{2}{\pi} \arctan(x)$$
- Gaussi sõlmfunktsioon
$$\sigma(x) = 2\Phi(x) - 1,$$
kus $\Phi(x)$ on standardse normaaljaotuse jaotusfunktsioon.

Graafiliselt esitub ühe varjatud kihiga närvivõrk järgmiselt.



Paneme tähele, et kui sõlmfunktsioon on lävefunktsioon, siis ühe varjatud kihiga närvivõrk on sisuliselt juba boostingust tuttav lineaarsete klassifikaatorite kombinatsioon. Lävefunktsioon pole sile, ülejäänud sõlmfunktsiooni esindavadki lävefunktsiooni siledaid "versioone". Funktsioone $\sigma(\psi_1(x)), \ldots, \sigma(\psi_k(x))$ nimetatakse **varjatud neuroniteks**, ühe varjatud kihiga närvivõrgul (6.4.4) on seega k varjatud neuronit.

Erijuhu ühe varjatud kihiga närvivõrkudest moodustavad nn, **komitee-reeglid** *committee machines*, kus $w_0 = 0$ ja $w_1 = \cdots = w_k = 1$ ja sõlmfunktsioon on lävefunktsioon. Seega komitee-reegel on sisuliselt järgmine $g_1, \ldots, g_k$ on lineaarsed klassifitseerijad (väljunditega +1, -1), nende väärtused kohal x liidetakse kokku ja otsustatakse summa märgi põhjal (häälteenamus).



Pildi pealt on hästi näha, et kui σ on lävefunktsioon, siis ühe varjatud kihiga närvivõrgul põhinev klassifitseerija (6.4.4) on konstantsed tükeldusel, mille defineerivad hüpertasandid $H_1, \ldots, H_k$, kus

$$H_j = \{x : \psi_j(x) = 0\}, \quad j = 1, \ldots, k,$$

sest varjatud neuronite väljund

$$\sigma(\psi_1(x)), \ldots, \sigma(\psi_k(x)) \tag{6.4.5}$$

on binaarne vektor, mis üheselt määrab ära hüpertasandite poolt moodustatud tükelduse selle tüki, kuhu x kuulub. Edaspidine määrab vaid tüki märgi. Sellist hüpertasandite kaudu defineeritud klassifitseerijat nimetatakse teinekord *arrangement classifier*. Kas selline klassifitseerija on tükeldusreegel või mitte, sõltub sellest, mis põhimõttel parameetrid valitakse (üldiselt mitte).

Kahe varjatud kihiga närvivõrk on klassifitseerija kuhul (6.4.7), st

$$g(x) = \begin{cases} 0 & \text{kui } \psi(x) \le 0.5, \\ 1 & \text{mujal.} \end{cases}$$

kus ψ defineeritakse analoogiliselt valemiga (6.4.4):

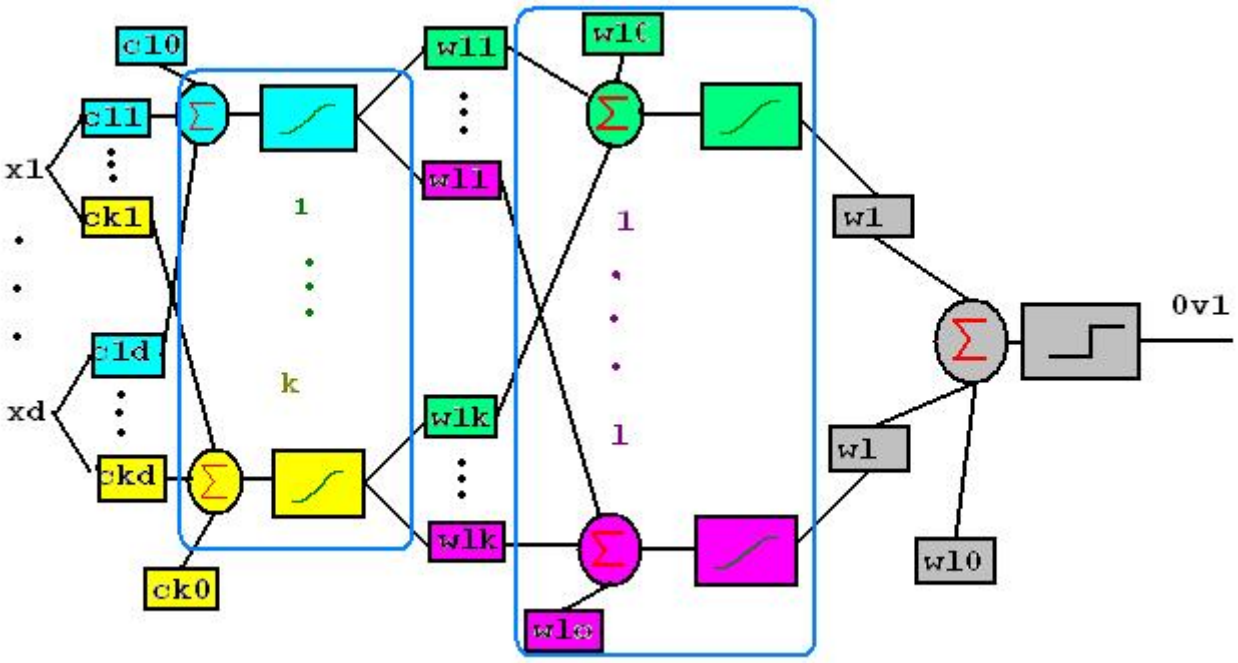
$$\psi(x) = \sum_{i=1}^{l} w_i \sigma(\psi_i(x)) + w_0, \tag{6.4.6}$$

kuid ka ψ_i on nüüd kujul (6.4.4):

$$\psi(x) = \sum_{i=1}^{k} w_i \sigma(\psi_i(x)) + w_0.$$

Seega kahe varjatud kihiga närvivõrk (6.4.6) kombineerib (kaaludega) omakorda l ühe varjatud kihiga närvivõrku. Närvivõrgu (6.4.6) esimeses varjatud kihis on k ja teises varjatud kihis l neuronit. Kombineerides kahe (varjatud) kihiga närvivõrgud, saame kolmekihilise närvivõrgu jne.

Graafiliselt esitub kahe varjatud kihiga närvivõrk järgmiselt.

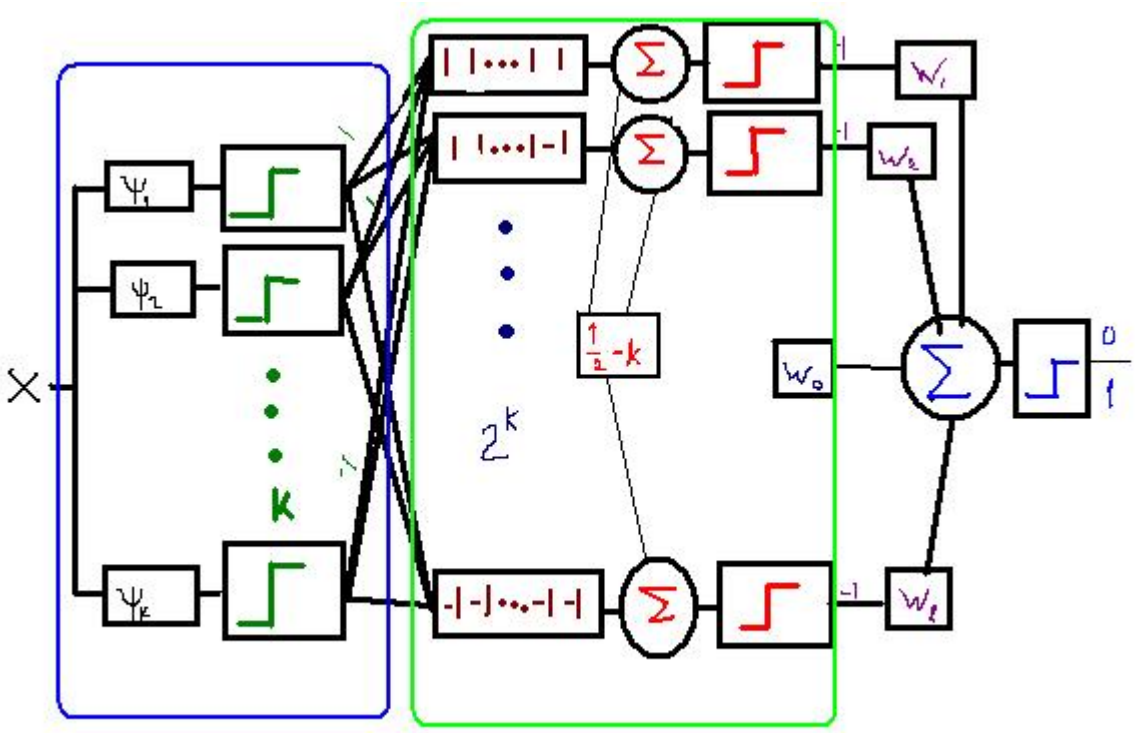


6.4.2 Arrangement-klassifitseerijad

Vaatleme kahe (või enama) varjatud kihiga närvivõrku, kus esimeses kihi sõlmfunktsioonid on kõik lävefunktsioonid. Sellisel juhul esimese kihi neuronite väljund on ikkagi vektor kujul $\sigma(\psi_1(x)), \ldots, \sigma(\psi_k(x))$, mis on konstantne igal tükil. Ükskõi mida edaspidised kihid

ka ei tee tulemus on alati konstantne hüpertasandite $H_1, \ldots, H_k$ defineeritud tükeldustel. Seega iga klassifitseerija (6.4.7), kus ψ on närvivõrk, mille esimeses kihis on k neuronit ja σ on lävefunktsioon on hüpertasandite $H_1, \ldots, H_k$ kaudu defineeritud *arrangement* klassifitseerija. Muidugi ei pruugi see tükeldus alati olla rikkaim võimalik tükeldus. Näitaks on vaid ühe hüpertasandi H_1 abil defineeritud lineaarne klassifitseerija ka üks võimalikest $H_1, \ldots, H_k$ tekitatud tükeldus-klassifitseerijatest, kuid mitte kõige rikkam.
Selliseid närvivõrke, mille esimeses kihis on k neuronit on lõpmatu hulk. Tasandite $H_1, \ldots, H_k$ abil defineeritud tükeldusi on aga lõplik arv. Intuitiivselt peaks nüüd olema selge, et mitmed mitmekihilised närvivõrk-klassifitseerijad on ekvivalentsed ühe ja sama *arrangement*-klassifitseerijaga.

Arrangement-klassifitseerija kui kahe varjatud kihiga närvivõrk. Püstitame nüüd küsimuse *arrangement*-klassifitseerijate hulga ja närvivõrkude hulga üksühelisest vastavusest järgmiselt: Milline on väikseim vajaminev neuronite ja kihtide arv, et esitada närvivõrgu kaudu suvalist *arrangement*-klassifitseerijat?



Olgu g suvaline k hüpertasandi $H_1, \ldots, H_k$ kaudu defineeritud *arrangement*-klassifitseerija. Veendume, et selle saab esitada kahe varjatud kihiga närvivõrgu abil, kus esimeses kihi on k neuronit ja teises ülimalt 2^k neuronit. Sobiva närvivõrgu defineerime järgmiselt: esimeses kihis on k neuronit $\sigma(\psi_1(x)), \ldots, \sigma(\psi_k(x))$, kus σ on lävefunktsioon. Olgu $b = (b_1, \ldots, b_k)$, $b_i \in \{+1, -1\}$ esimese kihi väljund, st

$$(\sigma(\psi_1(x)), \ldots, \sigma(\psi_k(x))) = (b_1, \ldots, b_k).$$

Olgu teises kihis 2^k neuronit, igaühes neist olgu kaaluvektor $a = (a_1, \ldots, a_k) \in \{+1, -1\}^k$, $a_o = -k + \frac{1}{2}$. Iga teise kihi neuron vastab ühele hüpertasandite $H_1, \ldots, H_k$ tekitatud tükelduse tükile. Muidugi, sõltuvalt tükeldusest, võib teises kihis olla ka vähem kui 2^k neuronit. Kuid alati võib teise kihi neuroneid olla ka 2^k, ning kui tükke ongi vähem, siis mõned neist ei lihtsalt ei aktiviseeru kunagi. Lihtsuse mõttes eeldame siinkohal, et tükeldus koosneb 2^k tükist.
Nüüd on selge, et

$$a_o + \sum_{i=1}^{k} b_i a_i \begin{cases} = \frac{1}{2}, & \text{kui } a = b; \\ \leq -\frac{3}{2}, & \text{kui } a \neq b; \end{cases}$$

Seega lävefunktsiooni korral

$$\sigma(a_o + \sum_{i=1}^{k} b_i a_i) = \begin{cases} 1, & \text{kui } a = b; \\ -1, & \text{kui } a \neq b. \end{cases}$$

Järelikult teise kihi väljund on 2^k-dimensionaalne binaarne vektor $u = (u_1, \ldots, u_l)$ ($l = 2^k$), kus vaid üks komponent on $+1$, ülejäänud on -1. Iga komponent vastab ühele klassifitseerija g tükile; kui $u_i = 1$, siis kuulub sisend x komponendile i vastavasse tükki, ütleme, et see tükk on aktiivne. Olgu g väljund 1 tükkidel $P \subset \{1, \ldots, l\}$. Defineerime närvivõrgu väljundi

$$\psi = w_o + \sum_{i=1}^{l} w_i u_i,$$

kus $w_i = 1$, kui $i \in P$ ja $w_i = -1$ mujal ning $w_o = |P| - (l - |P|) = 2|P| - l$.
Kui $i \in P$ on aktiivne, siis $\psi = (l - |P|) + 1 - (|P| - 1) + |P| - (l - |P|) = 2$. Kui $i \notin P$, siis $\psi = (l - |P|) - 1 - (|P| + 1) + |P| - (l - |P|) = -2$. Sellisel juhul seosega (6.4.1) defineeritud klassifitseerija annab 1 parajasti siis, kui x on selline, et hulka P kuuluv indeks on aktiivne. See aga tähendab, et x kuulub sellisesse tükki, mille korral tükeldus-klassifitseerija g annab väärtuseks 1. Seega oleme konstrueerinud kahe varjatud kihiga närvivõrgu, mis on ekvivalentne esialgse klassifitseerijaga g.

Kokkuvõttes: Iga *arrangement*-reegel on ekvivalentne teatud kahe varjatud kihiga närvivõrguga. Ehk kahe varjatud kihiga närvivõrkude klass on piisavalt lai sisaldamaks kõiki *arrangement*-reegleid. Järelikult, kui kasutada vaid lävefunktsioone, siis (klassifitseerimise seisukohast) ei anna ülejäänud kihid midagi juurde.

Et (lävefunktsiooni korral) on närvivõrgud sisuliselt *arrangement*-reeglid, tasub pilku heita nende omadustele. Järgmine teoreem ([1], Thm 30.1) väidab, et kui hüpertasandite arv kasvab kontrollitult, on empiirilist riski minimiseeriv *arrangement*-reegel universaalselt mõjus.

Teoreem 6.14 *Olgu g_n ülimalt k hüpertasandist konstrueeritud empiirilist riski minimiseeriv arrangement-klassifitseerija. Siis iga (X, Y) jaotuse korral $ER(g_n) \to R^*$, kui $k = o(\frac{n}{\ln n})$.*

6.4.3 ühe varjatud kihiga närvivõrk

Olgu $\mathcal{C}^{(k)}$ k varjatud neuroniga ühe varjatud kihiga närvivõrkude hulk. Täpsemalt, $\mathcal{C}^{(k)}$ on klassifitseerijad kujul (6.4.7):

$$g(x) = \begin{cases} 0 & \text{kui } \psi(x) \leq 0.5, \\ 1 & \text{mujal,} \end{cases} \tag{6.4.7}$$

kus $\psi(x)$ on kujul (6.4.4):

$$\psi(x) = \sum_{i=1}^{k} w_i \sigma(\psi_i(x)) + w_0.$$

ja ψ_i on lineaarne funktsioon. Olgu $\mathcal{F}_k$ nende funktsioonide hulk

Lähendamisviga

Eelpool nägime, et teatud tingimustel on kahe varjatud kihiga närvivõrk mõjus. Kas see kehtib ka ühe varjatud kihiga närvivõrgu korral? Mõjususeks peab nulliks koonduma nii lähendamisviga kui ka hindamisviga. Järgnevas uurime lähendamisviga

$$\inf_{g \in \mathcal{C}^{(k)}} R(g) - R^*.$$

Selgub, et iga sõlmfunktsiooni korral lähendavad ühekihilised närvivõrgud $\mathcal{C}^{(k)}$ sõlmede arvu k kasvades kuitahes hästi suvalist funktsiooni. Seetõttu läheneb lähendamisviga nullile ehk kehtib järgmine teoreem ([1], Cor. 30.1).

Teoreem 6.15 *Iga sõlmfunktsiooni σ ja (X, Y) jaotuse korral kehtib*

$$\lim_{k \to \infty} \inf_{g \in \mathcal{C}^{(k)}} R(g) - R^* = 0. \tag{6.4.8}$$

Püäme sellele natuke valgust heita sellele tulemusele. Definitsioonist järeldub, et klassifitseerija g on plug-in klassifitseerija. Seosest (6.1.3) järeldub, et iga

$$R(g) - R^* \leq 2 \int |\eta(x) - \psi(x)| F(dx).$$

Seega lähendamisviga koondub nulliks (st kehtib (6.4.8)) kui kujul (6.4.4) olevate funktsioonide klass, olgu see (üle kõikide k-de) on kõikjal tihe ruumis L_1. See aga on nii (ning selles pole raske veenduda), kui kõnealune funktsioonide klass on L_∞-kõikjal tihe pidevate funktsioonide ruumis $C[a, b]^d$ *iga* a ja b korral. Järelikult (6.4.8) kehtib, kui iga a ja b ja iga tõkestatud pideva funktsiooni g korral

$$\lim_{k \to \infty} \inf_{\mathcal{F}_k} \sup_{x \in [a,b]^d} |\psi(x) - g(x)| = 0. \tag{6.4.9}$$

Selgub, et (6.4.9) kehtib – igat pidevat ja tõkestatud funktsiooni g saab igal kuubil $[a, b]^d$ lähendada kuitahes hästi ühekihiliste närvivõrkudega, kui peidetud neuronite arv on piisavalt suur.

Koondumine (6.4.9), kui $d = 1$. Kui $d = 1$ ja σ on lävefunktsioon:

$$\sigma = \begin{cases} -1 & \text{kui } x \leq 0, \\ 1 & \text{mujal,} \end{cases} \tag{6.4.10}$$

siis koondumine (6.4.9) üsna loomulik: saab ju igat pidevat tõkestatud funktsiooni lõigul $[a, b]$ lähendada ühtlaselt tükati konstantsete funktsioonidega. Iga selline funktsioon on aga nn lihtne funktsioon kujul

$$\sum_{i=1}^{n} a_i I_{(c_i, d_i]}, \quad a_i \in \mathbb{R}.$$

Iga indikaator on aga esitatav kujul (6.4.4) (varjatud neuronite arvuga $2n$):

$$\sum_{i=1}^{2n} w_i \sigma(\psi_i(x)) + w_0,$$

sest

$$I_{(c,d]} = \frac{1}{2}\sigma(x - c) + \frac{1}{2}\sigma(-x + d),$$

millest

$$\sum_{i=1}^{n} a_i I_{(c_i, d_i]} = \sum_{i=1}^{n} a_i \Big(\frac{1}{2}\sigma(x - c_i) + \frac{1}{2}\sigma(-x + d_i) \Big).$$

Seega varjatud neuronite arvu kasvamisel koondub lähenemisviga nulliks. See kohtumine on universaalne, kuid seetõttu ei saa (üldiselt) midagi öelda koondumiskiiruse kohta, see võib olla kuitahes aeglane. Väitmaks midagi koondumiskiiruse kohta, tuleb teha eeldusi η kohta.

Pole raske veenduda, et ka komitee-reegli lähenemisviga koondub nulliks (neuronite arvu kasvamisel).

Empiirilise riski minimiseerimine: hindamisviga ja mõjusus

Teoreem 6.15 motiveerib järgmist küsimust: kas ühekihiliste võrkude korral on reegel $\{g_n\}$ on universaalselt mõjus, kui g_n minimiseerib empiirilise riskifunktsiooni üle $\mathcal{C}^{(k)}$ ja k kasvab? Teame, et mõjusus tähendab nii hindamisvea kui ka lähendamisvea koondumist nulliks, kui n ja kasvab. Lähendamisviga koondub nulliks, nii väidab teoreem 6.15. See tähendab, et klass $\mathcal{C}^{(k)}$ pole (k kasvamisel) liiga kitsendatud. Samas liiga suure \ komplekse klassi korral on raske hinnata hindamisviga – empiirilist riski minimiseeriv klassifitseerija g_n võib kergesti olla selline, et R_n on liiga suur. Selleks, et hinnata hindamisviga, on vaja teada või hinnata klassi $\mathcal{C}^{(k)}$ VC-dimensiooni. Seosetest (2.6.13) ja (2.6.14) saame antud

juhul

$$ER(g_n) - \inf_{g\in\mathcal{C}^{(k)}} R(g) \le 4\sqrt{\frac{V_k \ln(n+1) + \ln 2}{n}}; \tag{6.4.11}$$

$$\mathbf{P}\Big(|R(g_n) - \inf_{g\in\mathcal{C}^{(k)}} R(g)| > \epsilon\Big) \le 8n^{V_k} e^{\frac{-n\epsilon^2}{128}} \tag{6.4.12}$$

kus V_k on $\mathcal{C}^{(k)}$ dimensioon. Seega on $\{g_n\}$ universaalselt mõjus, kui

$$\frac{V_{k(n)} \ln n}{n} \to 0.$$

(siin k ja seega ka $V_{k(n)}$ kasvab koos n-iga). See garanteerib universaalse mõjususe. Samas tähendab see, et $V_{k(n)}$ kasvab koos n-ga ka nii aeglaselt, et

$$\sum_n n^{V_k} e^{\frac{-n\epsilon^2}{128}} < \infty, \tag{6.4.13}$$

(veendu selles!), millest saame (universaalse) tugeva mõjususe.

VC-dimensioon. Kõigepealt universaalne (sõlmfunktsioonist sõltumatu) alumine tõke VC dimensioonile: suvalise sõlmfunktsiooni korral ([1], Thm 30.5)

$$V_k \ge 2\lfloor \frac{k}{2} \rfloor d = O(kd).$$

Lävefunktsioon

Kuigi toodud alumine tõke on sõlmfunktsioonist sõltumatu, pole universaalset ülemist tõket nii kerge leida. Ühekihiliste võrkude klassi $\mathcal{C}^{(k)}$ VC dimensioon on väikseim lävefunktsiooni (6.4.10) korral. Sellise sõlmfunktsiooni korral kehtib hinnang ([1], Thm 30.6):

$$V_k \le 2(kd + 2k + 1)\log_2(e(kd + 2k + 1)) = O(kd \ln(kd)). \tag{6.4.14}$$

Seega lävefunktsiooni korral on ülaltoodud alumine tõke täpne $\ln(kd)$ faktorini. Seega, kui $k(n)$ kasvab nii aeglaselt, et $k(n)(\ln k(n))\frac{\ln n}{n} \to 0$, näiteks $k(n) \sim \sqrt{n}$ siis (6.4.13) koondub. Tegelikult pole raske näha, et (6.4.13) koondub ka siis, kui $k(n)\frac{\ln n}{n} \to 0$, s.t. $k(n) = o(\frac{n}{\ln n})$.

Seega sõlmfunktsiooni (6.4.10) ja mitte liiga kiiresti kasvava k korral:

- kui varjatud neuronite arv kasvab kontrollitult, näiteks $k(n)\frac{\ln n}{n} \to 0$, ja seda närvivõrku õigesti treenida (näiteks ERM), siis saame (universaalse tugevalt) mõjusa reegli;
- kuigi lähendamisviga koondub nulliks, puudub kontroll selle koondumise kiiruse üle, see võib olla aeglane;

- hindamisviga saab kontrollida universaalselt (jaotusest sõltumatult): (6.4.11) ja (6.4.12) annavad universaalse (jaotusest sõltumatu) hinnangu hindamisveale ja kleskmisele hindamisveale.
- Võrratusest (2.6.15) saame iga ühe varjatud kihiga värvivõrgu (mitte ilmtingimata ERM-printsiibil leitud klassifitseerija) korral riski hinnangu ja tõenäosusega $1-\delta$

$$R(g_n) \le R_n(g_n) + 4\sqrt{\frac{8(V_k \ln(n+1) - \ln\delta + \ln 8)}{n}}. \tag{6.4.15}$$

Teised sõlmfunktsioonid

Saab näidata, et teiste sõlmfunktsioonide korral pole VC dimensioon kunagi väiksem lävefunktsiooni VC dimensioonist. Paraku võib see olla oluliselt suurem. Enimkasutatavate sõlmfunktsioonide – logistiline, arctan, gaussi – korral on $V_k < \infty$, kuid enamikel juhtudel puudub V_k korralik hinnang, mistõttu puuduvad ka hinnanguid (6.4.11), (6.4.12) ja (6.4.15).
Leidub aga ka sigmoide, mille VC-dimensioon on lõpmatu. Sellistega olgem ettevaatlikud! Samas võivad nad siiski olla mõjusad.

6.4.4 L_1-kauguse minimiseerimine

Empiirilise riski minimiseerimine on raske. Seetõttu minimiseeritakse tihti teisi (empiirilisi) riskifunktsioone. Nagu ka lineaarste klassifikaatorite korral, on selleks enamasti

$$J_n^p(\psi) := \frac{1}{n}\sum_{i=1}^{n} |y_i - \psi(x_i)|^p, \tag{6.4.16}$$

kus ψ, nagu ikka, on (6.4.4):

$$\psi(x) = \sum_{i=1}^{k} w_i \sigma(\psi_i(x)) + w_0.$$

Funktsiooni (6.4.16) minimiseeriva funktsiooni ψ_n abil defineeritakse klassifitseerija (6.4.7). Tuntuim algoritm funktsiooni J_n^p minimiseerimiseks on nn. **back-propagation** algoritm.

Mõjususest. Funktsioon J_n^p on empiiriline versioon funktsioonist

$$J^p(\psi) := E|Y - \psi(X)|^p. \tag{6.4.17}$$

Nägime, et lineaarsete klassifitseerijate korral see lähenemisviis ei olnud alati õigustatud – teatud jaotuste korral võib funktsiooni (6.4.17) minimiseeriva lineaarse funktsiooni abil defineeritud klassifitseerija risk olla palju suurem parima lineaarse klassifitseerija riskist.

Selgub, et ühekihilise närvivõrgu puhul on olukord parem – juhul kui $p = 1$ ja neuronite arv kasvab sobiva kiirusega, moodustavad funktsiooni J_n^p minimiseerivate võrkude $\{\psi_n\}$ põhjal defineeritud (plug-in) klassifitseerijad (tugevalt) mõjusa reegli iga sõlmfunktsiooni korral

Teoreem 6.16 *([1], Thm 30.9) Olgu σ suvaline sõlmfunktsioon. Olgu $\psi_n \in \mathcal{C}^{(k)}$ ühekihiline närvivõrk, mis minimiseerib J_n^1 üle klassi $\mathcal{C}^{(k)}$, lisatingimusel, et*

$$\sum_{i=1}^{k} |w_i| \leq \beta_n.$$

Kui $k(n) \to \infty$, $\beta_n \to \infty$ ja

$$\lim_{n\to\infty} k(n)\beta_n^2 \ln(k(n)\beta_n) = 0,$$

siis reegel

$$g_n = \begin{cases} 0 & \text{if } \psi(x) \leq 0.5, \\ 1 & \text{otherwise}, \end{cases}$$

on universaalselt mõjus.

Teatud lisatingimustel kehtib ka universaalne tugev mõjusus.

Loomulikult on mõjusus pelgalt teoreetiline omadus – praktikas on varjatud neuronite arv alati lõplik.

Kirjandus: Närvivõrkudest loe [1], Ch 30.

Peatükk 7

Unsupervised learning: Hidden Markov models

So far, we have been considering the supervised learning – also called *learning with teacher* – where the training data are given with the labels, i.e. to every feature vector x_i, the corresponding label (or output on the regression) y_i is known. Now, we consider the case of **unsupervised learning**, where the training sample $x_1, \ldots, x_n$ consists of feature vectors without the labels. To do some meaningful predictions without any further information about the data is difficult, although several general methods like cluster analysis exists. Often the lack of labels is compensated by assuming or knowing the probabilistic (generic) model of the data. Often the model is known up to the parametrization, and the parameters are estimated from the data. Sometimes a part (typically a small fraction) of the training data is known with labels and the rest of the data consist of feature vectors, solely. This case is referred to as **semi-supervised learning**. In semi-supervised learning, typically, the supervised part of the data (with labels) is used for fitting the model, the rest of the data (without labels) are then analyzed with using the fitted model.

7.1 Definition of Hidden Markov model

We shall consider a simple yet general model for classification in the case of unsupervised learning. Let us start with the basic definitions.

Underlying Markov chain – the regime. As previously, let $\mathcal{Y} = \{0, \ldots, k-1\}$ be the set of classes; in this chapter the classes will be referred to as the **states**, and the set $\mathcal{Y}$ will be now called as the **state space.** Let now $Y = \{Y_t\}_{t \in \mathbb{N}}$ be a **homogeneous Markov chain** that takes values in $\mathcal{Y}$. As usually, we shall denote by $\pi_j = P(Y_1 = j)$, $j \in \mathcal{Y}$ the probability that the initial state is j; the vector π is thus the

distribution of Y_1. We shall denote via

$$\mathbb{P} = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0(k-1)} \\ p_{10} & p_{11} & \cdots & p_{1(k-1)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{(k-1)0} & p_{(k-1)1} & \cdots & p_{(k-1)(k-1)} \end{pmatrix}$$

the transition matrix of Y, hence, for any $i_1, \ldots, i_{t-2} \in \mathcal{Y}$

$$p_{ij} = \mathbf{P}(Y_t = j | Y_{t-1} = i) = \mathbf{P}(Y_t = j | Y_{t-1} = i, Y_{t-2} = i_{t-2}, \ldots, Y_1 = i_1),$$

where the last equality follows from the Markov property. The chain Y is often called as **regime** and it is uniquely determined by the parameters π and $\mathbb{P}$. In practice, the regime Y is often defined to be as simple as possible, hence also stationary, irreducible and aperiodic, but we do not need any of these assumptions in this chapter.

Hidden Markov process – the observations. The second component of our model is the random (feature) process $X = \{X_t\}_{t \in \mathbb{N}}$ with X_t taking values in $\mathcal{X} \subseteq \mathbb{R}^d$. We assume that the pair of random processes (Y, X) is such that:

1) given $\{Y_t\}$, the random variables $\{X_t\}$ are conditionally independent,

2) the distribution of X_t depends on $\{Y_t\}$ only through Y_t.

When **1)** and **2)** hold, then the process X is called a **hidden Markov process** and the pair (Y, X) is called a **hidden Markov model (HMM).** The hidden Markov process $X = \{X_1, X_2, \ldots\}$ models the observations so that the random features $X_1, \ldots, X_n$ are the first n vectors of it. The corresponding outcomes of $Y_1, \ldots, Y_n$ are not observed, the Markov chain is *hidden.*

Note that, in general, $X_1, \ldots, X_n$ are neither independent nor identically distributed any more, hence our model allows to drop both assumptions made so far.

Emission distributions. The assumption **2)** states that for any t, the distribution of X_t depends on the regime only through the value of Y_t. Hence the distribution of X_t is independent of t and of the outcomes Y_s, where $s \neq t$. Therefore, to every state $j \in \mathcal{Y}$ corresponds a probability measure P_j such that for every Borel set A and for every $t \geq 1$, it holds

$$P_j(A) = \mathbf{P}(X_t \in A | Y_t = j).$$

The probability measures P_j, $j \in \mathcal{Y}$ are called **emission distributions.** Without loss of generality, we will assume that the emission distributions P_j all have densities f_j with respect to a common reference measure dx. As usually, the Lebesgue'i measure and counting measure (both denoted by dx) are of particular interest.

Finite alphabet HMP. When the emission distributions P_j all have finite (countable) support, then X is referred to as **finite (countable)-alphabet HMP**. In this case, the set $\mathcal{X}$ can be taken as the union of all supports, hence finite (countable). It will be referred to as the *alphabet*. Then, for every state, the density $f_j(x)$ is just the probability to emit the observation or letter x from the state j.

Mixture model: a special case. In a special case, when the columns of the transition matrix $\mathbb{P}$ are all equal to π, i.e. $p_{ij} = \pi$, for all $i, j \in \mathcal{Y}$, then $Y_1, Y_2, \ldots$ are i.i.d. random variables with distribution π. Then also $X_1, X_2, \ldots$ are i.i.d. random variables with density function

$$f(x) = \sum_{j=0}^{k-1} \pi_j f_i(x).$$

This model is sometimes called **mixture model.** Hence HMM also covers the case studied in Chapter 1.

7.2 Forward-backward recursions

7.2.1 Some notations and preliminaries.

Given a set $\mathcal{A}$, integers m and n, $m < n$, and a sequence $a_1, a_2, \ldots \in \mathcal{A}^\infty$, we write a_m^n for the subsequence $(a_m, \ldots, a_n)$. When $m = 1$, it will be often suppressed.

Thus, $x^n := (x_1, \ldots, x_n)$ and $y^n := (y_1, \ldots, y_n)$ stand for the fixed observed and unobserved *realizations* of $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$, respectively. Any sequence $y^n \in \mathcal{Y}^n$ is called a **path**.

Let $1 \le r \le m \le n$. From the definition of HMM, it follows that the conditional density of observing x_r^m given the realization y^n is the product of emission densities. Hence, we define

$$p(x_r^m | y^n) := \prod_{t=r}^{m} f_{y_t}(x_t).$$

In the following, let for any y_u^v, where $1 \le u \le v$

$$p(y_u^v) := \mathbf{P}(Y_u = y_u, Y_{u+1} = y_{u+1}, \ldots, Y_v = y_v).$$

Hence, we shall denote the joint probability density of (x_r^m, y^n) by

$$p(x_r^m, y^n) := p(x_r^m | y^n) p(y^n) = \prod_{t=r}^{m} f_{y_t}(x_t) \mathbf{P}(Y^n = y^n) = \prod_{t=r}^{m} f_{y_t}(x_t) \pi_{y_1} \prod_{t=2}^{n} p_{y_{t-1}, y_t}.$$

The unconditional density of a piece of observations x_u^v is now

$$p(x_r^m) := \sum_{y^n \in \mathcal{Y}^n} p(x_r^m, y^n).$$

The joint probability density of (x_r^m, y_u^v), where $1 \le u \le v \le n$ is thus

$$p(x_r^m, y_u^v) := \sum_{s^n \in \mathcal{Y}^n : s_u^v = y_u^v} p(x_r^m, s^n).$$

With joint density $p(x_r^m, y_u^v)$ and unconditional $p(y_u^v)$ and $p(x_r^m)$ we can define the conditional probabilities

$$\mathbf{P}(Y_u = y_u, \ldots, Y_v = y_v | X_r = x_r, \ldots, X_m = x_n) =: p(y_u^v | x_r^m) := \frac{p(x_r^m, y_u^v)}{p(x_r^m)}$$

and conditional densities

$$p(x_r^m | Y_u = y_u, \ldots, Y_v = y_v) =: p(x_r^m | y_u^v) := \frac{p(x_r^m, y_u^v)}{p(y_u^v)}.$$

We shall more closely consider the conditional probabilities $p(y_t|x^n)$. Therefore, they will have a special notation: for any $t \ge 1$ and $j \in \mathcal{Y}$, let

$$p_t(j|x^n) := \mathbf{P}(Y_t = j | X^n = x^n).$$

The probabilities $p_t(\cdot|x^n)$ are called:

- **smoothing** (posterior) probabilities, when $t < n$;
- **filtering** (posterior) probabilities, when $t = n$;
- **prediction** (posterior) probabilities, when $t > n$.

We shall also define

$$p_t(j) := \mathbf{P}(Y_t = j).$$

For stationary chain, clearly $p_t(j) = \pi_j$ for any t, but in general not.

In a special case of finite-alphabet HMP, the joint and conditional densities are just

$$p(x_r^m, y_u^v) = \mathbf{P}(X_r = x_r, \ldots, X_m = x_n; Y_u = y_u, \ldots, Y_v = y_v) = \mathbf{P}(X_r^m = x_r^m; Y_u^v = y_u^v);$$
$$p(x_r^m | y_u^v) = \mathbf{P}(X_r = x_r, \ldots, X_m = x_n | Y_u = y_u, \ldots, Y_v = y_v) = \mathbf{P}(X_r^m = x_r^m | Y_u^v = y_u^v).$$

7.2.2 Forward and backward recursions

Let us for every $j \in \mathcal{Y}$ and $x^n \in \mathcal{X}^n$ define **forward** and **backward** variables

$$\alpha(j, x^t) := p_t(j|x^t)p(x^t), \quad \beta(x_{t+1}^n | j) := \begin{cases} 1, & \text{if } t = n \\ p(x_{t+1}^n | Y_t = j), & \text{if } t < n \end{cases}$$

Clearly with $y_t = j$

$$\alpha(j, x^t) = p(x^t, y_t) = \sum_{y^t : y_t = j} p(x^t, y^t), \tag{7.2.1}$$

$$\beta(x_{t+1}^n | j) = \sum_{y_{t+1}^n} p(x_{t+1}^n, y_{t+1}^n | Y_t = j) = \sum_{y_{t+1}^n} p(x_{t+1}^n, y_{t+1}^n | y_t). \tag{7.2.2}$$

For any $y^n \in \mathcal{Y}^n$ and $x^n \in \mathcal{X}^n$ the following **factorization** holds

$$\begin{aligned} p(x^n, y^n) =& p(x^n|y^n)p(y^n) = p(x^t|y^n)p(x_{t+1}^n|y^n)p(y^n) = p(x^t|y^t)p(x_{t+1}^n|y_t^n)p(y^n) \\ =& p(x^t|y^t)p(x_{t+1}^n|y_t^n)p(y^t)p(y_{t+1}^n|y_t) = p(x^t, y^t)p(x_{t+1}^n|y_t^n)p(y_{t+1}^n|y_t) \\ =& p(x^t, y^t)p(x_{t+1}^n, y_{t+1}^n|y_t). \end{aligned}$$

The second equality follows from the assumption 1) of HMM (conditional independence of x^n), the third inequlity follows from the assumption 2) and the fourth equality follows from the Markov property. The last equality follows from

$$p(x_{t+1}^n|y_t^n)p(y_{t+1}^n|y_t) = \frac{p(x_{t+1}^n, y_t^n)}{p(y_t^n)} \frac{p(y_t^n)}{p(y_t)} = \frac{p(x_{t+1}^n, y_t^n)}{p(y_t)} = p(x_{t+1}^n, y_{t+1}^n|y_t).$$

Summing over all paths y^n passing the state j at t, using (7.2.1) and (7.2.2), from the factorization $p(x^n, y^n) = p(x^t, y^t)p(x_{t+1}^n, y_{t+1}^n|y_t)$ we obtain with $y_t = j$

$$p(x^n, y_t) = \sum_{y^n : y_t = j} p(x^n, y^n) = \Big(\sum_{y^t : y_t = j} p(x^t, y^t) \Big) \Big(\sum_{y_{t+1}^n} p(x_{t+1}^n, y_{t+1}^n | y_t) \Big) = \alpha(j, x^t)\beta(x_{t+1}^n|j). \tag{7.2.3}$$

From (7.2.3), it follows that α and β-variables can be used for finding $p_t(j|x^n)$:

$$p(x^n) = \sum_{y_t \in \mathcal{Y}} p(x^n, y_t) = \sum_{j \in \mathcal{Y}} \alpha(j, x^t)\beta(x_{t+1}^n|j), \tag{7.2.4}$$

$$p_t(j|x^n) = \frac{\alpha(j, x^t)\beta(x_{t+1}^n|j)}{\sum_{j \in \mathcal{Y}} \alpha(j, x^t)\beta(x_{t+1}^n|j)}, \tag{7.2.5}$$

In a special case of finite-alphabet HMP the variables are

$$\alpha(j, x^t) = \mathbf{P}(X^t = x^t; Y_t = j), \quad \beta(x_{t+1}^n|j) = \mathbf{P}(X_{t+1}^n = x_{t+1}^n | Y_t = j).$$

and the equation (7.2.3) is, thus,

$$\mathbf{P}(X^n = x^n; Y_t = j) = \mathbf{P}(X^t = x^t; Y_t = j)\mathbf{P}(X_{t+1}^n = x_{t+1}^n | Y_t = j).$$

Standard recursions. From the factorization

$$p(x^n, y^n) = p(x^t, y^t)p(x_{t+1}^n, y_{t+1}^n|y_t) \tag{7.2.6}$$

with $n = t + 1$, we get

$$p(x^{t+1}, y^{t+1}) = p(x^t, y^t)p(x_{t+1}, y_{t+1}|y_t). \tag{7.2.7}$$

Summing over y^{t-1}, we get

$$p(x^{t+1}, y_t^{t+1}) = p(x^t, y_t)p(x_{t+1}, y_{t+1}|y_t)$$

and summing over y_t, we obtain

$$p(x^{t+1}, y_{t+1}) = \sum_{y_t} p(x^t, y_t)p(x_{t+1}, y_{t+1}|y_t).$$

Finally, since

$$p(x_{t+1}, y_{t+1}|y_t) = p(y_{t+1}|y_t)f_{y_{t+1}}(x_{t+1}) = p_{y_t y_{t+1}} f_{y_{t+1}}(x_{t+1})$$

we have obtained a **forward recursion** for calculation α-variables:

$$\alpha(j, x^{t+1}) = \sum_{i=0}^{k-1} \alpha(i, x^t)p_{ij}f_j(x_{t+1}).$$

Similarly, one can show the **backward recursion** for calculation β-variables (see [30]):

$$\beta(x_t^n|j) = \sum_{i=0}^{k-1} \beta(x_{t+1}^n|i)p_{ji}f_i(x_t).$$

In a special case of finite-alphabet HMP the forward and backward recursions are

$$\mathbf{P}(Y_{t+1} = j; X^{t+1} = x^{t+1}) = \sum_i \mathbf{P}(Y_t = i; X^t = x^t)\mathbf{P}(Y_{t+1} = j|Y_t = i)\mathbf{P}(X_{t+1} = x_{t+1}|Y_{t+1} = j)$$

$$\mathbf{P}(X_t^t = x_t^n|Y_{t-1} = j) = \sum_i \mathbf{P}(X_{t+1}^n = x_{t+1}^n|Y_t = i)\mathbf{P}(Y_t = i|Y_{t-1} = j)\mathbf{P}(X_t = x_t|Y_t = i).$$

Derin's recursion. The above-described (standard) backward and forward recursions are not numerically stable, hence they are not very practical. Therefore, their normalized versions as well as several alternative recursions for calculating $p_t(j|x^n)$ in simular fashion are used. For example, the so-called **Derin's recursion** for finding $p_t(j|x^n)$ backward is as follows:

$$p_t(j|x^n) = p_t(j|x^t) \sum_{i=0}^{k-1} \frac{p_{ji}p_{t+1}(i|x^n)}{p_{t+1}(i|x^t)}.$$

For proof, see [30]. Here the smoothing probabilities $p_t(\cdot|x^t)$ and prediction probabilities $p_{t+1}(\cdot|x^t)$ can be found by forward recursion as follows:

$$p_1(j|x_1) = \frac{\pi_j f_j(x_1)}{\sum_i \pi_i f_i(x_1)},\ p_{t+1}(j|x^t) = \sum_i p_{ij} p_t(i|x^t),\ p_{t+1}(j|x^{t+1}) = \frac{p_{t+1}(j|x^t) f_j(x_{t+1})}{\sum_i p_{t+1}(i|x^t) f_i(x_{t+1})} \tag{7.2.8}$$

Exercise: Prove (7.2.8).

7.2.3 Conditional chain

The following proposition shows an important property: Y is an conditioonally inhomogeneous Markov chain given X.

Lause 7.1 *Given the observations x^n, the regime has Markov property:*

$$\mathbf{P}(Y_{t+1} = j|Y_t = i, Y^{t-1} = y^{t-1}; X^n = x^n) = \mathbf{P}(Y_{t+1} = j|Y_t = i; X^n = x^n). \tag{7.2.9}$$

Moreover, the conditional transition probablities

$$\mathbf{P}(Y_{t+1} = j|Y_t = i; X^n = x^n)$$

depend on the observations x_{t+1}^n, only.

Tõestus. It suffices to show that for any $y^{t+1} \in \mathcal{Y}^{t+1}$, the following equality holds

$$p(y_{t+1}|y^t, x^n) = p(y_{t+1}|y_t, x_{t+1}^n). \tag{7.2.10}$$

Indeed, (7.2.10) shows that the conditional transition probability $p(y_{t+1}|y^t, x^n)$ does not depend on y^{t-1}, hence Markov property (7.2.9), holds; the equality (7.2.10) also shows that given y_t the conditional distribution of y_{t+1}^n is independent of x^t. To show (7.2.10), recall the factorization (7.2.6):

$$p(x^n, y^n) = p(x^t, y^t) p(x_{t+1}^n, y_{t+1}^n|y_t).$$

Summing over y_{t+2}^n and y_{t+1}^n, we obtain

$$p(x^n, y^{t+1}) = p(x^t, y^t) p(x_{t+1}^n, y_{t+1}|y_t), \quad p(x^n, y^t) = p(x^t, y^t) p(x_{t+1}^n|y_t).$$

Thus,

$$p(y_{t+1}|y^t, x^n) = \frac{p(x^n, y^{t+1})}{p(x^n, y^t)} = \frac{p(x^t, y^t) p(x_{t+1}^n, y_{t+1}|y_t)}{p(x^t, y^t) p(x_{t+1}^n|y_t)} = \frac{p(x_{t+1}^n, y_{t+1}|y_t)}{p(x_{t+1}^n|y_t)} = p(y_{t+1}|y_t, x_{t+1}^n).$$

■

7.3 Segmentation

Let $x^n = x_1, \ldots, x_n$ be the given observations. We do not know the corresponding regime (the Markov chain is hidden), instead we assume that our model (HMM) is exactly known. The **problem of segmentation (problem of decoding)** is to estimate or prognose the hidden state sequence $y_1, \ldots, y_n$. This can be regarded as a classification problem, where input $x^n \in \mathcal{X}^n$, the set of outputs (classes) is $\mathcal{Y}^n$ and the classifier is a function

$$g = (g_1, \ldots, g_n) : \mathcal{X}^n \to \mathcal{Y}^n. \tag{7.3.1}$$

However, since the input and the output (the set of classes) both depend on n, the segmentation is the problem in its own rights.

7.3.1 Decision theory for HMM's

What is the best classifier (7.3.1)? To answer that, let us apply some ideas of Bayesian decision theory also for HMM's case. Recall the Bayesian decision theory – the Bayes classifier g^* is the one that for every feature vector x minimizes the conditional risk at x, i.e.

$$g^*(x) = \arg\min_{j \in \mathcal{Y}} R(j|x).$$

Also recall that the conditional risk was obtained via loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, where $L(y, j)$ is the loss of misclassifying the true class y as j. With loss function, for any $j \in \mathcal{Y}$,

$$R(j|x) = \sum_{y \in \mathcal{Y}} L(y, j) p(y|x).$$

Suppose now that we have defined the conditional risk function also for the segmentation problem. Thus, for every x^n, we have the conditional risk

$$R(\cdot|x^n) : \quad \mathcal{Y}^n \to [0, \infty]$$

so that $R(s^n|x^n)$ measures the goodness of path $s^n \in \mathcal{Y}^n$ given x^n. Just like in Bayesian decision theory, the risk of classifier $g = (g_1, \ldots, g_n)$ is then the expectation of conditional risk

$$R(g) := ER(g(X^n)|X^n)$$

and the classifier with minimal risk (over all possible classifiers) is called **Bayes classifier** that can be obtained by minimizing the conditional risk:

$$g^*(x^n) := \arg\min_{s^n \in \mathcal{Y}^n} R(s^n|x^n).$$

Loss function. The conditional risk $R(\cdot|x^n)$ depends on the task and (just like in Bayesian desicion theory) it is often meaningful to define it via **loss function**

$$L : \mathcal{Y}^n \times \mathcal{Y}^n \to [0, \infty],$$

where $L(y^n, s^n)$ is loss, when the actual state sequence is y^n and the prognose is s^n. The conditional expectation

$$R(s^n|x^n) := E[L(Y^n, s^n)|X^n = x^n]$$

is the conditional risk of s^n. Then, for any classifier g, the risk $R(g)$ is just the expected loss

$$R(g) = EL(Y^n, g(X^n))$$

and the Bayes classifier is

$$g^*(x^n) = \arg\min_{s^n \in S^n} \sum_{y^n \in S^n} L(y^n, s^n)\mathbf{P}(Y^n = y^n|X^n = x^n).$$

7.3.2 Viterbi aligment

Let us start with symmetric loss-function:

$$L(y^n, s^n) = \begin{cases} 1, & \text{when } y^n \neq s^n; \\ 0, & \text{when } y^n = s^n. \end{cases} \tag{7.3.2}$$

Then the conditional risk (in this case denoted by R_∞) is

$$R_\infty(s^n|x^n) = 1 - \mathbf{P}(Y^n = s^n|X^n = x^n)$$

and the Bayes classifier – for symmetric loss denoted by v – maps every sequence of observations into sequence s^n with maximum likelihood.

$$v(x^n) := \arg\max_{s^n \in S^n} \mathbf{P}(Y^n = s^n|X^n = x^n).$$

The maximum-likelihood state sequence $v(x^n)$ is known as **Viterbi alignment**. It inherits its name from the **Viterbi algorithm** – a dynamic programming algorithm for finding $v(x^n)$. Partially due to the simplicity of Viterbi algorithm, Viterbi alignment is by far most popular classifier in segmentation.

Viterbi algorithm. Recall the problem: given observations x^n, we are trying to find (all) path(s) $v^n \in \mathcal{Y}^n$ that maximizes the joint density $p(x^n, y^n)$ over all $y^n \in \mathcal{Y}^n$. Since the size of the search space is k^n, the disrect optimization is imposible even for moderate n. The Viterbi algorithm allows to solve the problem with complexity $O(n)$.

Let us introduce some notation. Since x^n is fixed, we skip it from the notation. Let, for every $t = 1, \dots, n$ and state $j \in \mathcal{Y}$, the **(Viterbi) scores** be

$$\delta_t(j) = \max_{y^t, y_t = j} p(x^t, y^t).$$

Hence $\delta_t(j)$ is the maximum joint likelihood over all paths ending at the state j. The scores can be found recursively (in t). Indeed, let y^{t+1} be a path that ends with j, i.e $y_{t+1} = j$. Then, using the factorization again,

$$p(x^{t+1}, y^{t+1}) = p(x^t, y^t)p(x_{t+1}, y_{t+1}|y_t) = p(x^t, y^t)p_{y_t j} f_j(x_{t+1}).$$

Let $s^t \in \mathcal{Y}^t$ be the path that maximizes $p(x^t, y^t)p_{y_t j}$ over all paths. If it ends at the state i, i.e. $s_t = i$, then

$$p(x^t, s^t)p_{s_t j} = p(x^t, s^t)p_{ij}$$

so that s^t has to be the path that maximizes $p(x^t, y^t)$ over all paths y^t ending with i. This is **Bellman's optimality principle**. In other words, if $s_t = i$, then $p(x^t, s^t) = \delta_t(i)$. That holds for every state i, thus

$$\delta_{t+1}(j) = \max_i \Big(\max_{y^{t+1}: y_{t+1} = i} p(x^t, y^t)p_{ij} \Big) f_j(x^{t+1}) = \max_i \big(\delta_t(i)p_{ij}\big) f_j(x_{t+1}).$$

Hence, we have the following (Viterbi) recursion for finding $\delta_t(i)$ for every i and t:

$$\delta_1(j) = \pi_j f_j(x_1), \quad \delta_{t+1}(j) = \max_i \big(\delta_t(i)p_{ij}\big) f_j(x_{t+1}) \tag{7.3.3}$$

Viterbi algorithm is a standard dynaming programming algorithm: at each $t = 1, \dots, n$ the scores $\delta_t(j)$ are calculated using recursion (7.3.3). By that, the algorithm stores

$$i_t(j) := \arg\max_i \delta_t(i)p_{ij}, \quad t = 1, \dots, n-1.$$

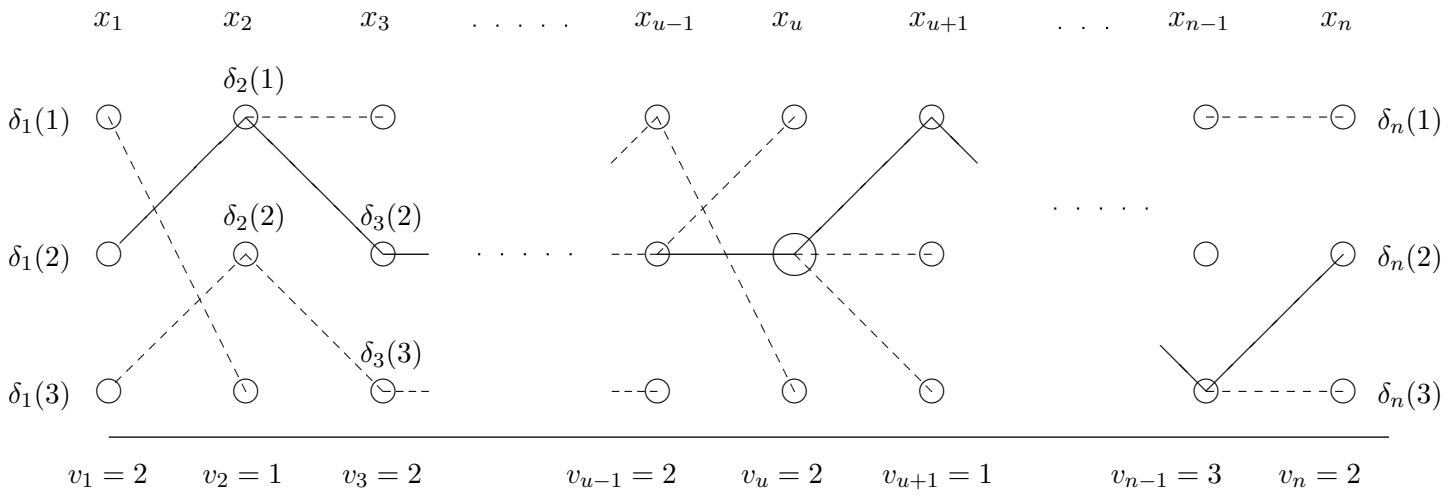
In case of ties, any choise will do. The solution can now found by *backtracking* as follows

$$v_n = \arg\max_j \delta_n(j), \quad v_t = i_t(v_{t+1}), \quad t = n-1, \dots, 1.$$

Viterbi algorithm

1. **Initialize:** For every $j \in \mathcal{Y}$, define $\delta_1(j) := \pi_j f_j(x_1)$;
2. **Do for** $t = 1, \ldots, n-1$**:**
   - Update
   $$\delta_{t+1}(j) = \max_i \big(\delta_t(i) p_{ij}\big) f_j(x_{t+1}); \tag{7.3.4}$$
   - Record
   $$i_t(j) := \arg\max_i \delta_t(i) p_{ij}$$
3. **Output:** Find Viterbi alignment v^n by backtracking:
   $$v_n := \arg\max_j \delta_n(j), \quad v_t = i_t(v_{t+1}), \quad t = n-1, \ldots, 1.$$

The following picture illustrates Viterbi algorithm in action. The solid lines indicates the output, the dashed lines indicate $i_t(j)$.



7.3.3 PMAP alignment

The symmetric loss (7.3.2) penalizes all differences alike: no matter whether y^n and s^n differ from one entry or all entries, the penalty is one. Often it is natural to penalize every different entry. This can be done by **pointwise loss-function**

$$l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty) \quad \text{where } l(s, s) = 0, \quad \forall s \in \mathcal{Y}. \tag{7.3.5}$$

Using l, we can define loss-function L as follows

$$L(y^n, s^n) := \sum_{t=1}^{n} l(y_t, s_t). \tag{7.3.6}$$

With (7.3.6), the conditional risk is

$$R(s^n|x^n) = E[L(Y^n, s^n)|X^n = x^n] = \sum_{t=1}^{n} E[l(Y_t, s_t)|X^n = x^n] \tag{7.3.7}$$

and minimizing $R(s^n|x^n)$ over s^n equals to minimizing $E[l(Y_t, s_t)|X^n = x^n]$ over s_t at every t. Hence, the Bayes classifier is obtained *pointwise*: $g^* = (g_1^*, \ldots, g_n^*)$, where

$$g_t^*(x^n) = \arg\min_{s \in \mathcal{Y}} E[l(Y_t, s)|X^n = x^n].$$

Counting errors. The most popular choice for l is, again, symmetric (pointwise) loss:

$$l(y, s) = \begin{cases} 0, & \text{if } y = s; \\ 1, & \text{if } y \neq s. \end{cases}.$$

Then the loss-function L counts the differences between y^n and s^n when compared pairwise: $L(y^n, s^n)$ is the number of pairwise errors or the difference in **Hamming distance**. Therefore, the conditional risk (in this case denoted by R_1) $R_1(s^n|x^n)$ measures expected number of misclassification errors of s^n given the observations are x^n. From (7.3.7), it follows that

$$R_1(s^n|x^n) := n - \sum_{t=1}^{n} \mathbf{P}(Y_t = s_t|X^n = x^n),$$

hence the Bayes classifier in this case – let us denote it by u – is the one that at each time t chooses the state with conditional probability:

$$u_t(x^n) = \arg\max_j \mathbf{P}(Y_t = j|X^n = x^n) = \arg\max_j p_t(j|x^n), \quad t = 1, \ldots, n.$$

We shall call u as **PMAP (pointwise maximum aposteriori) alignment**. Other names encountered: *marginal posterior mode*, *maximum posterior marginals*, *optimal symbol-by-symbol detection*, *symbol-by-symbol MAP estimation*, *MAP-state estimation*.

We know that the conditional probabilities $p_t(j|x^n)$ can be found with complexity $O(n)$ by several forward-backward recursions. This makes PMAP-classifier implementable and that is why PMAP alignment is the second most popular choice in practice.

Remark: Note that for mixture model these two alignments – Viterbi and PMAP alignment – coincide.

Logarithmic risks. Let us denote the logarithmic counterparts of R_∞ and R_1 risks:

$$\bar{R}_\infty(s^n|x^n) := -\ln \mathbf{P}(Y^n = s^n|x^n)$$
$$\bar{R}_1(s^n|x^n) := -\sum_{t=1}^{n} \ln \mathbf{P}(Y_t = s_t|x^n) = -\sum_{t=1}^{n} \ln p_t(s_t|x^n).$$

Clearly Viterbi alignment $v(x^n)$ minimizes $\bar{R}_\infty(\cdot|x^n)$ and PMAP alignment $u(x^n)$ minimizes $\bar{R}_1(\cdot|x^n)$.

7.3.4 Between PMAP and Viterbi

If the aim is to minimize the number of errors, then one should use PMAP-alignment. Unfortunately, it can be with very low or zero conditional likelihood, i.e. it might be that $p(u^n|x^n) = 0$. We call paths that have zero conditional likelihood **inadmissible**. This drawback is probably the main reason, why Viterbi alignment, although it might make in average more errors is preferred over PMAP alignment. In the following, we consider some options how to adjust PMAP alignment so that it still results an output with possible small number of expected misclassification errors, but at the same time remains admissible.

Restricted R_1-risk. The simplest solution is **restricted R_1-risk**:

$$\min_{y^n: p(y^n|x^n)>0} R_1(y^n|x^n) \quad \Leftrightarrow \quad \max_{y^n: p(y^n|x^n)>0} \sum_{t=1}^{n} p_t(y_t|x^n). \tag{7.3.8}$$

This problem can be solved by dynamic programming algorithm similar to the one of Viterbi algorithm.

Algorithm for restricted optimization:

1. **Initialize:**
   - Using forward-backward recursions, compute $p_t(j|x^n)$ for every $1 \le t \le n$ and $j \in \mathcal{Y}$;
   - For every $j \in \mathcal{Y}$, set
   $$\delta_1(j) := p_1(j|x^n);$$

2. **Do for** $t = 1, \ldots, n-1$**:**

   - For every $j \in \mathcal{Y}$, update

$$\delta_{t+1}(j) = \big(\max_i \delta_t(i) r_{ij} + p_{t+1}(j|x^n)\big) r_j^{t+1}, \tag{7.3.9}$$

     where

$$r_{ij} = \mathbb{I}_{\{p_{ij}>0\}}, \quad r_j^t = \mathbb{I}_{\{p_t(j|x^n)>0\}}.$$

   - Record

$$i_t(j) := \arg\max_i \delta_t(i) r_{ij}, \quad t = 1, \ldots, n-1.$$

3. **Output:** Find the optimal alignment s^n by backtracking:

$$s_n := \arg\max_j \delta_n(j), \quad s_t = i_t(s_{t+1}), \quad t = n-1, \ldots, 1.$$

Note that without the multipliers r_{ij} and r_j^t, the algorithm would indeed result PMAP-alignment, because, for every t and j, $\delta_{t+1}(j) = \max_i \delta_t(i) + p_{t+1}(j|x^n)$ so that

$$\max_j \delta_{t+1}(j) = \max_i \delta_t(i) + \max_j p_{t+1}(j|x^n).$$

If $p_{t+1}(j|x^n) > 0$, then $\arg\max_i \delta_t(i) r_{ij} > 0$, since otherwise there would no admissible path passing j at $t+1$ with positive likelihood. This contradicts $p_{t+1}(j|x^n) > 0$.

Finally note that the recursion (7.3.9) is equivalent to

$$\begin{aligned} \delta_1(j) &:= p_1(j|x^n), \\ \delta_{t+1}(j) &:= \max_i \left(\delta_t(i) + \log r_{ij}\right) + p_{t+1}(j|x^n) + \log r_j^{t+1}. \end{aligned} \tag{7.3.10}$$

Restricted $\bar{R}_1$-risk. In the presence of restrictions, (7.3.8) is not necessarily the solution of the following problem:

$$\min_{s^n : p(s^n|x^n)>0} \bar{R}_1(s^n|x^n) \quad \Leftrightarrow \quad \max_{s^n : p(s^n|x^n)>0} \sum_{t=1}^n \ln p_t(s_t|x^n). \tag{7.3.11}$$

The solution of (7.3.11) is sometimes called as the **posterior Viterbi decoding (PVD)** and it can be found the similar algorithm, where the recursion (7.3.9) is replaced by the following recursion

$$\begin{aligned} \delta_1(j) &:= p_1(j|x^n), \\ \delta_{t+1}(j) &:= \max_i \delta_t(i) r_{ij} \times p_{t+1}(j|x^n). \end{aligned} \tag{7.3.12}$$

Recursion (7.3.12) is clearly equivalent to

$$\begin{aligned} \delta_1(j) &:= \log p_1(j|x^n), \\ \delta_{t+1}(j) &:= \max_i \big(\delta_t(i) + \log r_{ij}\big) + \log p_{t+1}(j|x^n). \end{aligned} \tag{7.3.13}$$

Towards increasing the probability: $\bar{R}_1$-risk. Although admissible minimizers of R_1 and $\bar{R}_1$ risk are by definition of positive probability, this probability might still be very small. Indeed, in the above recursions, the weight r_{ij} is 1 even when p_{ij} is very small. Hence, in recursion (7.3.13) , we replace r_{ij} by the true transition (initial) probability p_{ij} and add $\log \pi_j$ to $\log p_1(j|x^n)$. Thus the modified recursion (7.3.13) is now

$$\begin{aligned} \delta_1(j) &:= \log p_1(j|x^n) + \log \pi_j \\ \delta_{t+1}(j) &:= \max_i \big(\delta_t(i) + \log p_{ij}\big) + \log p_{t+1}(j|x^n). \end{aligned} \tag{7.3.14}$$

Obviously, similar changes might be done in recursion (7.3.12). It is not difficult to see that with the recursion (7.3.14) the algorithm for restricted optimization solves the following problem

$$\max_{s^n} \Big[\sum_{t=1}^{n} \log p_t(s_t|x^n) + \log p(s^n) \Big] \quad \Leftrightarrow \quad \min_{s^n} \Big[\bar{R}_1(s^n|x^n) + h(s^n) \Big], \tag{7.3.15}$$

where the penalty term

$$h(s^n) \;\; = -\log p(s^n) \;\; =: \bar{R}_\infty(s^n) \tag{7.3.16}$$

is the prior log-likelihood risk which does not depend on the data.

More general problem. The recursion (7.3.14) immediately generalize as follows:

$$\begin{aligned} \delta_1(j) &:= \log p_1(j|x^n) + C \log \pi_j, \\ \delta_{t+1}(j) &:= \max_i \big(\delta_t(i) + C \log p_{ij}\big) + \log p_{t+1}(j|x^n), \end{aligned} \tag{7.3.17}$$

solving the following optimization problem

$$\min_{s^n} \Big[\bar{R}_1(s^n|x^n) + C h(s^n) \Big], \tag{7.3.18}$$

where $C > 0$ is a regularization or trade-off constant and $h(s^n) = \bar{R}_\infty(s^n)$. Then, PVD, i.e. the problem solved by the original recursions (7.3.12) and (7.3.13), can be recovered by taking C sufficiently small. Alternatively, the PVD problem can also be formally written in the form (7.3.18) with $C = \infty$ and $h(s^n)$ given, for example, by $\mathbb{I}_{\{p(s^n)=0\}}$.

Towards increasing the probability: R_1-risk. What if the actual probabilities p_{ij} (π_j) were also used in the optimal accuracy/PMAP decoding, i.e. optimization (7.3.9)-(7.3.10)? It appears more sensible to replace the indicators r_{ij} with p_{ij} (and adding π_j) in (7.3.10). The new recursion is now

$$\begin{aligned} \delta_1(j) &:= p_1(j|x^n) + \log \pi_j, \\ \delta_{t+1}(j) &:= \max_i \left(\delta_t(i) + \log p_{ij}\right) + p_{t+1}(j|x^n) + \log r_j^{t+1}. \end{aligned} \tag{7.3.19}$$

This solves the following problem:

$$\max_{s^n : p(s^n|x^n) > 0} \Big[\sum p_t(s_t|x^n) + \log p(s^n) \Big] \quad \Leftrightarrow \quad \min_{s^n : p(s^n|x^n) > 0} \Big[R_1(s^n|x^n) + \bar{R}_\infty(s^n) \Big]. \tag{7.3.20}$$

7.3.5 Combined risks

Motivated by the previous section, we consider the following general problem

$$\min_{s^n}\Big[C_1\bar{R}_1(s^n|x^n)+C_2\bar{R}_\infty(s^n|x^n)+C_3\bar{R}_1(s^n)+C_4\bar{R}_\infty(s^n)\Big], \tag{7.3.21}$$

where $C_i\geq 0$, $i=1,2,3,4$, $\sum_{i=1}^4 C_i>0$ and (recall)

$$\bar{R}_1(s^n|x^n)=-\sum_{t=1}^n\log p_t(s_t|x^n),\quad \bar{R}_\infty(s^n|x^n)=-\log p(s^n|x^n),$$

$$\bar{R}_1(s^n):=-\sum_{t=1}^n\log p_t(s_t),\quad \bar{R}_\infty(s^n)=-\log p(s^n)=-[\log\pi_{s_1}+\sum_{t=1}^{n-1}\log p_{s_ts_{t+1}}].$$

Some important special cases:

- the combination $C_1>0, C_2=C_3=C_4=0$ yields the PMAP case;
- the combination $C_2>0, C_1=C_3=C_4=0$ corresponds to the Viterbi decoding;
- the combination $C_3>0, C_1=C_2=C_4=0$ maximizes

  $$-\bar{R}_1(s^n)=\sum_{t=1}^n\log p_t(s_t)=\sum_{t=1}^n\log\mathbf{P}(Y_t=s_t),$$

  sometimes called *marginal prior mode* decoding;
- the combination $C_4>0, C_1=C_2=C_3=0$ maximizes $p(s^n)=\mathbf{P}(Y^n=s^n)$, sometimes called *maximum a priori* decoding;
- the combination case $C_1>0, C_4>0, C_2=C_3=0$ subsumes (7.3.18):

  $$\min_{s^n}\Big[C_1\bar{R}_1(s^n|x^n)-C_4\log p(s^n)\Big];$$

- the combination $C_1=C_3=0$ is the problem

  $$\min_{s^n}\Big[\bar{R}_\infty(s^n|x^n)+C\bar{R}_\infty(s^n)\Big]$$

  and its solution is a generalization of the Viterbi decoding that allows one to suppress ($C>0$) contribution of the data;
- the combination $C_1>0, C_2>0, C_3=C_4=0$ is the problem

  $$\min_{s^n}\Big[C_1\bar{R}_1(s^n|x^n)+C_2\bar{R}_\infty(s^n|x^n)\Big],$$

  and when $0<C_2\ll C_1$, it equals to minimizing $\bar{R}_1(s^n|x^n)$ under the condition

  $$\bar{R}_\infty(s^n|x^n)<\infty\quad\Leftrightarrow\quad p(s^n|x^n)>0.$$

  Thus the problem is the same as (7.3.11) and the solution of this problem is PVD-alignment.

Remark. It is important to note that with $C_2 > 0$ every solution of (7.3.21) is admissible. No less important, and perhaps a bit less obvious, is that $C_1, C_4 > 0$ also guarantees admissibility of the solutions.

Dynamic programming algorithm for solving (7.3.21) With suitable recursion, the algorithm for restricted optimization can be used to solve 7.3.21). To state the recursion, let

$$g_t(j) := C_1 \log p_t(j|x^n) + C_2 \log f_j(x_t) + C_3 \log p_t(j).$$

Note that the function g_t depends on the entire data x^n and they involve $p_t(j|x^n)$ as well as $p_t(j)$. The general recursion is the following

$$\begin{aligned} \delta_1(j) &:= C_1 \log p_1(j|x^n) + (C_2 + C_3 + C_4) \log \pi_j + C_2 \log f_j(x_1), \\ \delta_{t+1}(j) &:= \max_i \left(\delta_t(i) + (C_2 + C_4) \log p_{ij}\right) + g_{t+1}(j). \end{aligned} \tag{7.3.22}$$

Some important special cases (check):

- the combination $C_1 > 0, C_2 = C_3 = C_4 = 0$ yields the PMAP alignment ;
- the combination $C_2 > 0, C_1 = C_3 = C_4 = 0$ gives

  $$\begin{aligned} \delta_1(j) &:= C_2 \log \left(\pi_j f_j(x_1)\right), \\ \delta_{t+1}(j) &:= \max_i \left(\delta_t(i) + C_2 \log p_{ij}\right) + C_2 \log f_j(x_{t+1}) = \max_i \left(\delta_t(i) + C_2 \log \left(p_{ij} f_j(x_{t+1})\right)\right) \end{aligned}$$

  and that equals to Viterbi recursion (7.3.4);
- the combination $C_3 > 0, C_1 = C_2 = C_4 = 0$ gives the recursion

  $$\begin{aligned} \delta_1(j) &:= C_3 \log \pi_j, \\ \delta_{t+1}(j) &:= \max_i \delta_t(i) + C_3 \log p_{t+1}(j) \end{aligned}$$

  that, indeed, maximizes $\sum_{t=1}^n \log p_t(s_t)$;
- the combination $C_4 > 0, C_1 = C_2 = C_3 = 0$ gives the recursion

  $$\begin{aligned} \delta_1(j) &:= C_4 \log \pi_j, \\ \delta_{t+1}(j) &:= \max_i \left(\delta_t(i) + C_4 \log p_{ij}\right) \end{aligned}$$

  that, indeed, maximizes $\log p(s^n)$;
- the combination case $C_1 > 0, C_4 > 0, C_2 = C_3 = 0$ gives the recursion

  $$\begin{aligned} \delta_1(j) &:= C_1 \log p_1(j|x^n) + C_4 \log \pi_j, \\ \delta_{t+1}(j) &:= \max_i \left(\delta_t(i) + C_4 \log p_{ij}\right) + C_1 \log p_{t+1}(j|x^n) \end{aligned} \tag{7.3.23}$$

that for $C_1 = 1$ and $C_4 = C$ is exactly the same as (7.3.17). Clearly (7.3.23) solves the problem

$$\max_{s^n}\Big[C_1\sum_t \log p_t(s_t|x^n) + C_4\log p(s^n)\Big] = \min_{s^n}\Big[C_1\bar{R}_1(s^n|x^n) + C_4\bar{R}_\infty(s^n)\Big];$$

- the combination $C_2 > 0, C_4 > 0, C_1 = C_3 = 0$ gives the recursion

$$\begin{aligned}\delta_1(j) &:= C_2\log\big(\pi_j f_j(x_1)\big) + C_4\log\pi_j,\\ \delta_{t+1}(j) &:= \max_i\Big(\delta_t(i) + C_2\big(\log p_{ij}\log f_j(x_{t+1})\big) + C_4\log p_{ij}\Big),\end{aligned}$$

that solves

$$\max_{s^n}\Big[C_2\log p(s^n|x^n) + C_4\log p(s^n)\Big] = \min_{s^n}\Big[C_2\bar{R}_\infty(s^n|x^n) + C_4\bar{R}_\infty(s^n)\Big],;$$

- the combination $C_1 > 0, C_2 > 0, C_3 = C_4 = 0$ gives the recursion

$$\begin{aligned}\delta_1(j) &:= C_1\log p_1(j|x^n) + C_2\log\big(\pi_j f_j(x_1)\big),\\ \delta_{t+1}(j) &:= \max_i\Big(\delta_t(i) + C_2\log\big(p_{ij}f_j(x_{t+1})\big)\Big) + C_1\log p_{t+1}(j|x^n)\end{aligned} \quad (7.3.24)$$

that solves

$$\max_{s^n}\Big[C_1\sum_t \log p_t(s_t|x^n) + C_2\log p(s^n|x^n)\Big] = \min_{s^n}\Big[C_1\bar{R}_1(s^n|x^n) + C_2\bar{R}_\infty(s^n|x^n)\Big].$$

7.3.6 k-block alignments

Rabiner's k blocks. In his celebrated tutorial [28], L. Rabiner proposes instead of maximize the expected number of correctly decoded pairs or triples of (adjacent) states. Hence, with k being the length of the overlapping block ($k = 2, 3, \ldots$), he proposed to maximize the sum

$$p(s^k|x^n) + p(s_2^{k+1}|x^n) + p(s_2^{k+2}|x^n) + \cdots + p(s_{n-k+1}^n|x^n). \quad (7.3.25)$$

With $k = 1$, (7.3.25) is the sum $\sum_t p_t(s_t|x^n)$, hence in case $k = 1$, the maximizer of (7.3.25) is PMAP-alignment.

Formally, maximizing (7.3.25) equals to minimizing the conditional risk

$$R_k(s^n|x^n) := E\big[L_k(Y^n, s^n)|X^n = x^n\big], \quad (7.3.26)$$

where L_k is the following loss function:

$$L_k(y^n, s^n) := \mathbb{I}_{\{s^k \neq y^k\}} + \mathbb{I}_{\{s_2^{k+1} \neq y_2^{k+1}\}} + \mathbb{I}_{\{s_3^{k+2} \neq y_3^{k+1}\}} + \cdots + \mathbb{I}_{\{s_{n-k+1}^n \neq y_{n-k+1}^n\}}.$$

It is natural to think that minimizers of R_k-risk – **Rabiner's k-block alignment** – "move" towards Viterbi paths "monotonically" as k increases to n. Indeed, when $k = n$, then (7.3.25) is $p(s^n|x^n)$, hence the maximizer of it is Viterbi alignment. However, as the Example below shows, minimizers of $R_k(s^n|x^n)$ are not guaranteed to be admissible for $k > 1$.

Admissible k-blocks. The above-mentioned drawback (possible non-admissibility) is easily overcome when the sum in (7.3.25) is replaced by the product. Or, equivalently, the probabilities $p(s_t^{t+k-1}|x^n)$ are replaced by $\log p(s_t^{t+k-1}|x^n)$ in the sum (7.3.25). Certainly, except the case $k = 1$, these problems are not equivalent, but with the product in place of the sum the k-block idea works well. Namely, the longer the block, the larger the resulting path probability and, more importantly, the solution is clearly guaranteed to be positive already for $k = 2$. Indeed, if $p(s^n) = 0$, then there must exist a pair (transition) s_t^{t+1} such that $p(s_t^{t+1}) = 0$. Then $p(s_t^{t+1}|x^n) = 0$ and if one multiplier equals to zero, so does the whole product. If $p(s^n) > 0$, but $p(s^n|x^n) = 0$, then there must exists at least one x_t so that $f_{s_t}(x_t) = 0$. That, in turn implies $p(s_t|x^n) = 0$ and, therefore $p(s_t^{t+1}|x^n) = 0$ for *any pair* (s_t, s_{t+1}) that begins with s_t.

By replacing the probabilities $p(s_t^{t+k-1}|x^n)$ by $\log p(s_t^{t+k-1}|x^n)$, we would get a natural candidate for logarithmic version of R_k-risk as follows

$$\log p(s^k|x^n) + \log p(s_2^{k+1}|x^n) + \log p(s_2^{k+2}|x^n) + \cdots + \log p(s_{n-k+1}^n|x^n).$$

However, to get a nice connection with above defined general family of alignments, we define the logarithmic counterpart of R_k slightly different. Namely, let

$$\begin{aligned} U_1^k &:= p(s_1)\cdots p(s_1^{k-2})p(s_1^{k-1}) \\ U_2^k &:= p(s_1^k)p(s_2^{k+1})\cdots p(s_{n-k}^{n-1})p(s_{n-k+1}^n) \\ U_3^k &:= p(s_{n-k+2}^n)p(s_{n-k+3}^n)\cdots p(s_n). \end{aligned}$$

and let

$$U_k(s^n) := U_1^k(s^n)\cdot U_2^k(s^n)\cdot U_3^k(s^n).$$

The logarithmic version of R_k-risk will be defined as

$$\bar{R}_k(s^n) := -\log U_k(s^n) = -\log U_1^k(s^n) - \log U_2^k(s^n) - \log U_3^k(s^n).$$

Hence, with $k = 3$,

$$\begin{aligned} \bar{R}_3(s^n) = -\big(&\log p(s_1|x^n) + \log p(s_1^2|x^n) + \\ &\log p(s_1^3|x^n) + \log p(s_2^4|x^n) + \cdots + \log p(s_{n-3}^{n-1}|x^n) + \log p(s_{n-2}^n|x^n) \\ &+ \log p(s_{n-1}^n|x^n) + \log p(s_n|x^n)\big). \end{aligned}$$

Clearly when k is small in comparison with n, the modification is minor. Also note that for $k = 1$, the newly introduced risk equals to $\bar{R}_1$, hence the notation is correct. But the next theorem shows that being so defined, the $\bar{R}_k$-risk has a nice interpretation. Let

$$v(k) = \arg\min_{s^n} R_k(s^n|x^n).$$

Teoreem 7.1 *For every $x^n \in \mathcal{X}^n$, for every $s^n \in \mathcal{Y}^n$ and for every $k = 2, \ldots, n$, it holds*

$$\bar{R}_k(s^n|x^n) = (k-1)\bar{R}_\infty(s^n|x^n) + \bar{R}_1(s^n|x^n).$$

Moreover, $v(k)$ is admissible and

$$\bar{R}_\infty(v(k)|x^n) \leq \bar{R}_\infty(v(k-1)|x^n), \quad \bar{R}_1(v(k)|x^n) \geq \bar{R}_1(v(k-1)|x^n).$$

For proof, see [33]. The main statement of the above-stated theorem is the first one saying that our k-block alignment belongs to our general family of alignments (7.3.21) with the constants $C_1 = 1$ and $C_2 = k - 1$. Hence, for any k, the optimal path $v(k)$ can be found via recursion (7.3.24).
We already know that $C_2 > 0$ guarantees the admissibility of the solution, so the second statement of the theorem, is an immediate consequence of the first one.
The last two statements guarantee that if k increases, then the likelihood of $v(k)$ increases and

$$\sum_t \log p_t(v_t(k)|x^n)$$

decreases. This is, indeed, what one expects from a k-block alignment.

Example. Consider the following four-state MC transition matrix

$$\frac{1}{8}\begin{pmatrix} 0 & 4 & 2 & 2 \\ 4 & 1 & 1 & 2 \\ 2 & 1 & 1 & 4 \\ 2 & 2 & 4 & 0 \end{pmatrix}$$

Suppose observations x_1, x_2, x_3, x_4 and the emission densities f_s $s = 1, 2, 3, 4$ are such that

$$f_s(x_1) = f_s(x_4) = \begin{cases} 1, & \text{if } s = 2; \\ 0, & \text{if } s \neq 2. \end{cases}, \quad f_s(x_3) = f_s(x_2) = \begin{cases} A > 1, & \text{if } s = 1; \\ 1, & \text{if } s \neq 1. \end{cases}$$

Hence every admissible path begins and ends with 2.

Exercise: Show that

- Viterbi alignments are $(2, 1, 2, 2), (2, 2, 1, 2), (2, 1, 4, 2), (2, 4, 1, 2)$;
- PMAP-alignment is $(2, 1, 1, 2)$ – inadmissible;
- Rabiner's 2-block alignment is $(2, 1, 1, 2)$ – inadmissible;
- $v(2)$ alignments are $(2, 1, 4, 2)$ and $(2, 4, 1, 2)$, both admissible.

References: About HMM's in general, read [29, 30, 28, 31], about theory of segmentation read [32, 33]

Kirjandus

[1] A probabilistic theory of pattern recognition
L. Devroye, L. Györfi, G. Lugosi.
Springer, 1996.

[2] Pattern classification and learning theory.
G. Lugosi
In: Principles of Nonparametric Learning Springer, Wien, New York, pp. 1–56, Springer 2002.
http://www.econ.upf.edu/ lugosi/surveys.html.

[3] Theory of classification: a survey of some recent advances
S. Boucheron, O. Bousquet, G. Lugosi
ESAIM: Probability and Statistics, 9:323–375, 2005.
http://www.econ.upf.edu/ lugosi/surveys.html.

[4] Introduction to statistical learning theory
S. Boucheron, O. Bousquet, G. Lugosi
http://www.econ.upf.edu/ lugosi/surveys.html.

[5] Statistical learning theory
V. Vapnik
Wiley, 1998

[6] Pattern classification (2nd edition)
R. Duda, P. Hart, D. Stork
Wiley, 2000.

[7] The elements of statistical learning
T. Hastie, R. Tibshirani, T. Friedman.
Springer, 2001.

[8] The elements of statistical learning (2nd Ed)
T. Hastie, R. Tibshirani, T. Friedman.
Springer, 2009.

[9] Statistical pattern recognition
A. Webb
Wiley, 2002

[10] Machine learning: A probabilistic perspective
K. P. Murphi
MIT, 2012

[11] Introduction to machine learning
E. Alpaydin
MIT, 2004.

[12] An introduction to support vector machines and other kernel-based learning methods
N. Cristianini, J. Shawe-Taylor.
Cambridge University Press, 2003.

[13] Support vector machines
I. Steinwart,A. Christiann.
Springer, 2008.

[14] Kernel methods for pattern analysis
J. Shawe-Taylor, N. Cristianini.
Cambridge University Press, 2004.

[15] Learning with kernels: support vector machines, regularization, optimization, and beyond
B. Schölkopf ; A. J. Smola.
MIT Press, 2002.

[16] Kernel Fisher discriminants
S. Mika
PhD 2002
edocs.tu-berlin.de/diss/2002/mika-sebastian.pdf

[17] LARS software for R and Splus
B.Efron ja T.Hastie
www-stat.stanford.edu/∼ hastie/Papers/LARS.

[18] Large margin classifiers: convex loss, low noise and convergence rates
P. Bartlett, M. Jordan, J. McAuliffe
In: Advances of Neural Information Processing Systems, 16, 2004

[19] Convexity, Classification, and Risk bounds
P. Bartlett, M. Jordan, J. McAuliffe
Journal of the American Statistical Association, 101(473):138-156, 2006
http://www.stat.berkeley.edu/ bartlett/papers/bjm-ccrb-05.pdf

[20] On the existence of weak lerners and applications to boosting.
S. Mannor, R. Meir
Machine Learning, 48(1-2):219-251, 2002.

[21] An Introduction to Boosting and Leveraging
R. Meir, G. Rätsch
Lecture Notes In Artificial Intelligence, 2003.
http://www.face-rec.org/algorithms/Boosting-Ensemble/8574x0tm63nvjbem.pdf

[22] Empirical margin distributions and bounding the generalization error of combined classifiers.
V. Kolchinskii, D. Panchenko
Ann, Statis., 30(1), 2002

[23] Rademacher and Gaussian Complexities: Risk Bounds and Structural Results
P. Bartlett, S. Mendelson
Journal of Machine Learning Research, 3:463-487, 2002
http://www.ai.mit.edu/projects/jmlr/papers/volume3/bartlett02a/bartlett02a.pdf

[24] AdaBoost is Consistent
P. Bartlett and M. Traskin
Journal of Machine Learning Research, 8:2347-2368, 2007
http://jmlr.csail.mit.edu/papers/volume8/bartlett07b/bartlett07b.pdf

[25] The Boosting Approach to Machine Learning: An Overview.
Robert E. Schapire
2002.
http://tjure.sfs.uni-tuebingen.de/files/Kursmaterialien/Kuebler/ML-ss05/schapire.pdf

[26] A Short Introduction to Boosting.
Y. Freund and R. E. Schapire
In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1401–1406, 1999.
http://www.yorku.ca/gisweb/eats4400/boost.pdf

[27] Boosting Algorithms: Regularization, Prediction and Model Fitting.
P. Bühlmann, and T. Hothorn
Statistical Science 22, no. 4: 477-505, 2007.
http://projecteuclid.org/euclid.ss/1207580163.

[28] A tutorial on Hidden Markov Models and selected applications in speech recognition
L. Rabiner
Proceedings of the IEEE 77 (2): 257Ű286, 1989
http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial

[29] Inference in hidden Markov models
O. Cappe, E. Moulines, T. Ryden
Springer, 2005

[30] Hidden Markov models for bioinformatics
T. koski
Computational Biology Series, 2.
Kluwer Academic, 2001

[31] Y. Ephraim, N. Merhav
Hidden Markov processes
Special issue on Shannon theory: perspective, trends, and applications.
IEEE Trans. Inform. Theory 48(6), 2002

[32] J. Lember, K. Kuljus, A. Koloydenko
Theory of segmentation
Hidden Markov models: theory and applications
InTech, 2011
http://www.intechopen.com/books/hidden-markov-models-theory-and-applications

[33] J. Lember, A. Koloydenko
Bridging Viterbi and Posterior Decoding: A Generalized Risk Approach to Hidden Path Inference Based on Hidden Markov Models
Journal of Machine Learning Research, 2014
arXiv:1007.3622v1, 2010