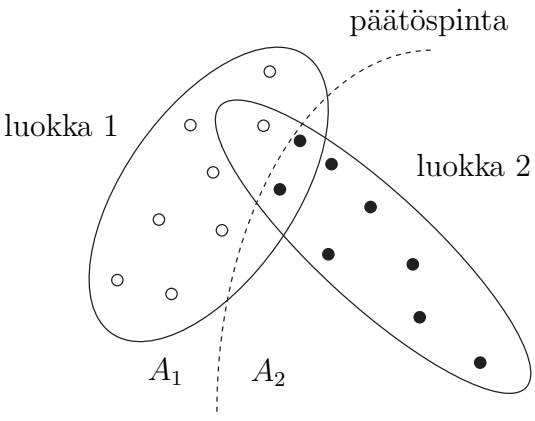
Tilastollinen hahmontunnistus

Lasse Holmström
Matemaattisten tieteiden laitos
Oulun yliopisto

Kevät 2010



Sisältö

Luku 1

Mitä on hahmontunnistus?

Eräs määritelmä:

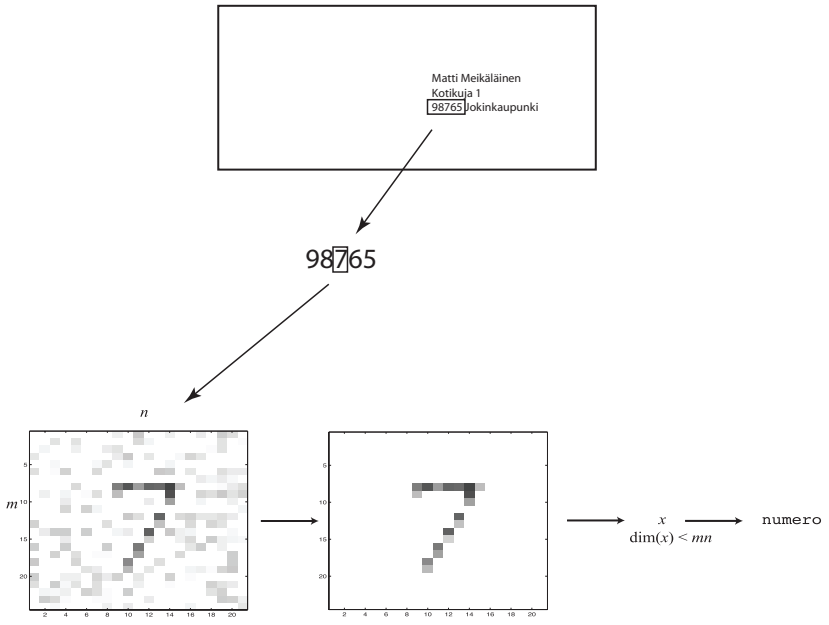
Hahmontunnistus on mittausten ja havaintojen tekemistä luonnollisista kohteista, näiden mittausten automaattista analyysiä ja kohteiden tunnistamista analyysin perusteella.

Esimerkki 1.1. Postinumeron automaattinen luku kirjekuoresta.

Tämä tehtävä voidaan ajatella suoritettavan seuraavien vaiheiden kautta (kuva 1.1):

- postinumero paikallistetaan digitaalisesta kuvasta
- yksittäiset numerot rajataan
- numerot tunnistetaan

Yksittäisen numeron tunnistaminen tapahtuu siten, että ensin sitä esittävästä digitaalisesta $m \times n$ kuvasta, ”hahmosta”, poistetaan häiriötä (”kohinaa”), ja sitten kuvassa oleva tieto tiivistetään ”piirrevektoriin” x, jonka di-



Kuva 1.1: Postinumeron automaattinen luku kirjekuoresta.

mensio d on tyypillisesti pienempi kuin alkuperäisen kuvan dimensio mn. Viime vaiheessa vektori x ”tunnistetaan” luokittelemalla se johonkin luokista $0, 1 \ldots, 9$. Oleellista on luonnollisesti pyrkiä mahdollisimman virheettömään luokitteluun.

Hahmovektorin $x = \left[x^{(1)}, x^{(2)}, \ldots, x^{(d)}\right]^T$ komponenttien $x^{(i)}$ tyyppi riippuu käytetystä järjestelmästä. Mustavalkoisessa kuvassa voidaan ottaa $x^{(i)} \in \{0, 1\}$ ja harmaasävyjä käytettäessä $x^{(i)} \in \{0, 1, \ldots, 255\}$, $i = 1, \ldots, d$. Luokitin voidaan ajatella kuvaukseksi $g : \mathbb{R}^d \to \{0, 1, \ldots, 9\}$, jossa

$$g(x) = \text{arvaus } x\text{:n luokaksi.}$$

$\|$

Eräitä tyypillisiä esimerkkejä hahmontunnistuksesta on esitetty taulukossa 1.1. Muista sovelluksista mainittakoon vielä DNA-sirut, joissa geenien aktiivisuuskuvio ennustaa esimerkiksi syövän tyypin, aivomagneettisten MEG-mittausten automaattinen tulkinta, lentokentällä käytettävät turvaportit ja

sovellus	x	**luokka**
konenäkö	kuva tai sen osa	esine kuvassa
puheentunnistus	puheen osan spektri	puhuttu foneemi
potilaan seuranta	EEG-käyrä	normaali/epänormaali

Taulukko 1.1: Esimerkkejä sovelluksista, joissa automaattista hahmontunnistusta hyödynnetään sekä niihin liittyvät hahmovektorit ja luokat.

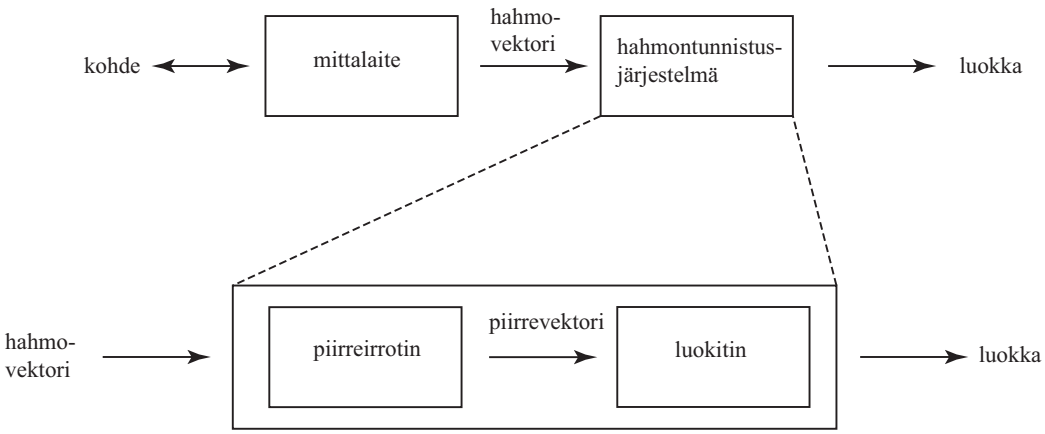
vaikkapa hiukkasfysiikan mittausten analyysi.

Hahmontunnistussovelluksen yleisrakenne sekä itse hahmontunnistujärjestelmän keskeiset osat on esitetty kuvassa 1.2. Siinä piirreirrotin pyrkii

- korostamaan luokittelun kannalta oleellista informaatiota hahmovektorissa
- vähentämään luokittelun kannalta epäoleellisen tiedon osuutta

Usein piirrevektorin dimensio on huomattavasti pienempi kuin hahmovektorin dimensio. Tämä auttaa luokittimen konstruktiossa ja nopeuttaa myös itse luokittelua mikä joissain sovelluksissa voi olla tärkeää. Toisin kuin luokitin, jotka usein hyödyntää jotain yleiskäyttöistä päätössääntöä, on piirreirrotusmenetelmä usein vahvasti käsillä olevaan sovellukseen sidottu. Tämän kurssin pääasiallisena kiinnostuksen kohteena ovat luokittimet. Taulukossa 1.2 on esitettty eräiden tällä kurssilla käytettävien termien englannikieliset vastineet.

Tilastollisessa hahmontunnistuksessa luokitellaan euklidisen avaruuden $\mathbb{R}^d$ vektoreita. Oma hahmontunnistuksen lajinsa on *strukturaalinen hahmontunnistus* (structural pattern recognition), jossa luokittelun kohteena ovat esimerkiksi graafit tai merkkijonot. Tätä aihepiiriä ei käsitellä tällä kurssilla.



Kuva 1.2: Hahmontunnistussovelluksen yleisrakenne sekä itse hahmontunnistusjärjestelmän rakenne.

hahmo	pattern
hahmontunnistus	pattern recognition
luokittelu	classification, discrimination (erityisesti tilastotieteessä)
piirre	feature
piirteiden irrotus	feature extraction
tilastollinen hahmontunnistus	statistical pattern recognition

Taulukko 1.2: Eräiden tällä kurssilla käytettävien termien englanninkieliset vastineet.

Luku 2

Matemaattisia apuneuvoja

2.1 Vektorit ja matriisit

Luokiteltava piirrevektori $x \in \mathbb{R}^d$ ajatellaan pystyvektorina

$$x = \begin{bmatrix} x^{(1)}, \ldots, x^{(d)} \end{bmatrix}^T = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} \in \mathbb{R}^d, \tag{2.1}$$

missä $x^{(i)}$ on x:n i:s komponentti ja T tarkoittaa transpoosia.

Vektorien $x = \begin{bmatrix} x^{(1)}, \ldots, x^{(d)} \end{bmatrix}^T$ ja $y = \begin{bmatrix} y^{(1)}, \ldots, y^{(d)} \end{bmatrix}^T$ *sisätulo* (skalaaritulo, pistetulo) on

$$x^T y = \begin{bmatrix} x^{(1)}, \ldots, x^{(d)} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(d)} \end{bmatrix} = \sum_{i=1}^{d} x^{(i)} y^{(i)}.$$

Vektorin x *normi* (pituus) määritellään tavalliseen tapaan kaavalla

$$\|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{d} \left(x^{(i)}\right)^2}.$$

Joukko $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ on *ortonormaali* (o.n.), jos

$$x_i^T x_j = \delta_{ij} \text{ kaikilla } i, j = 1, \ldots, n,$$

eli

$$\begin{cases} x_i^T x_j = 0, & \text{kun } i \neq j, \\ \|x_i\| = 1, & i = 1, \ldots, n. \end{cases}$$

Kun $x^T y = 0$, ovat x ja y ovat *ortogonaaliset* (kohtisuorat). Tätä merkitään $x \perp y$.

Käytämme $k \times d$ matriisille merkintää

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \\ a_{k1} & \cdots & a_{kd} \end{bmatrix} \in \mathbb{R}^{k \times d}.$$

*Diagonaalimatriisi*ssa muut kuin diagonaalialkiot ovat nollia,

$$A = \operatorname{diag}(a_1, \cdots, a_d) = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_d \end{bmatrix}.$$

Tästä erikoistapauksena on *yksikkömatriisi*,

$$I_d = \operatorname{diag}(1, \cdots, 1) = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}.$$

Matriisi $A \in \mathbb{R}^{d\times d}$ on *säännöllinen* (kääntyvä, ei-singulaarinen), jos on olemassa sellainen $B \in \mathbb{R}^{d\times d}$, että

$$AB = BA = I_d.$$

Tällöin merkitään $B = A^{-1}$. Helposti nähdään, että $(A_1A_2)^{-1} = A_2^{-1}A_1^{-1}$, kun A_1^{-1} ja A_2^{-1} ovat olemassa. Eräs ehto säännöllisyydelle on nollasta eroava determinantti, $\det(A) \neq 0$.

Matriisin $A \in \mathbb{R}^{k\times d}$ *transpoosi* on $B \in \mathbb{R}^{d\times k}$, jolle $b_{ij} = a_{ji}$ kaikilla i, j. Merkitsemme $B = A^T$. Tällöin pätee, että $(A_1A_2)^T = A_2^TA_1^T$.

Edellä *pystyvektori* (2.1) on myös $d \times 1$ matriisi. Vastaavasti matriisi $x^T = \left[x^{(1)}, \ldots, x^{(d)}\right]$ on *vaakavektori*.

Matriisi $A \in \mathbb{R}^{d\times d}$ on *symmetrinen*, jos $A^T = A$ ja *ortogonaalinen*, jos $A^T = A^{-1}$. Ortogonaalisen matriisin pystyrivit (ja myös vaakarivit) ovat ortonormaalit: jos

$$A = [a_1, \ldots, a_d], \quad a_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{dj} \end{bmatrix},$$

niin koska $A^TA = I_d$, on

$$\left[a_i^Ta_j\right] = \left[\delta_{ij}\right] \quad \text{eli} \quad a_i^Ta_j = \delta_{ij},\ i, j = 1, \ldots, d.$$

Kääntäen, jos matriisin $A \in \mathbb{R}^{d\times d}$ pystyrivit (vaakarivit) ovat ortonormaalit, on A ortogonaalinen.

Edelleen, jos A on ortogonaalinen, niin $\|Ax\| = \|x\|$ kaikilla x, sillä

$$Ax = \begin{bmatrix} a_{11} & \ldots & a_{1d} \\ \vdots & & \\ a_{d1} & \ldots & a_{dd} \end{bmatrix} \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^d a_{1j}x^{(j)} \\ \vdots \\ \sum_{j=1}^d a_{dj}x^{(j)} \end{bmatrix} = \sum_{j=1}^d x^{(j)}a_j,$$

joten

$$\begin{aligned}\|Ax\|^2 &= \left(\sum_{j=1}^{d} x^{(j)} a_j\right)^T \left(\sum_{k=1}^{d} x^{(k)} a_k\right) \\ &= \sum_{j,k=1}^{d} x^{(j)} x^{(k)} a_j^T a_k \\ &= \sum_{j=1}^{d} \left(x^{(j)}\right)^2 = \|x\|^2 .\end{aligned}$$

Matriisi $A \in \mathbb{R}^{d\times d}$ on *positiivisesti semidefiniitti*, jos $x^T Ax \geq 0$ kaikilla $x \in \mathbb{R}^d$ ja *positiivisesti definiitti*, jos $x^T Ax > 0$ kaikilla $x \neq 0$. Positiivisesti definiitti matriisi on säännöllinen, koska $Ax \neq 0$ aina kun $x \neq 0$, eli lineaarisena kuvauksena $A : \mathbb{R}^d \to \mathbb{R}^d$ A on injektio ja siis bijektio.

Luku $\lambda \in \mathbb{R}$ on matriisin $A \in \mathbb{R}^{d\times d}$ *ominaisarvo*, jos on olemassa sellainen $u \neq 0$, että $Au = \lambda u$. Tällöin u on A:n *ominaisvektori*.

2.2 Singulaariarvohajotelma ja pseudoinversio

Olkoon $A \in \mathbb{R}^{d\times d}$ symmetrinen. Lineaarialgebrasta tiedetään tällöin, että matriisilla A on reaaliset ominaisarvot $\lambda_1, \ldots, \lambda_d$ ja niitä vastaavat ortonormaalit ominaisvektorit $u_1, \ldots, u_d$,

$$\begin{aligned} Au_i &= \lambda_i u_i, \\ u_i^T u_j &= \delta_{ij}, \end{aligned} \tag{2.2}$$

$i, j = 1, \ldots, d$. On huomattava, että osa ominaisarvoista λ_i voi olla samoja; esimerkiksi yksikkömatriisin I_d kaikki d ominaisarvoa ovat arvoltaan 1.

Merkitään

$$U = [u_1, \ldots, u_d] \in \mathbb{R}^{d \times d},$$
$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d) = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

Silloin (2.2) voidaan kirjoittaa muodossa

$$AU = U\Lambda$$

josta

$$AUU^T = U\Lambda U^T.$$

Koska U on ortogonaalinen matriisi, $U^{-1} = U^T$, joten

$$A = U\Lambda U^T. \tag{2.3}$$

Tämä on symmetrisen matriisin *ominaisarvohajotelma*.

Olkoon sitten $A \in \mathbb{R}^{k \times d}$ mielivaltainen matriisi (ei siis välttämättä neliömatriisi). Silloin (HT)

- $A^T A \in \mathbb{R}^{d \times d}$ ja $AA^T \in \mathbb{R}^{k \times k}$ ovat symmetrisiä ja positiivisesti semidefiniittejä,
- $A^T A$:n ja AA^T:n ominaisarvot ovat ei-negatiivisia,
- matriiseilla $A^T A$ ja AA^T on samat positiiviset ominaisarvot (ja niillä samat kertaluvut).

Olkoot nyt $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d \geq 0$ matriisin $A^T A$ ominaisarvot ja $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_k \geq 0$ matriisin AA^T ominaisarvot. Olkoot edelleen $\mu_1 = \nu_1, \ldots, \mu_r = \nu_r$ näiden matriisien yhteiset positiiviset ominaisarvot ja $\lambda_i = \sqrt{\mu_i} = \sqrt{\nu_i}$, $i = 1, \ldots, r$. Tässä siis $r \leq \min(d, k)$.

Merkitään

$$\Lambda = \begin{bmatrix} \operatorname{diag}(\lambda_1, \ldots, \lambda_r) & 0_{r\times(d-r)} \\ 0_{(k-r)\times r} & 0_{(k-r)\times(d-r)} \end{bmatrix} \in \mathbb{R}^{k\times d}. \tag{2.4}$$

Tässä merkintä $0_{p\times q}$ tarkoittaa $p \times q$ nollamatriisia. Silloin matriisilla A on *singulaariarvohajotelma*

$$A = U\Lambda V^T, \tag{2.5}$$

missä $U \in \mathbb{R}^{k\times k}$ ja $V \in \mathbb{R}^{d\times d}$ ovat ortogonaalisia matriiseja, joiden pystyrivit koostuvat vastaavasti matriisien AA^T ja A^TA ortonormaaleista ominaisvektoreista,

$$\begin{aligned} AA^T u_i &= \nu_i u_i, \quad i = 1, \ldots, k, \\ A^T A v_i &= \mu_i v_i, \quad i = 1, \ldots, d. \end{aligned}$$

Luvut $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\min(k,d)} \geq 0$ ovat A:n *singulaariarvot*.

Esimerkki 2.1. Kun $k = 5$, $d = 3$, $r = 2$, on kaavan (2.4) diagonaalimatriisi Λ muotoa

$$\Lambda = \left[\begin{array}{cc|c} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

‖

Tarkastellaan sitten lineaarista $k \times d$ yhtälöryhmää

$$Ax = y, \quad x \in \mathbb{R}^d,\ y \in \mathbb{R}^k, \tag{2.6}$$

missä $A \in \mathbb{R}^{k\times d}$. Tässä yhtälössä on siis k yhtälöä ja d tuntematonta.

Esimerkki 2.2. Yhtälöryhmä

$$\begin{cases} 2x^{(1)} + 3x^{(2)} + 4x^{(3)} = 5 \\ 6x^{(1)} + 7x^{(2)} + 8x^{(3)} = 9 \end{cases}$$

voidaan kirjoittaa kaavan (2.6) tapaan matriisin ja vektoreiden avulla muodossa

$$\underbrace{\begin{bmatrix} 2 & 3 & 4 \\ 6 & 7 & 8 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 5 \\ 9 \end{bmatrix}}_{y}$$

eli $Ax = y$, missä $A \in \mathbb{R}^{2\times 3}$. ‖

Erotamme yhtälöryhmässä (2.6) kolme eri tapausta.

1. $k = d$ ja $\det(A) \neq 0$. Tällöin on olemassa yksikäsitteinen ratkaisu $x = A^{-1}y$.
2. $k = d$ ja $\det(A) = 0$ tai $k > d$ (ylimäärätty yhtälöryhmä). Tällöin ei yleensä ole ratkaisua.
3. $k < d$ (alimäärätty yhtälöryhmä). Tällöin on yleensä äärettömän monta ratkaisua. Voi kuitenkin myös olla, että ei ole yhtään ratkaisua.

Haluaisimme kuitenkin tilanteen, jossa aina löytyisi yksikäsitteinen ”ratkaisu”. Yhtälöryhmän (2.6) *yleistetty ratkaisu* $\tilde{x} \in \mathbb{R}^d$ määritellään ehdoilla

(i) $\min_{x\in\mathbb{R}^d} \|Ax - y\| = \|A\tilde{x} - y\|$,

(ii) jos $\|Ax - y\| = \|A\tilde{x} - y\|$, niin $\|x\| \geq \|\tilde{x}\|$.

Huomautus. Tapauksessa 1 yllä pätee selvästi, että $\tilde{x} = A^{-1}y$. Siksi vektoria $\widetilde{x}$ voi todella kutsua ”yleistetyksi” ratkaisuksi.

Olkoon $a_i^T = [a_{i1}, \ldots, a_{id}]$ matriisin A i:s vaakarivi ja $y = [y^{(1)}, \ldots, y^{(k)}]^T$. Silloin

$$Ax = \begin{bmatrix} a_1^T x \\ \vdots \\ a_k^T x \end{bmatrix}$$

ja

$$\|Ax - y\| = \sqrt{\sum_{i=1}^{k} \left(a_i x - y^{(i)}\right)^2}.$$

Siten kohdan (i) nojalla valinta $x = \tilde{x}$ minimoi lausekkeen

$$\sum_{i=1}^{k} \left(a_i x - y^{(i)}\right)^2,$$

eli $\tilde{x}$ on (2.6):n *pienimmän neliösumman ratkaisu*. Kohdan (ii) nojalla $\tilde{x}$ on itseasiassa tällaisista ratkaisuista *normiltaan pienin*.

Osoitetaan nyt, että on olemassa matriisi $A^+ \in \mathbb{R}^{d \times k}$ siten, että

$$\tilde{x} = A^+ y.$$

A^+ on nimeltään matriisin A (Mooren-Penrosen) *pseudoinverssi*.

Olkoon A ensin muotoa (2.4) oleva diagonaalimatriisi, $A = \Lambda$. Tällöin

$$\|Ax - y\|^2 = \|\Lambda x - y\|^2 = \sum_{i=1}^{r} \left(\lambda_i x^{(i)} - y^{(i)}\right)^2 + \sum_{i=r+1}^{k} \left(y^{(i)}\right)^2.$$

Nyt selvästi $\tilde{x} = [y^{(1)}/\lambda_1, \ldots, y^{(r)}/\lambda_r, 0, \ldots, 0]^T$, joten

$$\tilde{x} = \Lambda^+ y, \tag{2.7}$$

missä

$$\Lambda^+ = \begin{bmatrix} \operatorname{diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_r}\right) & 0_{r \times (k-r)} \\ 0_{(d-r) \times r} & 0_{(d-r) \times (k-r)} \end{bmatrix} \in \mathbb{R}^{d \times k}.$$

Yleisessä tapauksessa muodostetaan matriisille A ensin singulaariarvohajotelma,

$$A = U \Lambda V^T,$$

jolloin

$$\begin{aligned}\|Ax - y\| &= \left\|U\Lambda V^T x - y\right\| \\ &= \left\|U(\Lambda V^T x - U^T y)\right\| \\ &= \left\|\Lambda V^T x - U^T y\right\| \\ &= \|\Lambda z - w\|,\end{aligned}$$

missä $z = V^T x$, $w = U^T y$, toinen yhtäsuuruus seurasi ehdosta $U^{-1} = U^T$ ja kolmas matriisin U ortogonaalisuudesta. Edelleen, V on ortogonaalinen, joten V^T on ortogonaalinen ja siis $\|z\| = \|x\|$. Siten yhtälöryhmän

$$\Lambda z = w$$

yleistetty ratkaisu $\tilde{z}$ vastaa kaavan $\tilde{z} = V^T \tilde{x}$ välityksellä yhtälöryhmän (2.6) yleistettyä ratkaisua $\tilde{x}$. Mutta tuloksen (2.7) nojalla

$$\tilde{z} = \Lambda^+ w$$

joten

$$\tilde{x} = V\tilde{z} = V\Lambda^+ w = V\Lambda^+ U^T y.$$

Siten matriisin A pseudoinverssi voidaan määritellä kaavalla

$$A^+ = V\Lambda^+ U^T.$$

Helposti todetaan (HT):

(i) jos A on säännöllinen, niin $A^+ = A^{-1}$,

(ii) $(A^T)^+ = (A^+)^T$,

(iii) $A^+ = (A^T A)^+ A^T = A^T (AA^T)^+$.

Kohdista (i) ja (iii) saadaan käytännössä hyödylliset tulokset

(*iv*) jos $k \geq d$ ja $A^T A$ on säännöllinen, niin $A^+ = (A^T A)^{-1} A^T$,

(*v*) jos $k < d$ ja AA^T on säännöllinen, niin $A^+ = A^T (AA^T)^{-1}$.

Harjoitustehtävänä on osoittaa seuraavat tulokset:

1. jos $k > d$, ei AA^T voi olla säännöllinen,
2. jos $k < d$, ei $A^T A$ voi olla säännöllinen,
3. jos $k = d$, $A^T A$ on säännöllinen jos ja vain jos AA^T on säännöllinen.

2.3 Satunnaisvektorit

Hahmontunnistusjärjestelmässä piirrevektoria $x \in \mathbb{R}^d$ on luonnollista ajatella *satunnaissuureena*, koska

- x on peräisin satunnaisesta luokasta (esimerkiksi jokin luvuista $0, 1, \ldots, 9$),
- saman luokan piirrevektorit ovat erilaisia (esimerkiksi numerot kirjoitettu käsin),
- mittauksessa voi olla satunnaisvirhettä, ”kohinaa”.

Palautetaan siis ensin mieleen joitain todennäköisyyslaskennan perusasioita.

Todennäköisyysavaruus on kolmikko $(\Omega, \mathcal{F}, \mathbb{P})$, missä

- $\Omega \neq \emptyset$ on alkeistapahtumien joukko (perusjoukko),
- $\mathcal{F}$ on Ω:n osajoukkojen joukko (tapahtumat),

- $\mathbb{P} : \mathcal{F} \to [0, \infty[$ on todennäköisyys (mitta).

Lisäksi vaaditaan, että

- $\mathcal{F}$ on σ-algebra,
- $\mathbb{P}(\Omega) = 1$ ja jos $A_1, A_2, \ldots \in \mathcal{F}$ pistevieraita, niin

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \quad \text{(täysadditiivisuus)}.$$

Satunnaismuuttuja (sm) on kuvaus $X : \Omega \to \mathbb{R}$, jolle pätee

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}, \tag{2.8}$$

kun B on riittävän säännöllinen (ns. Borelin joukko). Siten näille B todennäköisyys

$$\mathbb{P}(X \in B) \equiv \mathbb{P}\big(X^{-1}(B)\big)$$

on määritelty.

Satunnaisvektori (sv) on vektoriarvoinen kuvaus $X : \Omega \to \mathbb{R}^d$, jolle ehto (2.8) pätee, kaikilla $B \subset \mathbb{R}^d$ (Borelin joukko). Voidaan osoittaa, että $X = \big[X^{(1)}, \ldots, X^{(d)}\big]^T : \Omega \to \mathbb{R}^d$ on satunnaisvektori jos ja vain jos $X^{(i)} : \Omega \to \mathbb{R}$ on satunnaismuuttuja kaikilla $i = 1, \ldots, d$. Satunnaismuuttuja on siis satunnaisvektorin (yksiulotteinen) erikoistapaus. Jatkossa merkitään satunnaisvektoreita yleensä *isoilla* kirjaimilla $(X, Y, \ldots)$.

Satunnaisvektorin saamien arvojen todennäköisyyttä kuvaa sen *jakauma*. Satunnaisvektorilla X on *diskreetti jakauma*, jos on olemassa sellainen numeroituva joukko $C \subset \mathbb{R}^d$, että $\mathbb{P}(X \in C) = 1$. Jos $C = \{c_1, c_2, \ldots\}$, X:n jakauman määräävät *pistetodennäköisyydet*

$$p_k = \mathbb{P}(X = c_k).$$

Esimerkki 2.3. Heitetään kahta noppaa. Olkoon $X =$ noppien silmälukujen summa. Kuvailemme koetta todennäköisyysavaruudella, jonka perusjoukko on $\Omega = \{(i,j) \mid i,j = 1,\ldots,6\}$. Tapahtumien kokoelmana $\mathcal{F}$ on kaikki Ω:n osajoukot. Olkoon $A \subset \Omega$ ja $n(A)$ joukon A:n alkioiden lukumäärä. Määritellään

$$\mathbb{P}(A) = \frac{n(A)}{n(\Omega)}.$$

Silloin $(\Omega, \mathcal{F}, \mathbb{P})$ on todennäköisyysavaruus. Kun $\omega = (i,j) \in \Omega$, on $X(\omega) = i+j$. Satunnaismuuttujan X arvojoukko on $C = X(\Omega) = \{2,3,\ldots,12\}$. Nyt $n(\Omega) = 36$ ja pistetodennäköisyydet saadaan kaavasta

$$p_k = \mathbb{P}(X = k) = \frac{6 - |k-7|}{36}, \quad k = 2,3,\ldots,12.$$

‖

Funktio $f : \mathbb{R}^d \to [0,\infty[$ on *tiheysfunktio* (tf), jos

$$\int_{\mathbb{R}^d} f(x)dx = 1.$$

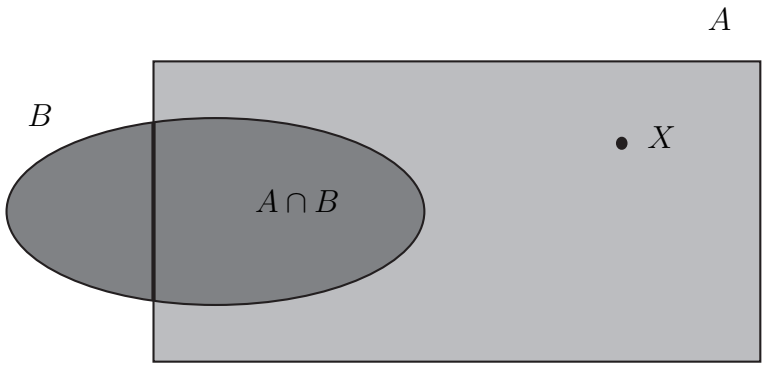
Tarkkaan ottaen lisäksi vaaditaan, että f on ns. Borelin funktio eli että $f^{-1}(B)$ on Borelin joukko kaikilla Borelin joukoilla $B \subset \mathbb{R}$. Kaikki tällä kurssilla eteen tuleva funktiot ovat Borelin funktiota – on itse asiassa melko vaikeaa keksiä esimerkkejä funktioista jotka eivät sitä ole.

Satunnaisvektorilla X on *jatkuva jakauma tiheysfunktiona* f_X, jos

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx$$

kaikilla $B \subset \mathbb{R}^d$ (Borel).

Esimerkki 2.4. (Tasainen jakauma avaruudessa $\mathbb{R}^2$)



Kuva 2.1: Joukossa A tasaiseen jakaumaan liittyvä todennäköisyys.

Olkoon $A \subset \mathbb{R}^2$ kiinteä ja $X = \left[X^{(1)}, X^{(2)}\right]^T$ joukosta A umpimähkään valittu piste. Silloin, jos $B \subset \mathbb{R}^2$ ja $m(\cdot)$ merkitsee joukon pinta-alaa, pätee

$$\mathbb{P}(X \in B) = \frac{m(A \cap B)}{m(A)}, \quad B \subset \mathbb{R}^2.$$

Oletamme tässä, että $m(A) > 0$. (Tarkkaan ottaen pitäisi puhua taas Borelin joukoista ja pinta-alan sijaan Lebesguen mitasta.) Tilannetta on havainnollistettu kuvassa 2.1. Olkoon nyt

$$f_X(x) = \begin{cases} \frac{1}{m(A)}, & x \in A, \\ 0, & \text{muulloin.} \end{cases}$$

Tällöin selvästi

$$\int_{\mathbb{R}^2} f_X(x)dx = \int_A \frac{1}{m(A)}dx = m(A) \cdot \frac{1}{m(A)} = 1,$$

joten f_X on tiheysfunktio. Edelleen, jos $B \subset \mathbb{R}^2$, on

$$\int_B f_X(x)dx = \int_{A \cap B} \frac{1}{m(A)}dx = \frac{m(A \cap B)}{m(A)} = \mathbb{P}(X \in B).$$

Siten f_X on X:n tiheysfunktio. Sanomme, että X:n jakauma on *tasainen* A:ssa. ‖

Satunnaisvektorin X jakaumaa voidaan myös luonnehtia erilaisilla *tunnusluvuilla*. Näitä ovat mm. odotusarvo, kovarianssi ja korrelaatio. Tarkastellaan seuraavassa vain jatkuvan satunnaisvektorin tapausta.

Palautetaan ensin mieleen jatkuvan satunnaismuuttujan $X : \Omega \to \mathbb{R}$ vastaavat tunnusluvut. Oletetaan, että kaikki alla olevat integraalit ovat olemassa.

Odotusarvo

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

kuvaa X:n keskimääräistä arvoa.

Varianssi

$$D^2(X) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)^2\big] = \int_{-\infty}^{\infty} \big(x - \mathbb{E}(X)\big)^2 f_X(x) dx$$

puolestaan kuvaa X:n arvojen keskittymistä $\mathbb{E}(X)$:n ympärille. Näitä käsitteitä on havainnollistettu kuvassa 2.2.

Olkoon sitten $X : \Omega \to \mathbb{R}^d$ jatkuva satunnaisvektori,

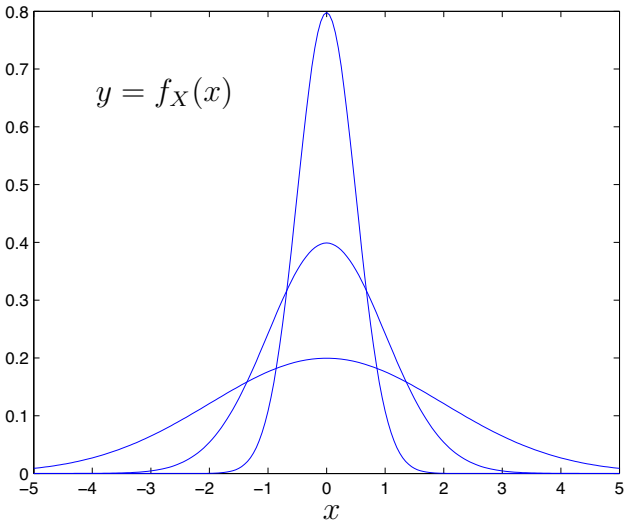
$$X = \big[X^{(1)}, \ldots, X^{(d)}\big]^T = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(d)} \end{bmatrix}.$$

X:n *odotusarvo* on vektori

$$\mathbb{E}(X) = \begin{bmatrix} \mathbb{E}(X^{(1)}) \\ \vdots \\ \mathbb{E}(X^{(d)}) \end{bmatrix} \in \mathbb{R}^d.$$

Vektorin $\mathbb{E}(X)$ komponentit ovat satunnaismuuttujien $X^{(i)}$ odotusarvoja,

$$\mathbb{E}(X^{(i)}) = \int_{\mathbb{R}^d} x^{(i)} f_X(x) dx = \int_{-\infty}^{\infty} x^{(i)} f_{X^{(i)}}(x^{(i)}) dx^{(i)},$$



Kuva 2.2: Kuvassa on kolmen satunnaismuttujan jakauman tiheysfunktiot. Kunkin satunnaismuuttujan odotusarvo $\mathbb{E}(X) = 0$ mutta varianssit eroavat, $D^2(X) = 0.5, 1$ tai 2, kapeimmasta jakaumasta leveimpään.

missä

$$f_{X^{(i)}}(x^{(i)}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x)\, dx^{(1)} \cdots dx^{(i-1)} dx^{(i+1)} \cdots dx^{(d)}$$

on $X^{(i)}$:n (reuna)tiheysfunktio. Merkitsemme odotusarvoa tavallisesti μ:llä tai μ_X:llä. Siis $\mu = \mathbb{E}(X)$, $\mu = [\mu^{(1)}, \ldots, \mu^{(d)}]^T$, $\mu^{(i)} = \mathbb{E}(X^{(i)})$, $i = 1, \ldots, d$.

X:n *kovarianssimatriisi* on

$$\mathrm{Cov}(X) = [\sigma_{ij}] \in \mathbb{R}^{d \times d}, \tag{2.9}$$

missä

$$\begin{aligned}\sigma_{ij} &= \mathbb{E}\big[(X^{(i)} - \mu^{(i)})(X^{(j)} - \mu^{(j)})\big] \\ &= \int_{\mathbb{R}^d} (x^{(i)} - \mu^{(i)})(x^{(j)} - \mu^{(j)}) f_X(x)\, dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x^{(i)} - \mu^{(i)})(x^{(j)} - \mu^{(j)}) f_{X^{(i)},X^{(j)}}(x^{(i)}, x^{(j)})\, dx^{(i)}\, dx^{(j)},\end{aligned}$$

missä $f_{X^{(i)},X^{(j)}}(x^{(i)}, x^{(j)})$ on parin $(X^{(i)}, X^{(j)})$ (reuna)tiheysfunktio, $i, j = 1, \ldots, d$.

Koska

$$\begin{aligned}(X-\mu)(X-\mu)^T &= \begin{bmatrix} X^{(1)} - \mu^{(1)} \\ \vdots \\ X^{(d)} - \mu^{(d)} \end{bmatrix} \big[X^{(1)} - \mu^{(1)}, \ldots, X^{(d)} - \mu^{(d)}\big] \\ &= \begin{bmatrix} \big(X^{(1)} - \mu^{(1)}\big)\big(X^{(1)} - \mu^{(1)}\big) & \cdots & \big(X^{(1)} - \mu^{(1)}\big)\big(X^{(d)} - \mu^{(d)}\big) \\ \vdots & & \vdots \\ \big(X^{(d)} - \mu^{(d)}\big)\big(X^{(1)} - \mu^{(1)}\big) & \cdots & \big(X^{(d)} - \mu^{(d)}\big)\big(X^{(d)} - \mu^{(d)}\big) \end{bmatrix},\end{aligned}$$

on itseasiassa

$$\mathrm{Cov}(X) = \mathbb{E}\big[\big(X - \mu\big)\big(X - \mu^T\big)\big].$$

Näemme myös, että kaavassa (2.9)

$$\begin{aligned}\sigma_{ii} &= \mathbb{E}\big[\big(X^{(i)} - \mu^{(i)}\big)^2\big] = D^2\big(X^{(i)}\big), \\ \sigma_{ij} &= \mathbb{E}\big[\big(X^{(i)} - \mu^{(i)}\big)\big(X^{(j)} - \mu^{(j)}\big)\big] = \mathrm{Cov}(X^{(i)}, X^{(j)}),\ i \neq j,\end{aligned}$$

missä $\mathrm{Cov}(X^{(i)}, X^{(j)})$ on satunnaismuuttujien $X^{(i)}$ ja $X^{(j)}$ on kovarianssi. Merkitsemme kovarianssimatriisia tavallisesti Σ:lla tai Σ_X:llä. Siis $\Sigma = \mathrm{Cov}(X) = [\sigma_{ij}]$.

Hahmontunnistuskirjallisuudessa *korrelaatiomatriisilla* tarkoitetaan usein matriisia

$$S = \mathbb{E}(XX^T) = [s_{ij}],$$
$$s_{ij} = \mathbb{E}\big(X^{(i)}X^{(j)}\big) = \int\limits_{\mathbb{R}^d} x^{(i)}x^{(j)} f_X(x)dx.$$

Sen sijaan tilastotieteessä "korrelaatiomatriisi" määritellään kaavalla

$$R = [r_{ij}],$$
$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \frac{\mathrm{Cov}(X^{(i)}, X^{(j)})}{\sqrt{D^2(X^{(i)})}\sqrt{D^2(X^{(j)})}}.$$

Suureet Σ, S ja R mittaavat eri tavoin X:n komponenttien $X^{(i)}$ välistä riippuvuutta. Kovarianssimatriisin Σ:n diagonaali puolestaan kuvaa satunnaismuuttujien $X^{(i)}$ keskittymistä odotusarvojensa $\mu^{(i)}$ ympärille.

Huomautus. Edellä oletettiin, että kaikki odotusarvot (integraalit) olivat olemassa. Oletamme aina jatkossakin, että kaikki satunnaisvektoreilla suoritetut operaatiot ovat hyvin määriteltyjä.

Käytännössä μ, S ja Σ joudutaan estimoimaan X:stä saadun *satunnaisotoksen* $X_1, \ldots, X_n \sim f_X$ avulla (X_i:t siis ovat riippumattomia ja samoin jakautuneita, tiheysfunktiona f_X). Tavallisia estimaattoreita ovat

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{(otoskeskiarvo)}$$
$$\widehat{S} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \qquad \text{(otoskorrelaatiomatriisi)}$$
$$\widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T \qquad \text{(otoskovarianssimatriisi)}$$

Tekijä $1/(n-1)$ yllä takaa *harhattomuuden*, $\mathbb{E}(\widehat{\Sigma}) = \Sigma$.

2.4 Multinormaalijakauma

Olkoon X jatkuva satunnaisvektori, $A \in \mathbb{R}^{d\times d}$ säännöllinen ja $b \in \mathbb{R}^d$. Määritellään uusi satunnaisvektori X:n affiinina muunnoksena

$$Y = AX + b.$$

Silloin

$$\begin{aligned}\mu_Y &= \mathbb{E}(Y) = \mathbb{E}(AX + b)\\ &= \mathbb{E}(AX) + \mathbb{E}(b)\\ &= A\,\mathbb{E}(X) + b = A\mu_X + b.\end{aligned}$$

Edelleen,

$$\begin{aligned}\Sigma_Y &= \mathbb{E}\big[(Y - \mu_Y)(Y - \mu_Y)^T\big]\\ &= \mathbb{E}\big[(AX + b - A\mu_X - b)(AX + b - A\mu_X - b)^T\big]\\ &= \mathbb{E}\big[A(X - \mu_X)\big(A(X - \mu_X)\big)^T\big]\\ &= \mathbb{E}\big[A(X - \mu_X)(X - \mu_X)^T A^T\big]\\ &= A\,\mathbb{E}\big[(X - \mu_X)(X - \mu_X)^T\big]A^T\\ &= A\,\Sigma_X A^T. \qquad (2.10)\end{aligned}$$

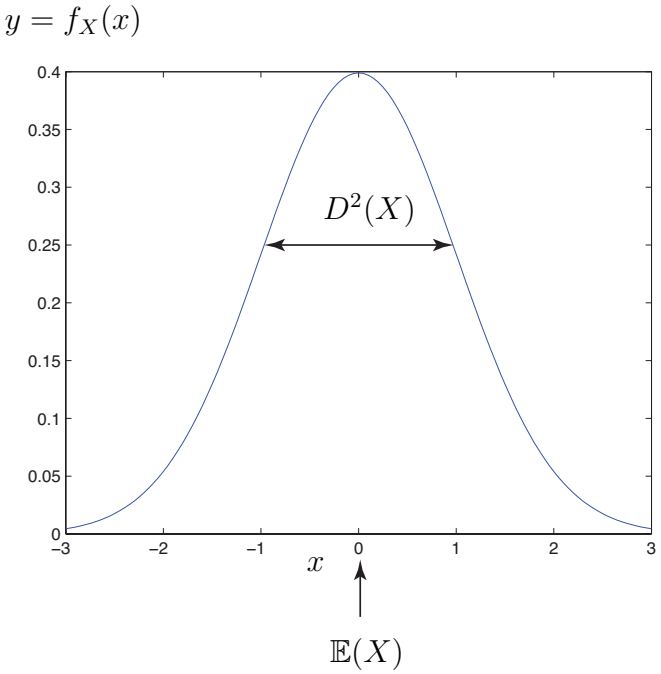
Kuten tunnettua, satunnaismuuttujalla X sanotaan olevan *standardi normaalijakauma*, jos sen jakaumalla on tiheysfunktio

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R}.$$

Tällöin merkitään $X \sim \mathrm{N}(0, 1)$ (kuva 2.3).

Sanomme, että satunnaisvektorilla $X = \big[X^{(1)}, \ldots, X^{(d)}\big]^T$ on *standardi multinormaalijkauma*, jos sen jakaumalla on tiheysfunktio

$$f_X(x) = \frac{1}{(2\pi)^{d/2}}\, e^{-\frac{1}{2}\|x\|^2}, \quad x \in \mathbb{R}^d.$$



Kuva 2.3: Standardinormaalijakauman tiheysfunktio.

Tällöin merkitsemme $X \sim \mathrm{N}(0, I_d)$. Kun $x = [x^{(1)}, \ldots, x^{(d)}]^T$, on siis

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}[(x^{(1)})^2 + \cdots + (x^{(d)})^2]} \\ &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^{(i)})^2}. \end{aligned} \tag{2.11}$$

Tästä voidaan laskea $X^{(i)}$:n tiheysfunktio,

$$\begin{aligned} f_{X^{(i)}}(x^{(i)}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x) dx^{(1)} \cdots dx^{(i-1)} dx^{(i+1)} \cdots dx^{(d)} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^{(i)})^2}, \end{aligned}$$

sillä

$$\int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^{(j)})^2} \, dx^{(j)} = 1$$

kaikilla $j \neq i$.

Siis $X^{(i)} \sim \mathrm{N}(0,1)$ ja erityisesti $\mu_X = [\mu_{X^{(1)}}, \ldots, \mu_{X^{(d)}}]^T = [0, \ldots, 0]^T \equiv 0$. Edelleen, kaavan (2.11) nojalla $X^{(1)}, \ldots, X^{(d)}$ ovat *riippumattomia*. Siten X:n kovarianssimatriisille $\Sigma = [\sigma_{ij}]$ saadaan

$$\begin{aligned}
\sigma_{ij} &= \mathrm{Cov}\big(X^{(i)}, X^{(j)}\big) \\
&= \mathbb{E}\big[\big(X^{(i)} - 0\big)\big(X^{(j)} - 0\big)\big] \\
&= \mathbb{E}\big(X^{(i)} X^{(j)}\big) = \mathbb{E}\big(X^{(i)}\big)\,\mathbb{E}\big(X^{(j)}\big) \\
&= 0 \cdot 0 = 0
\end{aligned}$$

kaikilla $i \neq j$ ja $\sigma_{ii} = D^2(X^{(i)}) = 1$ kaikilla i. Siten $\Sigma_X = I_d$.

Tehdään nyt satunnaisvektorille $X \sim \mathrm{N}(0, I_d)$ affiini muunnos

$$Y = AX + b,$$

missä A on säännöllinen. Edellä lasketun perusteella

$$\begin{aligned}
\mu_Y &= A0 + b = b, \\
\Sigma_Y &= AI_dA^T = AA^T.
\end{aligned} \tag{2.12}$$

Olkoon $g : \mathbb{R}^d \to \mathbb{R}^d$, $g(x) = Ax + b$. Silloin $Y = g(X)$. Johdetaan Y:n tiheysfunktio. Kun $B \subset \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{P}(Y \in B) &= \mathbb{P}\big(g(X) \in B\big) \\
&= \mathbb{P}\big(X \in g^{-1}(B)\big) \\
&= \int\limits_{g^{-1}(B)} f_X(x)dx.
\end{aligned}$$

Tehdään muuttujan vaihdos $y = g(x)$, jolloin $x = g^{-1}(y) = A^{-1}(y - b)$ ja

$$\int\limits_{g^{-1}(B)} f_X(x)dx = \int\limits_B f_X\big(A^{-1}(y-b)\big)|\det(A^{-1})|dy \int\limits_B = f_Y(y)dy,$$

missä

$$f_Y(y) = |\det(A^{-1})|\, f_X\big(A^{-1}(y-b)\big), \quad y \in \mathbb{R}^d,$$

joka on siis Y:n tiheysfunktio. Siis

$$f_Y(y) = \frac{|\det(A^{-1})|}{(2\pi)^{d/2}}\, e^{-\frac{1}{2}\|A^{-1}(y-b)\|^2}.$$

Tässä

$$\begin{aligned}
\big\|A^{-1}(y-b)\big\|^2 &= \big[A^{-1}(y-b)\big]^T A^{-1}(y-b) \\
&= (y-b)^T (A^{-1})^T A^{-1}(y-b) \\
&= (y-b)^T (AA^T)^{-1}(y-b) \\
&= (y-\mu_Y)^T\, \Sigma_Y^{-1}(y-\mu_Y),
\end{aligned}$$

missä viimeinen yhtäsuuruus seuraa kaavoista (2.12). Edelleen,

$$\det(\Sigma_Y) = \det(AA^T) = [\det(A)]^2,$$

joten

$$\big|\det(A^{-1})\big| = \frac{1}{|\det(A)|} = \frac{1}{\sqrt{\det(\Sigma_Y)}}.$$

Siten

$$f_Y(y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_Y)}}\, e^{-\frac{1}{2}(y-\mu_Y)^T\, \Sigma_Y^{-1}(y-\mu_Y)}, \quad y \in \mathbb{R}^d.$$

Sanomme, että Y:llä on *multinormaalijakauma parametrein* μ_Y *ja* Σ_Y ja merkitsemme

$$Y \sim \mathrm{N}(\mu_Y, \Sigma_Y).$$

Matriisi $\Sigma_Y = AA^T$ on symmetrinen ja positiivisesti definiitti (HT). Multinormaalijakaumaan päädyttäisiin myös määrittelemällä sen tiheysfunktioksi yleisesti

$$f_Y(y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(C)}}\, e^{-\frac{1}{2}(y-b)^T\, C^{-1}(y-b)}, \tag{2.13}$$

missä $C \in \mathbb{R}^{d\times d}$ on symmetrinen ja positiivisesti definiitti ja $b \in \mathbb{R}^d$. Nimittäin, olkoon $C \in \mathbb{R}^{d\times d}$ mikä tahansa symmetrinen ja positiivisesti definiitti matriisi ja

$$C = U\Lambda U^T$$

matriisin C ominaisarvohajotelma, missä $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_i > 0$ kaikilla i. Silloin voidaan määritellä matriisin Λ "neliöjuuri"

$$\Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2})$$

ja edelleen $A = U\Lambda^{1/2}$, jolloin

$$\begin{aligned} AA^T &= (U\Lambda^{1/2})(U\Lambda^{1/2})^T \\ &= U\Lambda^{1/2}(\Lambda^{1/2})^T U^T \\ &= U\Lambda U^T = C. \end{aligned}$$

Siten, jos $X \sim \mathrm{N}(0, I_d)$, on satunnaisvektorilla $Y = AX + b$ tiheysfunktio (2.13).

Huomautus. Edellä olevien tulosten perusteella on tietokoneella helppo simuloida satunnaisvektorin $Y \sim \mathrm{N}(b, C)$ arvoja, kunhan on käytössä $\mathrm{N}(0,1)$-jakautuneita satunnaislukuja tuottava satunnaislukugeneraattori:

1. Generoi riippumattomat $X^{(1)}, \ldots, X^{(d)} \sim \mathrm{N}(0,1)$, jolloin $X = [X^{(1)}, \ldots, X^{(d)}]^T \sim \mathrm{N}(0, I_d)$.
2. Laske $Y = AX + b$, missä $C = AA^T$.
3. Toista n kertaa.

Tuloksena on silloin satunnaisotos $Y_1, \ldots, Y_n \sim \mathrm{N}(b, C)$.

Minkä "muotoinen" on jakauman $\mathrm{N}(\mu_X, \Sigma_X)$ tiheysfunktio f_X? Nyt siis

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_X)}}\, e^{-\frac{1}{2}(x-\mu_X)^T \Sigma_X^{-1}(x-\mu_X)}, \quad x \in \mathbb{R}^d.$$

”Muodon” määräävät tasa-arvopinnat $f_X(x) =$ vakio eli pinnat

$$(x - \mu_X)^T \Sigma_X^{-1} (x - \mu_X) = c,$$

$c =$ vakio. Olkoon $\Sigma_X = U \Lambda U^T$ ja tehdään muuttujan vaihto

$$\begin{aligned} y &= U^T (x - \mu_X), \\ x &= \mu_X + U y. \end{aligned} \tag{2.14}$$

Silloin

$$\begin{aligned} (x - \mu_X)^T \Sigma_X^{-1} (x - \mu_X) &= (Uy)^T U \Lambda^{-1} U^T U y \\ &= y^T U^T U \Lambda^{-1} U^T U y \\ &= y^T \Lambda^{-1} y = \sum_{i=1}^{d} \frac{(y^{(i)})^2}{\lambda_i}, \end{aligned}$$

koska $U^T U = I_d$.

Kaava (2.14) voidaan tulkita siten, että origo siirretään μ_X:ään ja koordinaatisto kierretään siten, että uusiksi kantavektoreiksi tulevat $u_1, \ldots, u_d$ (U:n pystyrivit). Tasa-arvopinnoilla on tällöin yhtälö

$$\sum_{i=1}^{d} \frac{(y^{(i)})^2}{\lambda_i} = c$$

eli

$$\sum_{i=1}^{d} \left(\frac{y^{(i)}}{\sqrt{c\lambda_i}} \right)^2 = 1.$$

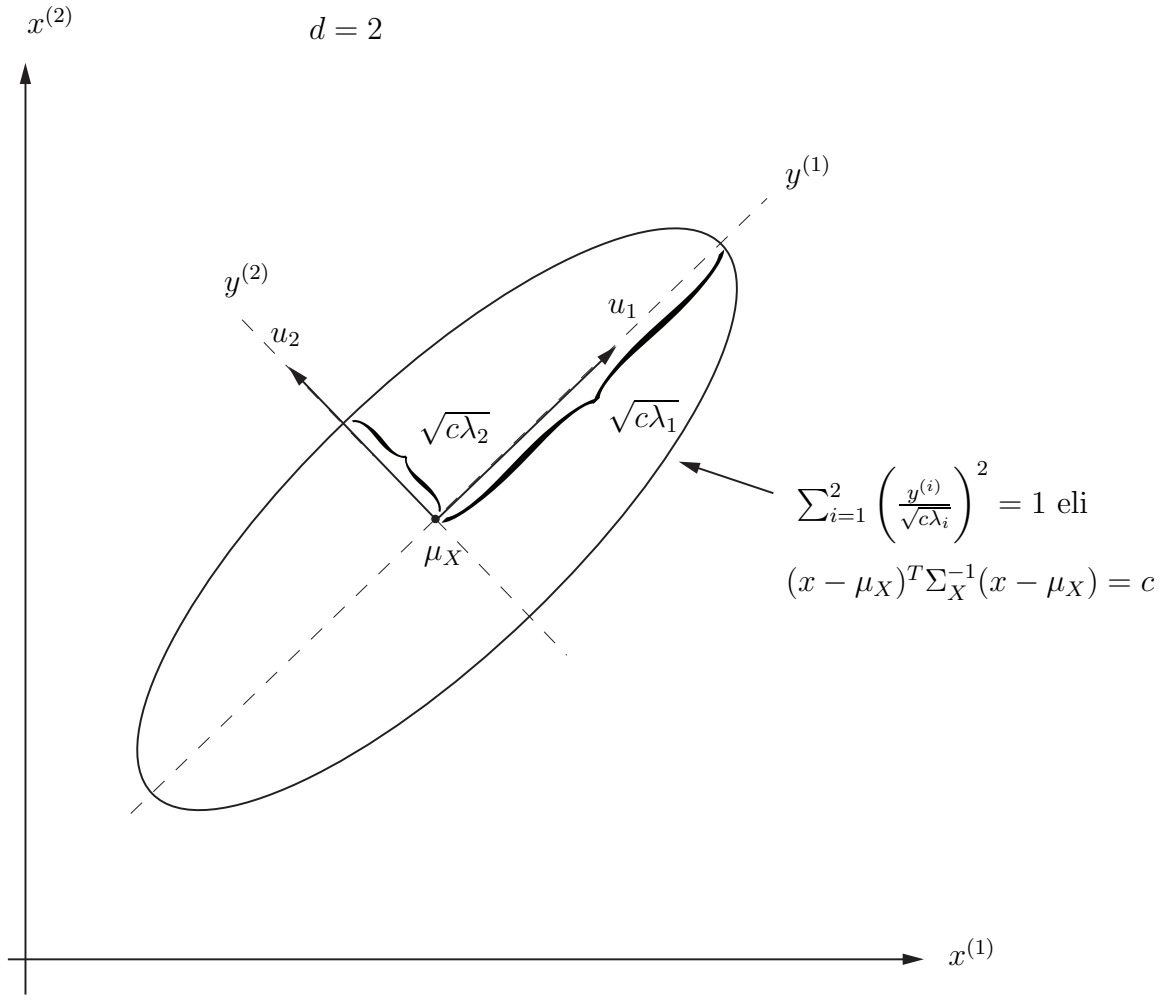
Tämä on avaruuden $\mathbb{R}^d$ *ellipsoidi* puoliakselien pituuksin $\sqrt{c\lambda_i}$. Tilannetta on havainnollistettu kuvassa 2.4 tapauksessa $d = 2$.

Edelleen, jos

$$Y = U^T (X - \mu_X) = [Y^{(1)}, \ldots, Y^{(d)}]^T,$$

on

$$\mu_Y = \mathbb{E}\big[U^T (X - \mu_X)\big] = 0$$



Kuva 2.4: Kaksiulotteisen normaalijakauman tasa-arvokäyrä.

ja, koska Y on muotoa $AX + b$, missä $A = U^T$, on (2.10):n nojalla

$$\Sigma_Y = U^T \Sigma_X U = U^T U \Lambda U^T U = \Lambda,$$

joten $\lambda_i = D^2(Y^{(i)})$.

2.5 Valkaisu

Olkoon satunnaisvektorin X kovarianssimatriisi Σ_X positiivisesti definiitti. Silloin

$$\Sigma_X = U \Lambda U^T,$$

missä $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, ja $\lambda_i > 0$ kaikilla i. Määritellään

$$\Lambda^{-1/2} = \operatorname{diag}(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2}).$$

Silloin selvästi $\Lambda^{-1/2} = (\Lambda^{1/2})^{-1}$, joten

$$(U\Lambda^{1/2})^{-1} = \Lambda^{-1/2} U^T.$$

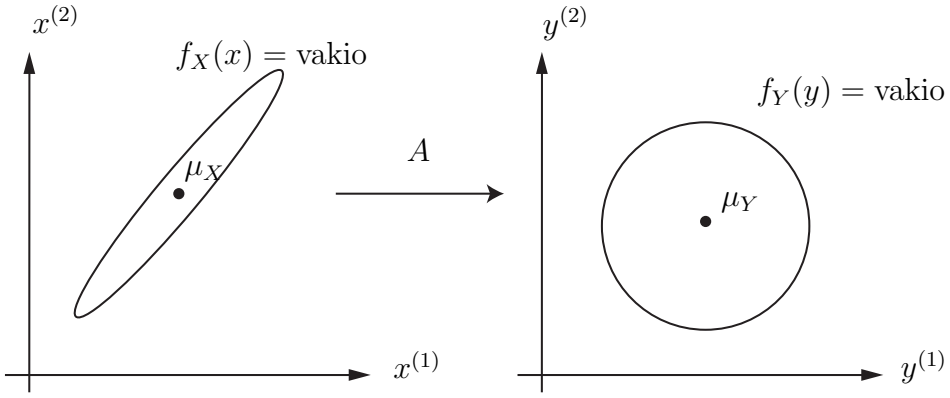
Olkoon $A = \Lambda^{-1/2} U^T$ ja $Y = AX$. Silloin (vrt. (2.10))

$$\begin{aligned} \Sigma_Y &= A \Sigma_X A^T \\ &= \Lambda^{-1/2} U^T U \Lambda U^T U \Lambda^{-1/2} \\ &= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I_d, \end{aligned}$$

koska $U^T U = I_d$. Sanomme, että muunnos $Y = AX$ *valkaisee* X:n. Satunnaisvektorille $Y = [Y^{(1)}, \ldots, Y^{(d)}]^T$ pätee:

- $D^2\big(Y^{(i)}\big) = 1$ kaikilla i,
- $\operatorname{Cov}\big(Y^{(i)}, Y^{(j)}\big) = 0$ eli $Y^{(i)}$ ja $Y^{(j)}$ ovat *korreloimattomia* kaikilla $i \neq j$.

Huomautus. Jos $X \sim \mathrm{N}(\mu_X, \Sigma_X)$, ovat $Y^{(1)}, \ldots, Y^{(d)}$ itseasiassa *riippumattomia*: $f_Y = f_{Y^{(1)}} \cdots f_{Y^{(d)}}$.



Kuva 2.5: Valkaisu havainnollistettuna tiheysfunktioden tasa-arvokäyrillä.

Valkaisua on havainnollistettu kuvassa 2.5. Siinä erityisesti Σ_X *diagonalisoidaan*. Hahmontunnistuksessa on joskus hyötyä siitä, että kaksi kovarianssimatriisia voidaan myös diagonalisoida *samanaikaisesti*. Olkoot C_1 ja C_2 symmetrisiä ja C_1 lisäksi positiivisesti definiitti ominaisarvohajotelmalla

$$C_1 = U\Lambda U^T.$$

Merkitään $C_1^{-1/2} = U\Lambda^{-1/2}U^T$. Tällöin $C_1^{-1/2}C_2C_1^{-1/2}$ on symmetrinen:

$$\left(C_1^{-1/2}C_2C_1^{-1/2}\right)^T = (C_1^{-1/2})^T C_2^T (C_1^{-1/2})^T$$

ja

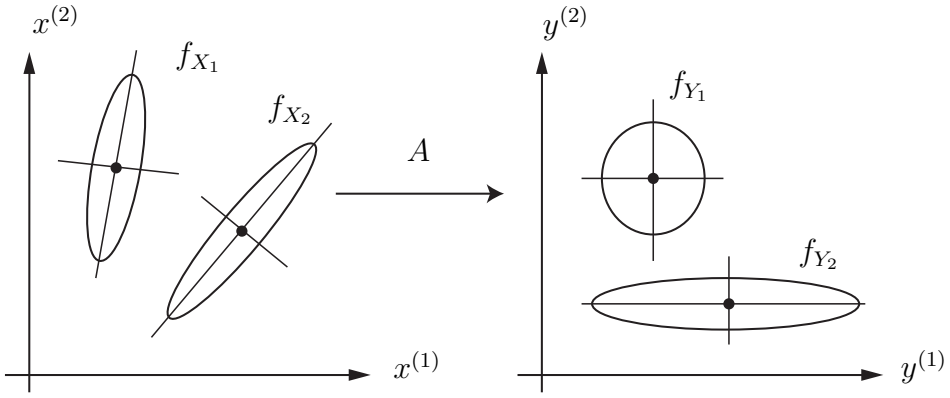
$$(C_1^{-1/2})^T = (U\Lambda^{-1/2}U^T)^T = (U^T)^T(\Lambda^{-1/2})^T U^T = U\Lambda^{-1/2}U^T = C_1^{-1/2},$$

sekä $C_2^T = C_2$, koska C_2 on symmetrinen. Siten matriisilla $C_1^{-1/2}C_2C_1^{-1/2}$ on ominaisarvohajotelma,

$$\begin{aligned} &C_1^{-1/2}C_2C_1^{-1/2} = VDV^T, \\ &D = \text{diag}(\delta_1, \ldots, \delta_d). \end{aligned} \tag{2.15}$$

Olkoon nyt

$$A = V^T C_1^{-1/2} = V^T U\Lambda^{-1/2}U^T.$$



Kuva 2.6: Kahden kovarianssin samanaikainen diagonalisointi havainnollistettuna tiheysfunktioiden tasa-arvokäyrillä.

Silloin

$$AC_1A^T = \left(V^TU\Lambda^{-1/2}U^T\right)\left(U\Lambda U^T\right)\left(U\Lambda^{-1/2}U^TV\right) = I_d,$$
$$AC_2A^T = V^TC_1^{-1/2}C_2C_1^{-1/2}V = D,$$

missä viimeisessä yhtälössä sovellettiin kaavaa (2.15). Näemme, että A diagonalisoi sekä matriisin C_1 että matriisin C_2. Erityisesti, jos X_1 ja X_2 ovat satunnaisvektoreita kovarianssimatriiseilla C_1 (positiivisesti definiitti) ja C_2, on

$$\begin{cases} Y_1 = AX_1 & \text{valkaisu} \quad (\Sigma_{Y_1} = I_d) \\ Y_2 = AX_2 & \text{diagonalisointi} \quad (\Sigma_{Y_2} = D). \end{cases}$$

Sekä Y_1:n että Y_2:n komponentit ovat *korreloimattomia*. Multinormaalissa tapauksessa ne ovat jopa *riippumattomia*. Tilannetta on havainnollistettu kuvassa 2.6.

Huomautus. Multinormaalissa tapauksessa tasa-arvo ellipsoidien akselit ovat kovarianssimatriisin ominaisvektoreita; diagonaalisen kovarianssin tapauksessa ominaisvektorit ovat tietysti kantavektorit $[0, \ldots, 0, \overset{i}{1}, 0, \ldots, 0]^T$, $i = 1, \ldots, d$.

Luku 3

Päätösteoriaa

3.1 Hahmontunnistuksen matemaattinen malli

Olkoon $X : \Omega \to \mathbb{R}^d$ satunnaisvektori, jolla on jatkuva jakauma. Piirrevektoria $x \in \mathbb{R}^d$ ajatellaan X:n arvona, eli *realisaationa*,

$$X(\omega) = x,$$

missä $\omega \in \Omega$ vastaa tehtyä mittausta tai havaintoa. Tämä on tapa mallintaa x:n satunnaisuutta. Olkoon $c \in \mathbb{N}$ ja oletetaan, että x voi olla peräisin jostain c:stä mahdollisesta luokasta eli x on peräisin jostain luokasta $j \in \{1, 2, \ldots, c\}$. Piirrevektorin X luokkaa mallinnetaan diskreetillä satunnaismuuttujalla $J : \Omega \to \{1, \ldots, c\}$. Siis

$$J(\omega) = j.$$

Koko piirrevektoria mallinnetaan satunnaismuuttujaparilla (X, J).

Huomautus. On syytä korostaa, että satunnaismuuttujat X ja J riippuvat toisistaan. Lisäksi yleisesti ottaen J ei suinkaan määräydy yksikäsitteisesti

X:n perusteella: täsmälleen sama x voi periaatteessa olla peräisin useammasta kuin yhdestä luokasta j.

Hahmontunnistuksen luokittelua koskevan ongelman voikin pelkistää muotoon ”kun x on havaittu, on arvattava siihen liittyvä luokka j”.

Esimerkki 3.1. Postinumeron automaattinen luku kirjekuoresta.

Yksittäistä numeroa tunnistettaessa meillä on $c = 10$ luokkaa:

luokka j	merkitys	
j	luku j	$(j = 1, \ldots, 9)$
10	luku 0	

Piirrevektori x on muodostettu tunnistettavaa numeroa esittävästä digitaalisesta kuvasta. ‖

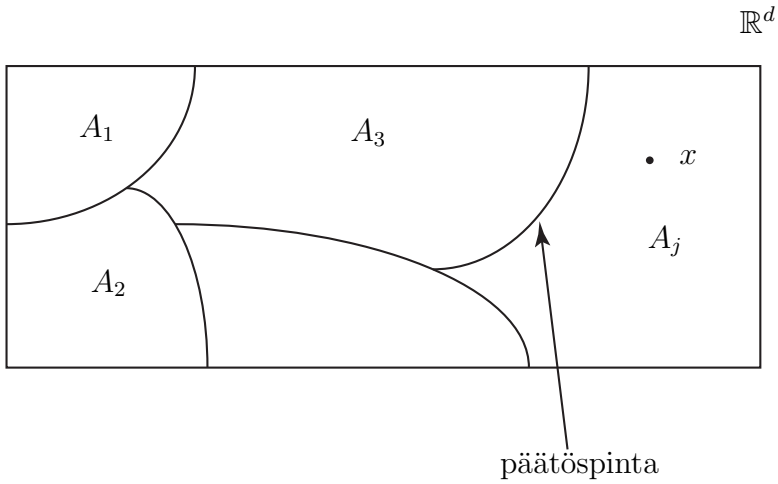
Esimerkki 3.2. Joskus on kehitelty ajatusta lentokentällä käytettävästä turvaportista, joka voisi erottaa toisistaan toisaalta joukon pieniä metalliesineitä ja toisaalta ison metalliesineen, joka voisi olla vaikka käsiase. Näin voitaisiin toivottavasti pienentää väärien hälytysten määrää. Tällöin siis meillä olisi $c = 3$ luokkaa:

luokka j	merkitys
1	ei metallia
2	iso metalliesine
3	paljon pieniä metalliesineitä

Piirrevektori x muodostetaan portin keloissa mitatun sähkömagneettinen heijastuksen (kelojen jännitteiden) avulla. ‖

Luokitteluongelma ratkaistaan määrittelemällä *luokitin* $g : \mathbb{R}^d \to \{1, \ldots, c\}$ (Borelin funktio) ja käyttämällä sitten päätössääntöä

$$g(x) = x\text{:n luokka.}$$



Kuva 3.1: Luokittimen päätösalueiden synnyttämä avaruuden $\mathbb{R}^d$ ositus.

Päätössääntö on siis täysin deterministinen: tietty havaittu x luokitellaan aina samaan luokkaan. Luokitin g itseasiassa jakaa piirreavaruuden $\mathbb{R}^d$ pistevieraisiin *päätösalueisiin*

$$A_j = g^{-1}(\{j\}) = \{x \in \mathbb{R}^d \mid g(x) = j\}, \quad j = 1, \ldots, c.$$

Siten $g(x) = j$ kaikilla $x \in A_j$ eli kaikki joukon A_j vektorit luokitellaan luokkaan j. Päätöalueet A_j ovat Borelin joukkoja ja ne muodostavat avaruuden $\mathbb{R}^d$ pistevieraan osituksen,

$$\mathbb{R}^d = \bigcup_{j=1}^{c} A_j, \quad A_i \cap A_j = \emptyset, \text{ kun } i \neq j.$$

Päätösalueita erottavat toisistaan *päätöspinnat*. Kun g on riittävän säännöllinen, ovat päätöspinnat myös matemaattisessa mielessä ”pintoja”. Tilannetta on havainnollistettu kuvassa 3.1.

Huomautus. Yleensä kaikki luokasta j peräisin olevat piirrevektorit *eivät* satu täsmälleen alueeseen A_j. Edelleen, kuten edellä huomautettiin, sama x voi myös olla peräisin usesta eri luokasta. Tällöin ei yleensä voida välttää sitä, että tietyllä todennäköisyydellä jotkut havainnot luokitellaan väärin. Pyrkimyksenä tavallisesti onkin valita g siten, että tämä virhetodennäköisyys minimoituu.

Esimerkki 3.3. Sukupuolta voi yrittää määrätä ihmisen pituuden perusteella. Vaikka miehet yleensä ovatkin vähän pitempiä, voi miehellä ja naisella kuitenkin olla täsmälleen sama pituus. ‖

Esimerkki 3.4. Turvaportin tapaukset $j = 2, 3$ voivat helposti synnyttää samanlaisen sähkömagneettisen heijastuksen. ‖

3.2 Bayesin luokitin

Luokan j *prioritodennäköisyys* on todennäköisyys, että satunnainen piirrevektori X on peräisin luokasta j,

$$P_j = \mathbb{P}(J = j), \qquad j = 1, \ldots, c.$$

Esimerkki 3.5. Postinumeron tunnistuksessa luonteva oletus voi olla $P_j = 1/10$ kaikilla $j = 1, \ldots, 10$. ‖

Esimerkki 3.6. Turvaporttiesimerkissä tilanne on selvästikin paljon ongelmallisempi. Mikä esimerkiksi on todennäköisyys, että henkilö kantaa asetta? ‖

Oletetaan jatkossa, että $P_j > 0$ kaikilla j. Luokan j *ehdollinen tiheysfunktio* eli *luokkatiheysfunktio* f_j on tiheysfunktio, jolle

$$\mathbb{P}(X \in B | J = j) = \int_B f_j(x)dx, \qquad B \subset \mathbb{R}^d. \tag{3.1}$$

Tässä

$$\mathbb{P}(X \in B | J = j) = \frac{\mathbb{P}(X \in B \text{ ja } J = j)}{\mathbb{P}(J = j)} \tag{3.2}$$

on todennäköisyys, että $X \in B$ ehdolla $J = j$. Tiheysfunktio f_j siis kuvaa piirrevektoreiden X jakauman luokassa j. Tätä voidaan merkitä $(X|J = j) \sim f_j$ ja myös esimerkiksi vaikka $(X|J = j) \sim \mathrm{N}(\mu_j, \Sigma_j)$, kun luokan jakauma

on multinormaali. Oletamme jatkossa, että luokkatiheysfunktiot f_j löytyvät kaikille luokille $j = 1, \ldots, c$.

Osoitetaan, että kaava

$$f(x) = \sum_{j=1}^{c} P_j f_j(x), \quad x \in \mathbb{R}^d$$

määrittelee satunnaisvektorin X tiheysfunktion. Jos nimittäin $B \subset \mathbb{R}^d$, on ensinnäkin

$$\begin{aligned}
\int_B f(x)dx &= \int_B \sum_{j=1}^{c} P_j f_j(x)dx \\
&= \sum_{j=1}^{c} P_j \int_B f_j(x)dx \\
&= \sum_{j=1}^{c} P_j \, \mathbb{P}(X \in B | J = j) \\
&= \sum_{j=1}^{c} \mathbb{P}(J = j) \, \mathbb{P}(X \in B | J = j). \qquad (3.3)
\end{aligned}$$

Todennäköisyyslaskennan tavanomaisin merkinnöin $X \in B$ tarkoittaa tapahtumaa

$$\{\omega \in \Omega \mid X(\omega) \in B\} \equiv \{X \in B\}$$

ja $J = j$ tarkoittaa tapahtumaa

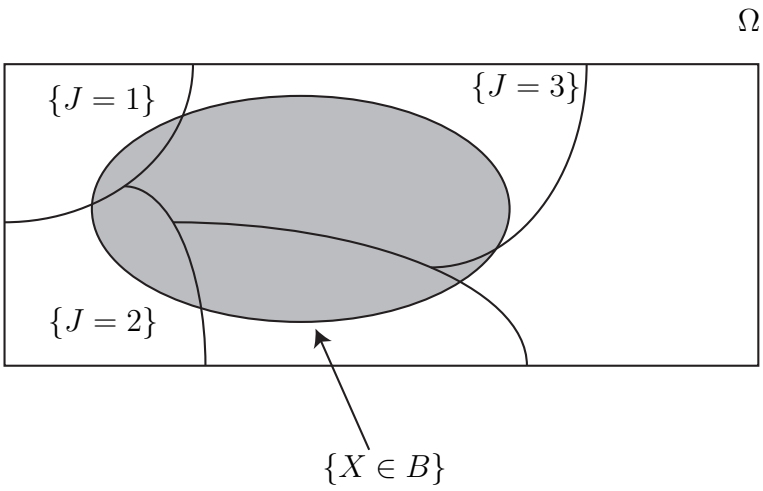
$$\{\omega \in \Omega \mid J(\omega) = j\} \equiv \{J = j\}.$$

Tapahtumat $\{J = j\}$, $j = 1, \ldots, c$ selvästi muodostavat perusjoukon Ω pistevieraan osituksen (kuva 3.2),

$$\Omega = \bigcup_{j=1}^{c} \{J = j\},$$

joten kokonaistodennäköisyyden kaavan nojalla saadaan

$$\mathbb{P}(X \in B) = \sum_{j=1}^{c} \mathbb{P}(J = j) \, \mathbb{P}(X \in B | J = j). \qquad (3.4)$$



Kuva 3.2: Kokonaistodennäköisyyden kaavan soveltaminen päätösalueiden muodostamaan ositukseen.

Siten

$$\mathbb{P}(X \in B) = \int_B f(x)dx,$$

joten f on todellakin kaavojen (3.3) ja (3.4) nojalla X:n tiheysfunktio.

Luokan j *posterioritodennäköisyys* (pisteessä x) on ehdollinen todennäköisyys

$$\mathbb{P}(J = j|X = x).$$

Se on todennäköisyys, että havaittu x tulee luokasta j. Merkitsemme jatkossa lyhyesti

$$\mathbb{P}(J = j|X = x) = P(j|x).$$

$P(j|x)$ voidaan laskea kaavasta (katso huomautus alla)

$$P(j|x) = \begin{cases} \frac{P_j f_j(x)}{f(x)}, & \text{kun } f(x) > 0, \\ 0, & \text{kun } f(x) = 0. \end{cases} \tag{3.5}$$

Nyt $P(j|x) \geq 0$ ja

$$\sum_{j=1}^{c} P(j|x) = \frac{1}{f(x)} \sum_{j=1}^{c} P_j f(x) = \frac{f(x)}{f(x)} = 1,$$

kun $f(x) > 0$.

Huomautus. Todennäköisyyttä $\mathbb{P}(J = j|X = x)$ *ei* voi suoraan tulkita tapahtumien ehdollisena todennäköisyytenä sillä kaava

$$\mathbb{P}(J = j|X = x) = \frac{\mathbb{P}(J = j \text{ ja } X = x)}{\mathbb{P}(X = x)}$$

ei ole järkevä, koska

$$\mathbb{P}(X = x) = \int\limits_{\{x\}} f(u)du = 0.$$

Voidaan osoittaa, että (3.5) kuitenkin yhtyy ehdollisen todennäköisyyden tarkkaan, mittateoreettisen todennäköisyyslaskennan määritelmään. Tämän kurssin tarpeisiin tyydymme kuitenkin vain seuraavaan heuristiseen perusteluun.

Olkoon $\varepsilon > 0$ pieni ja

$$B(x, \varepsilon) = \big\{u \in \mathbb{R}^d \mid \|u - x\| < \varepsilon\big\}$$

pieni x:n ympäristö. Silloin

$$\begin{aligned}\int\limits_{B(x,\varepsilon)} P(j|u)f(u)du &\approx P(j|x) \int\limits_{B(x,\varepsilon)} f(u)du \\ &= P(j|x)\, P\big(X \in B(x, \varepsilon)\big).\end{aligned}$$

Toisaalta (3.5):n nojalla

$$\begin{aligned}\int\limits_{B(x,\varepsilon)} P(j|u)f(u)du &= \int\limits_{B(x,\varepsilon)} \frac{P_j f_j(u)}{f(u)} \cdot f(u)du \\ &= P_j \int\limits_{B(x,\varepsilon)} f_j(u)du \\ &= P_j \mathbb{P}\big(X \in B(x, \varepsilon)|J = j\big) \\ &= P_j \frac{\mathbb{P}\big(X \in B(x, \varepsilon) \text{ ja } J = j\big)}{P_j} \\ &= \mathbb{P}\big(X \in B(x, \varepsilon) \text{ ja } J = j\big),\end{aligned}$$

missä toiseksi viimeinen ja viimeinen yhtäsuuruus seurasivat vastaavasti kaavoista (3.1) ja (3.2). Siten

$$P(j|x)\,\mathbb{P}\big(X \in B(x,\varepsilon)\big) \approx \mathbb{P}\big(X \in B(x,\varepsilon) \text{ ja } J = j\big)$$

eli

$$P(j|x) \approx \frac{\mathbb{P}\big(X \in B(x,\varepsilon) \text{ ja } J = j\big)}{\mathbb{P}\big(X \in B(x,\varepsilon)\big)} = \mathbb{P}\big(J = j | X \in B(x,\varepsilon)\big)$$

mikä on todennäköisyys, että $J = j$ ehdolla $X \in B(x,\varepsilon)$. Kun $\varepsilon \to 0$, voidaan $P(j|x)$ siis todellakin tulkita todennäköisyydeksi, että havaittu x on peräisin luokasta j. Kaavassa (3.5) määritelmä $P(j|x) = 0$, kun $f(x) = 0$, on puolestaan puhtaasti vain sopimus.

Olkoon $z_1, \ldots, z_n \in \mathbb{R}$. Merkintä

$$k = \operatorname*{argmax}_{j=1,\ldots,n} z_j$$

tarkoittaa, että z_k on suurin luvuista $z_1, \ldots, z_n$,

$$z_k = \max\{z_1, \ldots, z_n\}.$$

Saadaksemme yksikäsitteisesti sen k:n, sovimme, että k on *pienin* l, jolla $z_l = \max\{z_1, \ldots, z_n\}$. Vastaavasti määritellään merkintä

$$k = \operatorname*{argmin}_{j=1,\ldots,n} z_j.$$

Esimerkki 3.7. Olkoon $(z_1, z_2, z_3, z_4, z_5, z_6, z_7) = (2, 1, 3, 1, 4, 2, 4)$. Silloin

$$\operatorname*{argmax}_{j=1,\ldots,7} z_j = 5, \quad \operatorname*{argmin}_{j=1,\ldots,7} z_j = 2.$$

‖

Määritelmä 3.8. *Bayesin luokitin* on kuvaus $g^* : \mathbb{R}^d \to \{1, \ldots, c\}$,

$$g^*(x) = \operatorname*{argmax}_{j=1,\ldots,c} P(j|x), \quad x \in \mathbb{R}^d. \tag{3.6}$$

Siis: $g^*(x) \in \{1, \ldots, c\}$ on luokka, jolle

$$\mathbb{P}(g^*(x)|x) = \max_{j=1,\ldots,c} P(j|x) \tag{3.7}$$

eli $g^*(x)$ on luokka, jolla on suurin posterioritodennäköisyys.

Koska

$$P(j|x) = \frac{P_j\, f_j(x)}{f(x)},$$

voitaisiin yhtä hyvin määritellä

$$g^*(x) = \operatorname*{argmax}_{j=1,\ldots,c} P_j\, f_j(x) \tag{3.8}$$

eli

$$P_{g^*(x)} f_{g^*(x)}(x) = \max_{j=1,\ldots,c} P_j\, f_j(x). \tag{3.9}$$

Tämä pätee myös, kun $f(x) = 0$. Silloin $P(j|x) = 0$ ja $P_j f_j(x) = 0$ kaikilla j, joten $g^*(x) = 1$ sekä (3.6):ssa, että (3.8):ssa.

Määritelmä 3.9. Luokittimen $g : \mathbb{R}^d \to \{1, \ldots, c\}$ *virhetodennäköisyys* eli *luokitteluvirhe* on

$$e(g) = \mathbb{P}\big(g(X) \neq J\big).$$

Tässä symboli "e" tulee sanasta "error" (virhe).

Huomautus. Koska J on X:n oikea luokka, on $g(X) \neq J$ jos ja vain jos luokitin g tekee virheen. Siten

$$e(g) = \text{ todennäköisyys, että } g \text{ tekee virheen.}$$

Lause 3.10. *Bayesin luokittimella on pienin virhetodennäköisyys.*

Todistus. Olkoon $g^* : \mathbb{R}^d \to \{1, \ldots, c\}$ Bayesin luokitin (3.6) (tai (3.8)) ja $g : \mathbb{R}^d \to \{1, \ldots, c\}$ mielivaltainen toinen luokitin. On näytettävä, että

$$e(g^*) \leq e(g).$$

Saadaan,

$$e(g) = \mathbb{P}\big(g(X) \neq J\big) = 1 - \mathbb{P}\big(g(X) = J\big).$$

Nyt joukot

$$\{J = j\} \equiv \{\omega \in \Omega \mid J(\omega) = j\}$$

muodostavat perusjoukon Ω pistevieraan osituksen:

$$\Omega = \bigcup_{j=1}^{c} \{J = j\} \quad \text{ja} \quad \{J = j\} \cap \{J = k\} = \emptyset, \text{ kun } j \neq k.$$

Kokonaistodennäköisyyden kaavan nojalla saadaan silloin

$$\begin{aligned}
\mathbb{P}\big(g(X) = J\big) &= \sum_{j=1}^{c} \mathbb{P}\big(g(X) = J | J = j\big)\mathbb{P}(J = j) \\
&= \sum_{j=1}^{c} P_j \, \mathbb{P}\big(g(X) = j | J = j\big) \\
&= \sum_{j=1}^{c} P_j \, \mathbb{P}\big(X \in g^{-1}(j) | J = j\big) \\
&= \sum_{j=1}^{c} P_j \int\limits_{A_j} f_j(x)dx.
\end{aligned}$$

Tässä

$$A_j = g^{-1}(j) = \{x \in \mathbb{R}^d \mid g(x) = j\},$$

ja viimeisessä yhtäsuuruudessa käytettiin kaavaa (3.1). Edelleen,

$$P_j \int\limits_{A_j} f_j(x)dx = \int\limits_{A_j} P_j f_j(x)dx = \int\limits_{A_j} P_{g(x)} f_{g(x)}(x)dx.$$

Toisaalta

$$\mathbb{R}^d = \bigcup_{j=1}^{c} A_j, \qquad A_j \cap A_k = \emptyset, \text{ kun } j \neq k,$$

joten

$$\sum_{j=1}^{c} P_j \int\limits_{A_j} f_j(x)dx = \sum_{j=1}^{c} \int\limits_{A_j} P_{g(x)} f_{g(x)}(x)dx = \int\limits_{\mathbb{R}^d} P_{g(x)} f_{g(x)}(x)dx.$$

Siten

$$e(g) = 1 - \int_{\mathbb{R}^d} P_{g(x)} f_{g(x)}(x) dx. \tag{3.10}$$

Kaavasta (3.9) seuraa Bayesin luokittelijalle g^*, että

$$P_{g(x)}\, f_{g(x)}(x) \leq P_{g^*(x)}\, f_{g^*(x)}(x) \quad \text{kaikilla } x \in \mathbb{R}^d.$$

Tuloksen (3.10) nojalla saadaan silloin

$$e(g) \geq 1 - \int_{\mathbb{R}^d} P_{g^*(x)}\, f_{g^*(x)}(x)\, dx = e(g^*).$$

□

Suure $e(g^*)$ on pariin (X, J) liittyvä *Bayesin virhetodennäköisyys.*

Huomautus. Sama virhetodennäköisyys saataisiin, jos valittaisiin $g^*(x)$ tasapelien tapauksessa jotenkin toisinkin sillä tasapelithän eivät mitenkään näkyneet todistuksessa.

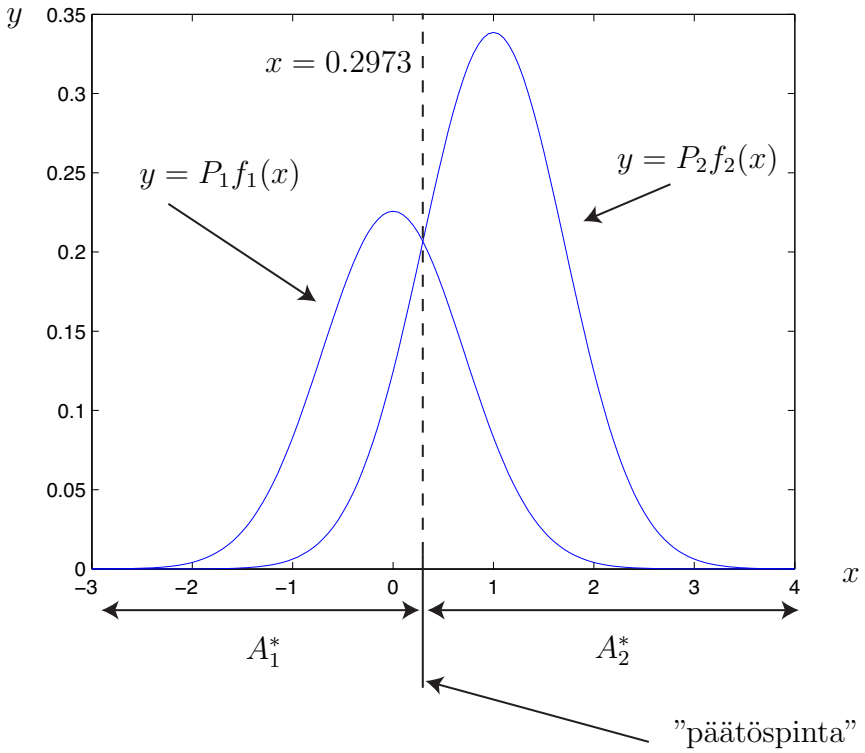
Voitaisiin kysyä miksi jatkaa teorian kehittelyä enää tästä eteenpäin, kun kerran optimaalinen luokitin on nyt löydetty? Syy on siinä, että harvoin suureet P_j ja f_j ovat tarkkaan tiedossa. Niitä joudutaan käytännössä jotenkin aina arvioimaan, estimoimaan ja Bayesin luokitin jää vain ideaaliksi, jota kuitenkaan ei yleensä voida tavoittaa. Tähän palataan myöhemmin.

Esimerkki 3.11. Olkoon $d = 1$, $c = 2$ ja luokkien jakaumat normaaleja,

$$\text{Luokka 1}: \quad f_1(x) = \frac{1}{\sqrt{\pi}}\, e^{-x^2}, \quad P_1 = 0.4$$
$$\text{Luokka 2}: \quad f_2(x) = \frac{1}{\sqrt{\pi}}\, e^{-(x-1)^2}, \quad P_2 = 0.6$$

Siis,

$$(X|J = 1) \sim \mathrm{N}\big(0, (1/\sqrt{2})^2\big),$$
$$(X|J = 2) \sim \mathrm{N}\big(1, (1/\sqrt{2})^2\big).$$



Kuva 3.3: Esimerkin 3.11 luokkatiheysfunktiot ja päätösalueet.

Bayesin luokittimen päätössääntö on

$$\text{jos } P_1\, f_1(x) \geq P_2\, f_2(x),\ x \text{ luokitellaan luokkaan 1}$$
$$\text{jos } P_1\, f_1(x) < P_2\, f_2(x),\ x \text{ luokitellaan luokkaan 2}$$

Merkitään tätä lyhyesti

$$P_1\, f_1(x) \underset{2}{\overset{1}{\gtrless}} P_2\, f_2(x).$$

Olkoon $l(x) = f_1(x)/f_2(x)$ (ns. *uskottavuusosamäärä*). Silloin Bayesin päätössääntö voidaan kirjoittaa muodossa

$$l(x) \underset{2}{\overset{1}{\gtrless}} \frac{P_2}{P_1} \quad \text{eli} \quad s(x) = -\log\big(l(x)\big) \underset{2}{\overset{1}{\lessgtr}} \log\Big(\frac{P_1}{P_2}\Big),$$

missä log tarkoittaa luonnollista logaritmia.

Nyt

$$\begin{aligned} s(x) &= -\log\big(l(x)\big) \\ &= -\log\left(\frac{\frac{1}{\sqrt{\pi}}\, e^{-x^2}}{\frac{1}{\sqrt{\pi}}\, e^{-(x-1)^2}}\right) \\ &= -\log e^{-x^2+(x-1)^2} \\ &= -\big(-x^2 + (x-1)^2\big) = 2x - 1. \end{aligned}$$

Siten Bayesin luokitin on

$$2x - 1 \underset{2}{\overset{1}{\lessgtr}} \log\Big(\frac{P_1}{P_2}\Big) \qquad \text{eli} \qquad x \underset{2}{\overset{1}{\lessgtr}} \frac{1}{2} + \frac{1}{2}\log\Big(\frac{P_1}{P_2}\Big).$$

Näin ollen

$$g^*(x) = \begin{cases} 1, & \text{kun } x \leq \frac{1}{2} + \frac{1}{2}\log\big(\frac{P_1}{P_2}\big), \\ 2, & \text{kun } x > \frac{1}{2} + \frac{1}{2}\log\big(\frac{P_1}{P_2}\big). \end{cases}$$

Tässä

$$\frac{1}{2} + \frac{1}{2}\log\big(\frac{P_1}{P_2}\big) = \frac{1}{2} + \frac{1}{2}\log\big(\frac{0.4}{0.6}\big) \approx 0.2973.$$

Siten optimaaliset päätösalueet ovat likimain (kuva 3.3)

$$A_1^* = \{x \mid g^*(x) = 1\} =]-\infty, 0.2973]$$
$$A_2^* = \{x \mid g^*(x) = 2\} =]0.2973, \infty[$$

Luokitteluvirhe voidaan laskea kaavasta (3.10),

$$\begin{aligned} e(g^*) &= 1 - \int_{-\infty}^{\infty} P_{g^*(x)} f_{g^*(x)}(x)dx \\ &= 1 - \Big[\int_{-\infty}^{0.2973} P_1 f_1(x)dx + \int_{0.2973}^{\infty} P_2 f_2(x)dx \Big]. \end{aligned}$$

Olkoon

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}u^2} du$$

jakauman $\mathrm{N}(0,1)$ kertymäfunktio. Silloin

$$\begin{aligned} \int_{-\infty}^{0.2973} f_1(x)dx &\approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.4204} e^{-\frac{1}{2}u^2} du = \Phi(0.4204) \approx 0.6629, \\ \int_{0.2973}^{\infty} f_2(x)dx &\approx \frac{1}{\sqrt{2\pi}} \int_{-0.9938}^{\infty} e^{-\frac{1}{2}u^2} du = 1 - \Phi(-0.9938) \approx 0.8398, \end{aligned}$$

missä ensimmäisessä ja toisessa integraalissa tehtiin vastaavasti muuttujan vaihdokset $x = u/\sqrt{2}$ ja $x - 1 = u/\sqrt{2}$. Siten

$$e(g^*) \approx 1 - 0.4 \cdot 0.6629 - 0.6 \cdot 0.8398 \approx 0.2310.$$

Tätä pienempään virheeseen ei siis millään luokittimella voi päästä. ‖

Edellinen esimerkki yleistyy välittömästi avaruuteen $\mathbb{R}^d$ ja mielivaltaisille luokkatiheysfunktioille f_1, f_2. Määritellään

$$\begin{aligned} l(x) &= \frac{f_1(x)}{f_2(x)}, \\ s(x) &= -\log\big(l(x)\big), \end{aligned}$$

missä oletetaan, että $l(x) > 0$ kaikilla x. Bayesin luokittimen päätössääntönä on

$$l(x) \underset{2}{\overset{1}{\gtrless}} \frac{P_2}{P_1} \qquad \text{eli} \qquad s(x) \underset{2}{\overset{1}{\lessgtr}} \log\Big(\frac{P_1}{P_2}\Big)$$

jolloin päätöspinnaksi tulee

$$\big\{x \in \mathbb{R}^d \mid l(x) = P_2/P_1\big\} = \big\{x \in \mathbb{R}^d \mid s(x) = \log(P_1/P_2)\big\}.$$

Tarkastellaan vielä c:n luokan tapausta ja johdetaan luokitteluvirheelle uusi kaava. Olkoon $g : \mathbb{R}^d \to \{1, \ldots, c\}$ luokitin ja

$$A_j = \{x \in \mathbb{R}^d \mid g(x) = j\}, \quad j = 1, \ldots, c.$$

Silloin kokonaistodennäköisyyden kaavasta saadaan

$$\begin{aligned} e(g) &= \mathbb{P}\big(g(X) \neq J\big) \\ &= \sum_{j=1}^{c} P_j \, \mathbb{P}\big(g(X) \neq J | J = j\big) \\ &= \sum_{j=1}^{c} P_j \, \mathbb{P}\big(g(X) \neq j | J = j\big) \\ &= \sum_{j=1}^{c} P_j \int\limits_{\mathbb{R}^d \setminus A_j} f_j(x) dx \end{aligned}$$

eli

$$e(g) = \sum_{j=1}^{c} P_j \, e_j(g),$$

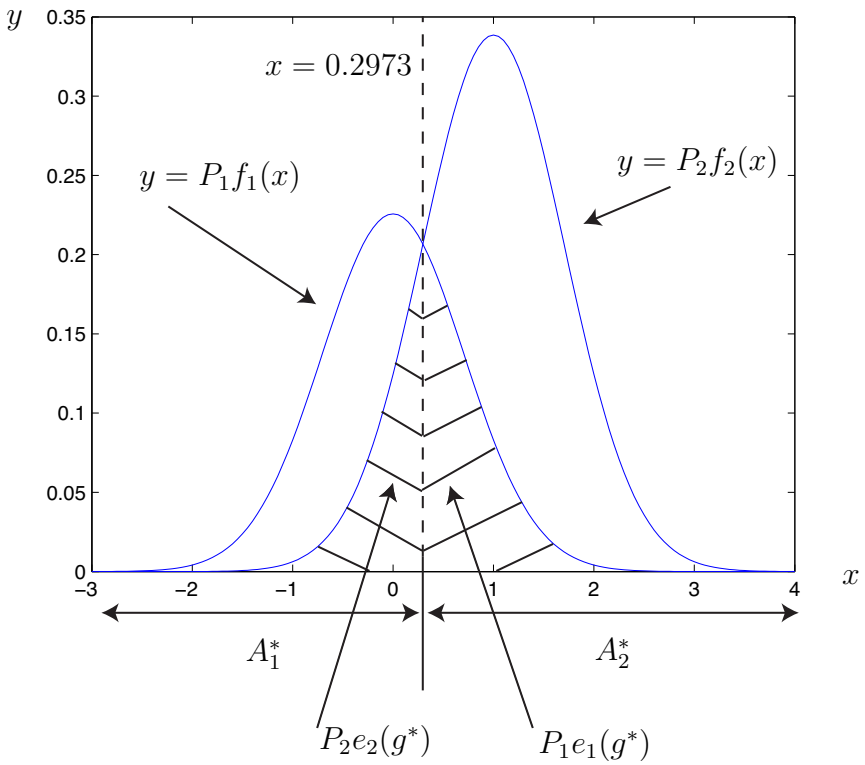
missä

$$e_j(g) = \int\limits_{\mathbb{R}^d \setminus A_j} f_j(x) dx = \mathbb{P}\big(g(X) \neq j | J = j\big)$$

on todennäköisyys, että luokan j piirrevektori luokitellaan väärin. Siten $e(g)$ on painotettu keskiarvo kunkin luokan ”sisäisestä” virhetodennäköisyydestä.

Erityisesti Bayesin virhetodennäköisyydelle saadaan

$$\begin{aligned} e(g^*) &= \sum_{j=1}^{c} P_j \, e_j(g^*), \qquad (3.11) \\ e_j(g^*) &= \int\limits_{\mathbb{R}^d \setminus A_j^*} f_j(x) dx, \\ A_j^* &= \big\{x \in \mathbb{R}^d \mid g^*(x) = j\big\}. \end{aligned}$$



Kuva 3.4: Esimerkin 3.11 luokkakohtaiset virheet.

Luokkakohtaisia virheitä $e_j(g^*)$ on havainnollistettu kuvassa 3.4. Siinä viivoitettujen pinta-alojen summa on $e(g^*) \approx 0.2310$.

Bayesin virhetodennäköisyys voidaan periaatteessa laskea kaavasta (3.11), jos P_j:t ja f_j:t tunnetaan. Käytännössä tämä on useinmiten mahdotonta, koska alueet A_j^* ovat monimutkaisia ja integraaleja $\int_{\mathbb{R}^d \setminus A_j^*} f_j(x)dx$ ei saada laskettua.

Olkoon kuitenkin $c = 2$ ja f_1 ja f_2 tunnettuja. Silloin on joskus mahdollista määrätä satunnaismuuttujan

$$s(X) = -\log\big(l(X)\big) = -\log\left(\frac{f_1(X)}{f_2(X)}\right)$$

luokkatiheysfunktiot g_j,

$$\mathbb{P}\big(s(X) \in B | J = j\big) = \int_B g_j(s)ds, \quad j = 1, 2,\ B \subset \mathbb{R}.$$

Tällöin

$$\begin{aligned} e_1(g^*) &= \mathbb{P}\big(g^*(X) \neq 1 | J = 1\big) = \mathbb{P}\big(g^*(X) = 2 | J = 1\big) \\ &= \mathbb{P}\left(s(X) > \log\left(\frac{P_1}{P_2}\right) \Big| J = 1\right) \\ &= \int_{\log(P_1/P_2)}^{\infty} g_1(s)ds \end{aligned}$$

ja vastaavasti $e_2(g^*)$:lle. Nämä ovat yksiulotteisia integraaleja ja siten käytännössä helpommin laskettavissa.

3.3 Muita optimaalisia luokittimia

3.3.1 Bayesin minimiriskiluokitin

Turvaportti on esimerkki luokitteluongelmasta, jossa eri virhetyypit eivät ole vakavuudeltaan samanarvoisia. Väärä hälytys, jossa matkustaja otetaan tarpeettomasti tarkempaan tutkimukseen on vähemmän haitallista kuin asetta kantavan terroristin päästäminen läpi.

Olkoon siis $\lambda(j, k) \geq 0$, $j, k = 1, \ldots, c$, *tappio*, kun oikea luokka on $J = j$ mutta luokitin g antaa $g(X) = k$.

Esimerkki 3.12. Turvaportti, jossa $c = 3$ ja

$j = 1$: ei metallia

$j = 2$: iso metalliesine

$j = 3$: paljon pieniä metalliesineitä

Jos toimitaan niin, että luokitukset $g(x) = 1$ ja $g(x) = 2$ eivät johda erliliseen manuaaliseen ruumiintarkastukseen, tappio λ voitaisiin nyt määritellä esimerkiksi seuraavasti:

		k		
		1	2	3
j	1	0	100	0
	2	1000	0	1000
	3	0	100	0

‖

Luokittimen hyvyyttä voidaan nyt mitata keskimääräisellä tappiolla eli *riskillä*,

$$R(g) = \mathbb{E}\,\lambda\big(J, g(X)\big).$$

Määritelmä 3.13. *Bayesin minimiriskiluokitin* on kuvaus $g^*_\lambda : \mathbb{R}^d \to \{1, \ldots, c\}$,

$$g^*_\lambda(x) = \operatorname*{argmin}_{k=1,\ldots,c} \sum_{j=1}^{c} \lambda(j,k)\, P(j|x), \quad x \in \mathbb{R}^d. \tag{3.12}$$

Lause 3.14. *Bayesin minimiriskiluokittimella on pienin riski.*

Todistus. Harjoitustehtävä. □

Bayesin luokitin g^* on erikoistapaus luokittimesta g^*_λ, kun valitaan

$$\lambda(j,k) = \begin{cases} 0, & j = k \quad \text{(oikea luokitus)} \\ 1, & j \neq k \quad \text{(virhe)}. \end{cases}$$

Tämän osoittaminen on myös harjoitustehtävä.

3.3.2 Hylkääminen

Joskus voi olla viisasta *kieltäytyä* luokittelemasta automaattisesti sellaista piirevektoria x, joka ei selvästi kuulu mihinkään luokkaan ja käsitellä tällaiset hankalat tapaukset vaikkapa manuaalisesti. Esimerkiksi erittäin epäselvästi kirjoitetun postinumeron luku voi onnistua ihmiseltä luotettavammin, koska hän saattaa pystyä hyödyntämään kaikkea kirjekuoressa olevaa osoiteinformaatiota.

Luokitin *hylkäysmahdollisuudella* on kuvaus $g : \mathbb{R}^d \to \{0, 1, \ldots, c\}$, missä 0 on *hylkäysluokka*. Hylkäys johtaa tyypillisesti piirrevektorin erityiskäsittelyyn, josta seuraa kustannus. Sopiva tappion muoto tällöin on esimerkiksi

$$\lambda(j,k) = \begin{cases} 0, & j = k \qquad \text{(oikea luokitus)} \\ t, & k = 0 \qquad \text{(hylkäys)} \\ 1, & \text{muuten} \quad \text{(virhe)}, \end{cases}$$

missä $j, k = 1, \ldots, c$, j on oikea luokka ja k on luokittimen antama luokka.

Lauseen 3.14 ja kaavan (3.12) nojalla optimaalinen luokitin g_t^* tällöin on

$$g_t^*(x) = \operatorname*{argmin}_{k=0,1,\ldots,c} \sum_{j=1}^{c} \lambda(j,k)\, P(j|x).$$

Lausetta 3.14 ja kaavaa (3.12) sovelletaan siis tässä $c+1$ luokan tilanteeseen, jossa $P(0|x) = 0$ kaikilla x, jolloin yllä olevassa summassa $j = 0$ voidaan jättää pois.

Kun $k = 0$,

$$\sum_{j=1}^{c} \lambda(j,k)\, P(j|x) = t \sum_{j=1}^{c} P(j|x) = t,$$

ja kun $k \neq 0$,

$$\sum_{j=1}^{c} \lambda(j,k) P(j|x) = \sum_{\substack{j=1 \\ j \neq k}}^{c} P(j|x) = 1 - P(k|x).$$

Siten

$$g_t^*(x) = \begin{cases} 0, \text{ kun } t \leq \min\limits_{j=1,\ldots,c} \big(1 - P(j|x)\big), \\ \operatorname{argmin}\limits_{j=1,\ldots,c} \big(1 - P(j|x)\big), \text{ muuten} \end{cases}$$

eli

$$g_t^*(x) = \begin{cases} 0, \text{ kun } \max\limits_{j=1,\ldots,c} P(j|x) \leq 1 - t, \\ \operatorname{argmax}\limits_{j=1,\ldots,c} P(j|x), \text{ muuten.} \end{cases}$$

Sääntö on siis seuraava:

- Piirrevektori hylätään, jos kaikki posterioritodennäköisyydet alle kynnysarvon $1 - t$ (mikään luokka ei ole kovin varma).
- Muuten tehdään Bayesin luokittimen mukainen päätös.

Havaitaan, että

- arvo $t < 0$ ei ole järkevä, koska tällöin $1 - t > 1$ ja kaikki piirrevektorit x hylätään,
- arvo $t > \frac{c-1}{c}$ ei myöskään ole ole järkevä, koska tällöin

  $$1 - t < 1 - \frac{c-1}{c} = \frac{c - c + 1}{c} = \frac{1}{c},$$

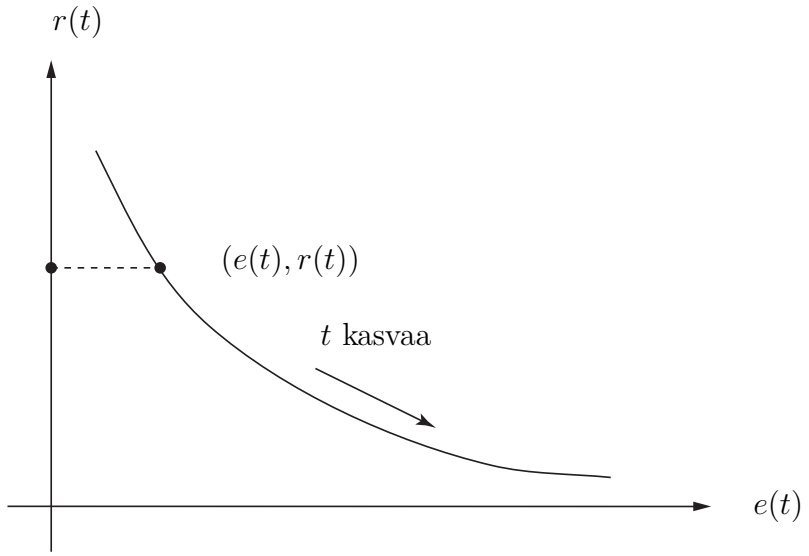
  jolloin $\max\limits_{j=1,\ldots,c} P(j|x) > 1 - t$ kaikilla x, joten mitään piirrevektoria x ei hylätä.

Siten tulee olla $0 \leq t \leq (c-1)/c$.

Olkoon nyt

$$\begin{aligned} r(t) &= \mathbb{P}\big(g_t^*(X) = 0\big), \\ e(t) &= \mathbb{P}\big(g_t^*(X) \neq 0 \text{ ja } g_t^*(X) \neq J\big) \end{aligned}$$

.



Kuva 3.5: Luokittimen virhe-hylkäyskäyrä. Sopiva t:n arvo voidaan löytää esimerkiksi kiinnittämällä haluttu hylkäystodennäköisyys $r(t)$.

Tässä $r(t)$ on todennäköisyys, että luokitin g_t^* tekee hylkäyksen ja $e(t)$ on puolestaan todennäköisyys, että g_t^* ei hylkää mutta tekee virheen. Tavoitteena on valita sellainen t, että $r(t)$ ja $e(t)$ ovat kummatkin pieniä. Nyt siis t ei enää ole tunnettu kustannus (so. jokin kiinteä arvo) vaan kynnysparamteri, jota säätämällä saadaa halutut $r(t)$ ja $e(t)$. Kuvaus $t \mapsto \big(e(t), r(t)\big)$ määrittelee ns. *virhe-hylkäyskäyrän* (kuva 3.5). Se voidaan likimain määrätä kokeellisesti suorittamalla eri t:n arvoilla luokitteluja luokittimella g_t^*. Käytännössä voidaan sopiva t:n arvo löytää kiinnittämällä $r(t)$:lle haluttu arvo ja katsomalla millä t:n arvolla se saavutetaan.

3.4 Neymanin-Pearsonin luokitin

Tämä luokitin sopii kahden luokan tapaukseen, kun prioritodennäköisyyksiä P_1, P_2 ei tunneta ja halutaan kontrolloida luokkakohtaisia virhetodennäköisyyksiä $e_1(g)$ ja $e_2(g)$.

Olkoon $u \in [0, \infty[$ ja määritellään luokittelija $g_u : \mathbb{R}^d \to \{1, 2\}$ kaavalla

$$\frac{f_1(x)}{f_2(x)} \underset{2}{\overset{1}{\gtrless}} u \tag{3.13}$$

mikä tarkoittaa, että

$$g_u(x) = \begin{cases} 1, & \text{kun } f_1(x) \geq u\, f_2(x), \\ 2, & \text{muuten.} \end{cases}$$

Huomautus. $l(x) = f_1(x)/f_2(x)$ on jo luvussa 3.2 esiintynyt uskottavuusosamäärä. Jos P_1 ja P_2 tunnettaisiin, saataisiin valinnalla $u = P_2/P_1$ itse asiassa Bayesin luokitin.

Merkitään nyt

$$e_j(u) = e_j(g_u) = \mathbb{P}\big(g_u(X) \neq j | J = j\big), \quad j = 1, 2,$$

joka on todennäköisyys, että g_u tekee virheen, kun $J = j$. Käyrä

$$u \mapsto (e_2(u), 1 - e_1(u))$$

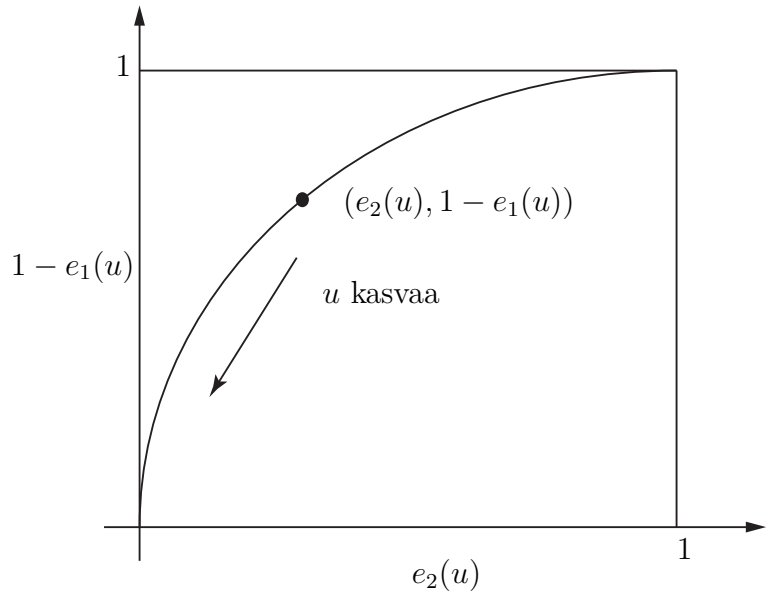
on luokittimen (3.13) *toimintaominaiskäyrä* (kuva 3.6). Se voidaan määrätä kokeellisesti ja valita sitten u, joka tuottaa halutut virheet $e_1(u)$ ja $e_2(u)$.

Luokitin (3.13) tunnetaan *Neymanin-Pearsonin luokittimena* ja sillä on seuraava optimaalisuusominaisuus.

Lause 3.15 (Neymanin-Pearsonin lemma)**.** *Olkoon g_u Neymanin-Pearsonin luokitin ja g jokin toinen luokitin. Silloin, jos $e_2(g) \leq e_2(g_u)$, niin $e_1(g) \geq e_1(g_u)$.*

Todistus. Olkoon 1_B joukon $B \subset \mathbb{R}^d$ karakteristinen funktio,

$$1_B(x) = \begin{cases} 1, & \text{kun } x \in B, \\ 0, & \text{muulloin.} \end{cases}$$



Kuva 3.6: Luokittimen toiminta-ominaiskäyrä.

Merkitään

$$A_j = \{x \in \mathbb{R}^d \mid g(x) = j\},$$
$$A_j(u) = \{x \in \mathbb{R}^d \mid g_u(x) = j\}.$$

Silloin

$$\begin{aligned}
e_1(g_u) - e_1(g) &= \int\limits_{A_2(u)} f_1(x)dx - \int\limits_{A_2} f_1(x)dx \\
&= 1 - \int\limits_{A_1(u)} f_1(x)dx - \left[1 - \int\limits_{A_1} f_1(x)dx\right] \\
&= \int\limits_{A_1} f_1(x)dx - \int\limits_{A_1(u)} f_1(x)dx \\
&= \int\limits_{\mathbb{R}^d} 1_{A_1}(x) f_1(x)dx - \int\limits_{\mathbb{R}^d} 1_{A_1(u)}(x) f_1(x)dx \\
&= \int\limits_{\mathbb{R}^d} \left[1_{A_1}(x) - 1_{A_1(u)}(x)\right] f_1(x)dx.
\end{aligned}$$

Tässä toinen yhtälö seurasi siitä, että

$$\mathbb{R}^d = A_1(u) \cup A_2(u), \qquad A_1(u) \cap A_2(u) = \emptyset$$

ja

$$1 = \int\limits_{\mathbb{R}^d} f_1(x)dx = \int\limits_{A_1(u)} f_1(x)dx + \int\limits_{A_2(u)} f_1(x)dx$$

sekä vastaavasti A_1, A_2:lle.

Jos $x \in A_1(u)$, on kaavan (3.13) nojalla $f_1(x) \geq u\, f_2(x)$. Siten

$$\left[1_{A_1}(x) - 1_{A_1(u)}(x)\right] f_1(x) \leq \left[1_{A_1}(x) - 1_{A_1(u)}(x)\right] u f_2(x).$$

Samoin, jos $x \in \mathbb{R}^d \setminus A_1(u) = A_2(u)$, on kaavan (3.13) nojalla $f_1(x) < u\, f_2(x)$. Siten

$$\left[1_{A_1}(x) - 1_{A_1(u)}(x)\right] f_1(x) \leq \left[1_{A_1}(x) - 1_{A_1(u)}(x)\right] u f_2(x).$$

Näin ollen saadaan

$$\begin{aligned} e_1(g_u) - e_1(g) &\leq \int\limits_{\mathbb{R}^d} \left[1_{A_1}(x) - 1_{A_1(u)}(x)\right] u f_2(x)dx \\ &= u\left[\int\limits_{A_1} f_2(x)dx - \int\limits_{A_1(u)} f_2(x)dx\right] \\ &= u\left[e_2(g) - e_2(g_u)\right]. \end{aligned}$$

Nyt oletuksen mukaan $e_2(g) \leq e_2(g_u)$ ja koska $u \geq 0$, on siis $e_1(g_u) - e_1(g) \leq 0$ eli

$$e_1(g) \geq e_1(g_u).$$

□

Jos erityisesti $e_2(g_u) = \alpha$, niin g_u minimoi virheen $e_1(g)$ kaikkien sellaisten luokittimien joukossa, joille $e_2(g) \leq \alpha$.

Esimerkki 3.16. Tarkastellaan jälleen lentokentän turvaporttia ja olkoon nyt $c = 2$,

$j = 1$: ei metallia
$j = 2$: metallia

Olettakaamme että halutaan esimerkiksi $e_2(g) \leq 0.01$, eli että metalliesinettä kuljettava matkustaja havaitaaan todennäköisyydellä ≥ 0.99. Nyt voidaan valita u siten, että $e_2(g_u) = 0.01$, jolloin g_u:lla on pienin "väärien hälytysten" todennäköisyys $e_1(g)$. ‖

Joskus u voidaan myös ratkaista yhtälöstä

$$e_2(u) = \int_{A_1(u)} f_2(x)dx = \alpha.$$

Käytännössä toimintaominaiskäyrä estimoidaan ja u määrätään siitä.

3.5 Teoriasta käytäntöön

Todellisuudessa emme tunne suureita P_j, f_j tarkasti, vaan ne on *estimoitava* sopivasta aineistosta. Olkoot $\widehat{P}_j$ ja $\hat{f}_j$ näiden suureiden estimaatit, $j = 1, \ldots, c$. Tällöin esimerkiksi likimääräinen Bayesin luokitin (vrt. (3.8)) saadaan kaavasta

$$g(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}} \widehat{P}_j \hat{f}_j(x).$$

Toinen mahdollisuus on estimoida posterioritodennäköisyydet $P(j|x)$ (vrt. (3.6)) jollain regressiotekniikalla ja asettaa

$$g(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}} \widehat{P}(j|x).$$

Kolmas suosittu tapa konstruoida luokitin on esittää luokka "prototyyppien" avulla ja luokitella x luokkaan, jonka prototyyppiä se on lähinnä (jossain mielessä).

Yleisemmin voidaan estimoida jotkin *erotinfunktiot* t_j, $j = 1, \ldots, c$, ja asettaa

$$g(x) = \operatorname*{argmax}_{j=1,\ldots,c} t_j(x).$$

Suureiden P_j, f_j, $P(j|\cdot)$ estimointiin, sekä erotinfunktioiden t_j ja prototyyppien määräämiseen käytetään *opetusaineistoa* eli *opetusdataa*, joka on joukko jollain tavalla saatuja valmiiksi luokiteltuja piirrevektoreita eli *opetusvektoreita*

$$(X_1, J_1), \ldots, (X_n, J_n),$$

missä J_i on X_i:n luokka.

Esimerkki 3.17. Turvaportin tapauksessa voitaisiin opetusaineisto saada periaatteessa kahdella eri tavalla.

(a) Tutkitaan manuaalisesti portin läpi menneet n satunnaista ihmistä pitäen kirjaa eri luokkien esiintymisistä.

(b) Järjestään koetilanne, jossa eri luokkia edustavia ihmisiä kävelytetään portin läpi.

‖

Tässä esimerkissä vaihtoehto (a) edustaa *sekoiteotantaa*, jossa muodostetaan satunnaisotos vektorin X jakaumasta ja jossa kuhunkin luokkaan j siten tulee satunnainen määrä n_j edustajia. Tällöin voidaan $\widehat{P}_j$ luontevasti estimoida suhtella n_j/n. Luokkatiehysfunktion estimaatti $\hat{f}_j$ puolestaan voidaan konstruoida niiden X_i:den perusteella, joilla $J_i = j$ eli jotka tulevat luokasta j.

Esimerkin vaihtoehto (b) edustaa *erillistä otantaa*. Siinä voidaan etukäteen päättää luokasta j kerättävien esimerkkien määrä n_j. Prioritodennäköisyydet

P_j oletetaan silloin tunnetuiksi tai käytetään luokitinta, joka ei tarvitse niitä. Luokkatiheysfunktiot estimoidaan kuten sekoiteotannassa (vaihtoehto (a)).

Turvaporttiesimerkissä (a) selvästikin toimii huonosti, erityisesti jos pyritään erottamaan kolme luokkaa. Selvästikin luokka $j = 2$ (suuri metallisesine) esiintyy harvoin ja siksi siihen liittyvät suureet olisi vaikea estimoida tällä menetelmällä. Helpompaa on käyttää erillistä otantaa, jota voidaan vielä tehostaa korvaamalla koehenkilöt ja metalliesineet niiden aiheuttamien sähkömagneettisten heijastusten tietokonesimuloinneilla. Luonteva luokitintyyppi olisi Neymanin-Pearsonin lähestymistapa, jossa tärkeimmän luokan ($j = 2$) kohdalla tehtävälle virheelle voidaan asettaa yläraja.

Luokkatiheysfunktion $\hat{f}_j$ estimoinnissa joudutaan usein estimoimaan esimerkiksi luokan j odotusarvo μ_j ja kovarianssimatriisi Σ_j. Tavallisesti tällöin otetaan

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji},$$
$$\widehat{\Sigma}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ji} - \hat{\mu}_j)(X_{ji} - \hat{\mu}_j)^T,$$

missä $X_{j1}, \ldots, X_{jn_j}$ ovat luokasta j peräisin olevat opetusaineiston piirrevektorit.

Tässä siis X_{ji}:t ovat satunnaisvektoreita, vaikka opetusvektorit todellisuudessa ovatkin tietysti kiinteitä numeerisia vektoreita, $x_{ji} = X_{ji}(\omega)$. Teorian kehittelyn kannalta on kuitenkin tarpeen pitää opetusvektoreita satunnaissuureina.

Jatkossa pyrimme kehittelemään luokittelijoita, jotka minimoivat nimenomaan virhetodennäköisyyden eli luokitteluvirheen kiinnittämättä huomiota eri virhetyyppien kustannuksiin tai hylkäysmahdollisuuteen. Tämä yksinkertaistaa käsittelyä jonkin verran mutta on kuitenkin syytä korostaa, että eri suuruisten kustannusten ja hylkäysmahdollisuuden käyttö voivat monessa

käytännön tilanteessa olla tärkeitä näkökohtia hahmontunnistusjärjestelmän suunnittelussa.

Luku 4

Parametrisia luokittimia

4.1 Kahden luokan tapaus

4.1.1 Normaalijakauman käyttö

Olkoon $X = [X^{(1)}, \ldots, X^{(d)}]^T$ d-ulotteinen piirrevektori. Luokan j *ehdollinen odotusarvo* on

$$\mu_j = \mathbb{E}(X|J = j)$$

ja *ehdollinen kovarianssimatriisi* vastaavasti

$$\Sigma_j = \mathbb{E}\left[(X - \mu_j)(X - \mu_j)^T | J = j\right],$$

$j = 1, 2$. Suureet μ_j ja Σ_j selvästi kuvaavat piirrevektorin X jakaumaa *luokan j sisällä*. Ne voidaan laskea luokkatiheysfunktion f_j avulla,

$$\mu_j = \left[\mu_j^{(1)}, \ldots, \mu_j^{(d)}\right], \quad \mu_j^{(i)} = \int_{\mathbb{R}^d} x^{(i)} f_j(x) dx,$$

$$\Sigma_j = [\sigma_{kl}^j], \quad \sigma_{kl}^j = \int_{\mathbb{R}^d} (x^{(k)} - \mu_j^{(k)})(x^{(l)} - \mu_j^{(l)}) f_j(x) dx,$$

ja integraalit tulevat vastaavasti yksi- ja kaksiulotteisiksi, kun ”ylimääräiset” muttujat integroidaan pois.

Oletetaan nyt, että

$$(X|J=j) \sim \mathrm{N}(\mu_j, \Sigma_j),$$

$j = 1, 2$, eli että luokkien jakaumat ovat multinormaalit. Siis,

$$f_j(x) = \frac{1}{(2\pi)^{d/2}(\det(\Sigma_j))^{1/2}}\, e^{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)}, \qquad x \in \mathbb{R}^d.$$

Olkoot luokkien prioritodennäköisyydet P_1 ja P_2. Bayesin luokitin (vrt. luku 3.2) saadaan nyt päätössäännöstä

$$s(x) = -\log\left[\frac{f_1(x)}{f_2(x)}\right] \underset{2}{\overset{1}{\lessgtr}} \log\left(\frac{P_1}{P_2}\right)$$

eli

$$-\log\left[\frac{(2\pi)^{d/2}(\det(\Sigma_2))^{1/2}}{(2\pi)^{d/2}(\det(\Sigma_1))^{1/2}}\, e^{\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}\right] \underset{2}{\overset{1}{\lessgtr}} \log\left(\frac{P_1}{P_2}\right)$$

mikä voidaan kirjoittaa muodossa

$$t(x) \underset{2}{\overset{1}{\lessgtr}} 0, \tag{4.1}$$

missä t on *erotinfunktio*,

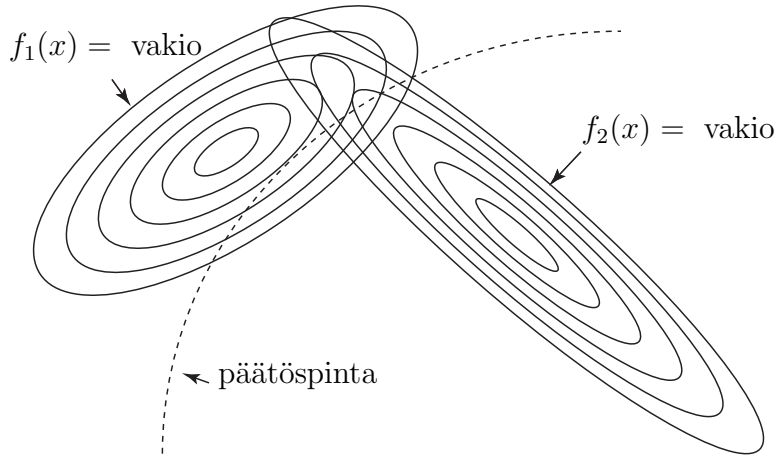
$$\begin{aligned} t(x) =& \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) \\ & - \frac{1}{2}\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) - \log\left(\frac{P_1}{P_2}\right). \end{aligned} \tag{4.2}$$

Päätöspintana on $\{x \in \mathbb{R}^d \mid t(x) = 0\}$, joka on $\mathbb{R}^d$:n toisen asteen pinta (kuva 4.1). Ehto (4.1) määrittelee *kvadraattisen luokittimen.*

Erikoistapauksessa $\Sigma_1 = \Sigma_2 = \Sigma$ saadaan *lineaarinen luokitin*, jolle erotinfunktiona on

$$t(x) = (\mu_2 - \mu_1)^T\Sigma^{-1}x + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2) - \log\left(\frac{P_1}{P_2}\right). \tag{4.3}$$

Sen päätöspinta $\{x \in \mathbb{R}^d \mid t(x) = 0\}$ on $\mathbb{R}^d$:n *(hyper)taso.*



Kuva 4.1: Kvadraattisen luokittimen luokkatiheysfunktiot tasa-arvopintoina esitettynä sekä luokittimen päätöspinta.

Esimerkki 4.1 (Korrelaatioluokitin)**.**

Tarkastellaan ajasta t riippuvaa signaalia $\mu(t)$, josta tehdyssä havainnossa $X(t)$ on satunnainen mittausvirhe $\varepsilon(t)$,

$$X(t) = \mu(t) + \varepsilon(t).$$

Signaali $\mu(t)$ voi olla esimerkiksi sähköjännite ja virhe $\varepsilon(t)$ mittauslaitteen aiheuttama häiriö ('kohinaa").

Tehdään mittaukset hetkinä $t_1 < t_2 < \cdots < t_d$ ja merkitään

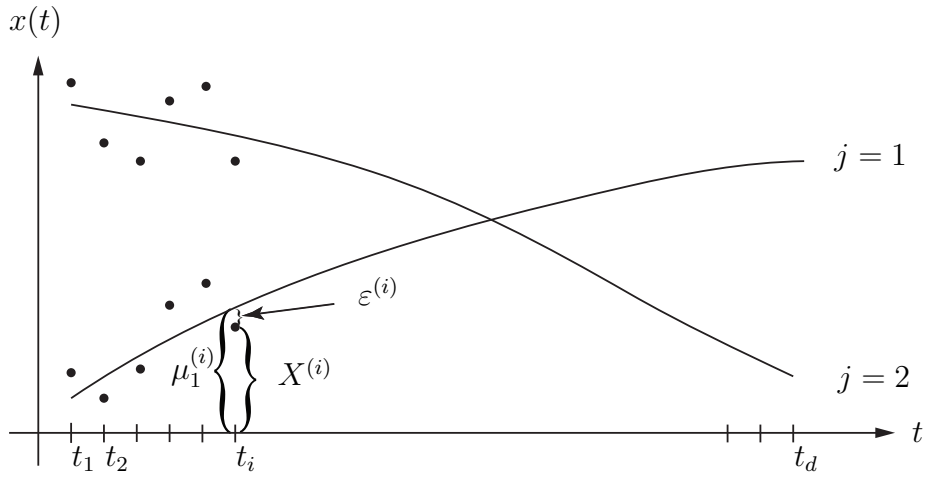
$$X^{(i)} = X(t_i), \qquad \mu^{(i)} = \mu(t_i), \qquad \varepsilon^{(i)} = \varepsilon(t_i), \qquad i = 1, \ldots, d.$$

Olkoon vielä

$$X = [X^{(1)}, \ldots, X^{(d)}]^T, \quad \mu = [\mu^{(1)}, \ldots, \mu^{(d)}]^T, \quad \varepsilon = [\varepsilon^{(1)}, \ldots, \varepsilon^{(d)}]^T,$$

jolloin

$$X = \mu + \varepsilon.$$



Kuva 4.2: Kaksi kohinaista signaalia.

Oletetaan nyt, että mittaus voi tulla kahdesta eri luokasta, eli

$$X = \mu_J + \varepsilon \tag{4.4}$$

missä $J = 1$ tai $J = 2$ ja

$$\mu_j = [\mu_j^{(1)}, \ldots, \mu_j^{(d)}]^T, \qquad j = 1, 2.$$

(vrt. kuva 4.2). Oletetaan vielä, että $\varepsilon \sim \mathrm{N}(0, \Sigma_\varepsilon)$, jollain Σ_ε (mahdollisesti korreloiva, normaalijakautunut mittausvirhe).

Silloin kaavan (4.4) nojalla luokan j jakauma on

$$(X|J = j) \sim \mathrm{N}(\mu_j, \Sigma_\varepsilon).$$

Suorittamalla tarvittaessa valkaisu voidaan olettaa, että $\Sigma_\varepsilon = I_d$. Tällöin lineaarinen luokitin (vertaa (4.3)) on muotoa

$$(\mu_2 - \mu_1)^T x + \frac{1}{2}\left(\mu_1^T \mu_1 - \mu_2^T \mu_2\right) - \log\left(\frac{P_1}{P_2}\right) \underset{2}{\overset{1}{\lessgtr}} 0 \tag{4.5}$$

eli

$$\mu_2^T x - \mu_1^T x \underset{2}{\overset{1}{\lessgtr}} \alpha,$$

missä

$$\alpha = \frac{1}{2}\big(\|\mu_2\|^2 - \|\mu_1\|^2 \big) + \log\big(\frac{P_1}{P_2}\big).$$

Tässä

$$\mu_j^T\, x = \sum_{i=1}^{d} \mu_j^{(i)} x^{(i)}$$

on mitatun signaalin "korrelaatio" luokan j virheettömän signaalin μ_j kanssa. Luokitin on siis optimaalinen Bayesin luokitin, jos oletukset ovat voimassa ja μ_j:t sekä P_j:t tunnetaan. ‖

Esimerkki 4.2 (Etäisyysluokitin)**.**

Olkoon

$$(X|J = j) \sim \mathrm{N}(\mu_j, \Sigma),$$

$j = 1, 2$. Suoritetaan valkaisu, jolloin voidaan olettaa, että $\Sigma = I_d$. Nyt siis Bayesin luokitin on (4.5). Yhtäpitävä sääntö saadaan kertomalla 2:lla,

$$2(\mu_2 - \mu_1)^T x + (\mu_1^T\mu_1 - \mu_2^T\mu_2) - 2\log\left(\frac{P_1}{P_2}\right) \underset{2}{\overset{1}{\lessgtr}} 0$$

eli, lisäämällä ja vähentämällä $x^T x$,

$$x^T x - 2\mu_1^T x + \mu_1^T\mu_1 - (x^T x - 2\mu_2^T x + \mu_2^T\mu_2) \underset{2}{\overset{1}{\lessgtr}} 2\log\left(\frac{P_1}{P_2}\right)$$

mikä voidaan kirjoittaa muotoon

$$(x - \mu_1)^T(x - \mu_1) - (x - \mu_2)^T(x - \mu_2) \underset{2}{\overset{1}{\lessgtr}} 2\log\left(\frac{P_1}{P_2}\right)$$

eli lopulta

$$\|x - \mu_1\|^2 - \|x - \mu_2\|^2 \underset{2}{\overset{1}{\lessgtr}} 2\log\left(\frac{P_1}{P_2}\right). \tag{4.6}$$

Tässä $\|x - \mu_j\|$ on x:n etäisyys luokan j "keskuksesta" (odotusarvosta) μ_j. Näin johdettu *etäisyysluokitin* on selvästi ekvivalentti korrelaatioluokittimen kanssa. ‖

Palataan sitten kvadraattiseen luokittimeen (4.1),

$$\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) \\ - \frac{1}{2}\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) - \log\left(\frac{P_1}{P_2}\right) \underset{2}{\overset{1}{\lessgtr}} 0.$$

Merkitään

$$\|x\|_{\Sigma_j^{-1}} = \sqrt{x^T\Sigma_j^{-1}x}, \qquad j = 1, 2.$$

Harjoituksen 3 tehtävässä 2 osoitetaan, että $\|\cdot\|_{\Sigma_j^{-1}}$ on $\mathbb{R}^d$:n normi. Kvadraattinen luokitin voidaan nyt kirjoittaa muotoon

$$\|x-\mu_1\|^2_{\Sigma_1^{-1}} - \|x-\mu_2\|^2_{\Sigma_2^{-1}} \underset{2}{\overset{1}{\lessgtr}} \log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) + 2\log\left(\frac{P_1}{P_2}\right). \tag{4.7}$$

Nähdään, että kyseessä on ”etäisyysluokitin”, kun ”etäisyyttä” mitataan normeilla $\|\cdot\|_{\Sigma_j^{-1}}$, $j = 1, 2$. Tällaista normia kutsutaan myös jakaumaan $\mathrm{N}(\mu_j, \Sigma_j)$ liityväksi *Mahalanobiksen etäisyydeksi*.

Olkoon sitten $P_1 = P_2$. Silloin tavallisen etäisyysluokittimen (4.6) päätöspinta on

$$\left\{x \in \mathbb{R}^d \mid \|x-\mu_1\| = \|x-\mu_2\|\right\}$$

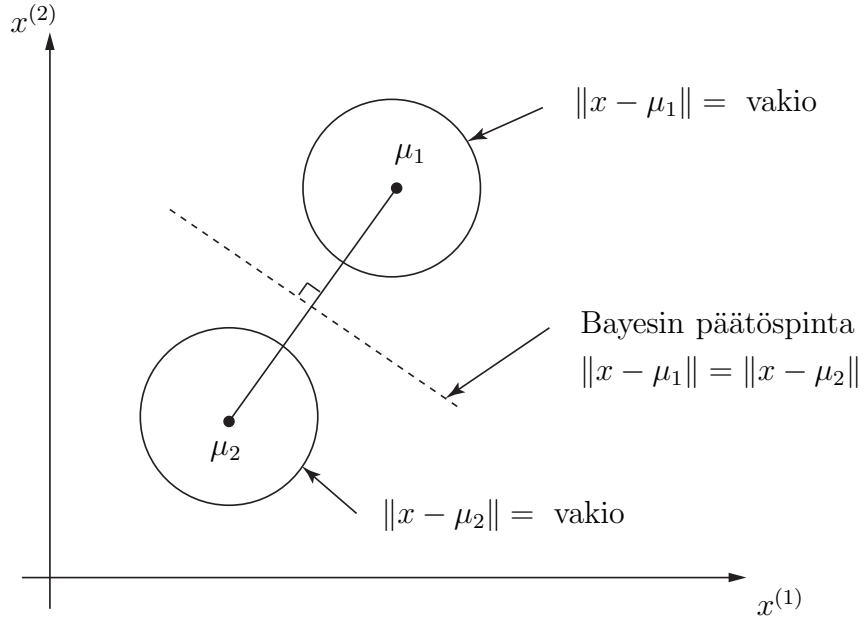
ja x luokitellaan siihen luokkaan, jonka keskiarvo μ_j on sitä lähempänä. Luokitinta on havainnollistettu kuvassa 4.3. On kuitenkin huomattava, että vaikka $\Sigma_1 = \Sigma_2 = \Sigma$, tämän säännön käyttö *ei* ole optimaalista, jos Σ ei ole muotoa $\sigma^2 I_d$ jollain $\sigma > 0$, koska optimaalinen päätöspinta on (ks. (4.7))

$$\left\{x \in \mathbb{R}^d \mid \|x-\mu_1\|_{\Sigma^{-1}} = \|x-\mu_2\|_{\Sigma^{-1}}\right\}.$$

Tilannetta valaisee kuva 4.4.

4.1.2 Yleinen lineaarinen luokitin

Lineaariseen luokittimeen päädyttiin edellä olettamalla multinormaalisuus ja $\Sigma_1 = \Sigma_2$. Vaikka nämä oletukset eivät toteudu, voidaan kuitenkin tietenkin



Kuva 4.3: Etäisyysluokitin kun $P_1 = P_2$. Se on optimaalinen kun luokkatiheysfunktiot ovat muotoa $\mathrm{N}(\mu_j, \sigma^2 I_d)$

.

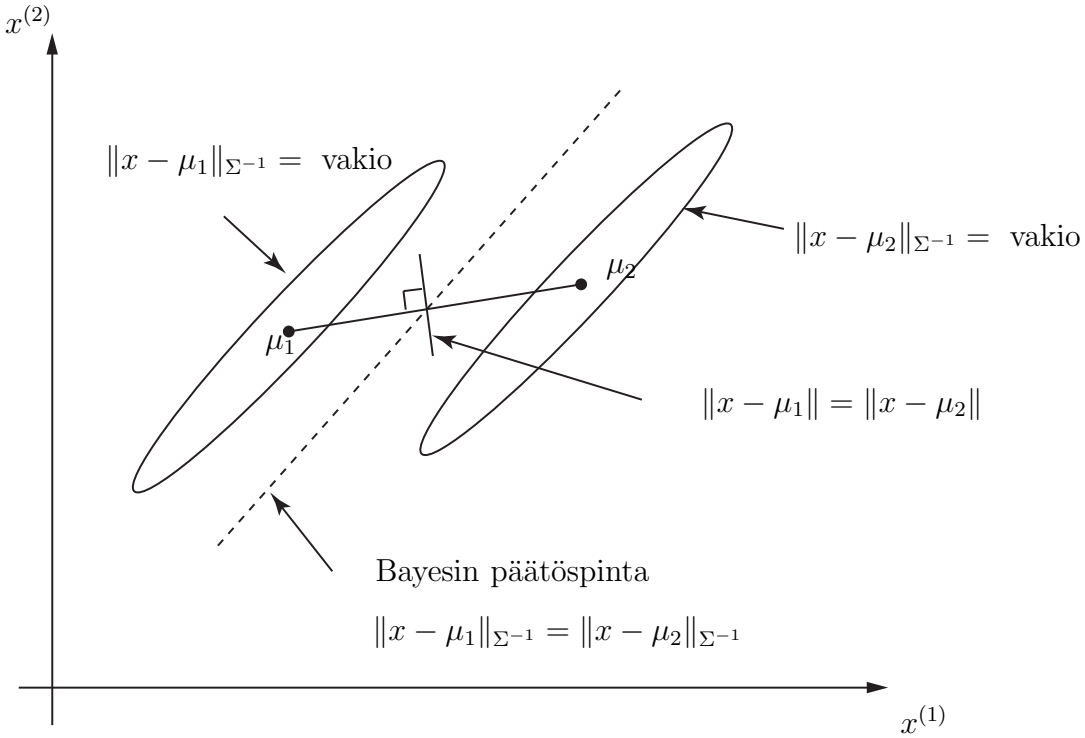
postuloida lineaarinen luokitin

$$t(x) = a^T x + \alpha \underset{2}{\overset{1}{\lessgtr}} 0, \tag{4.8}$$

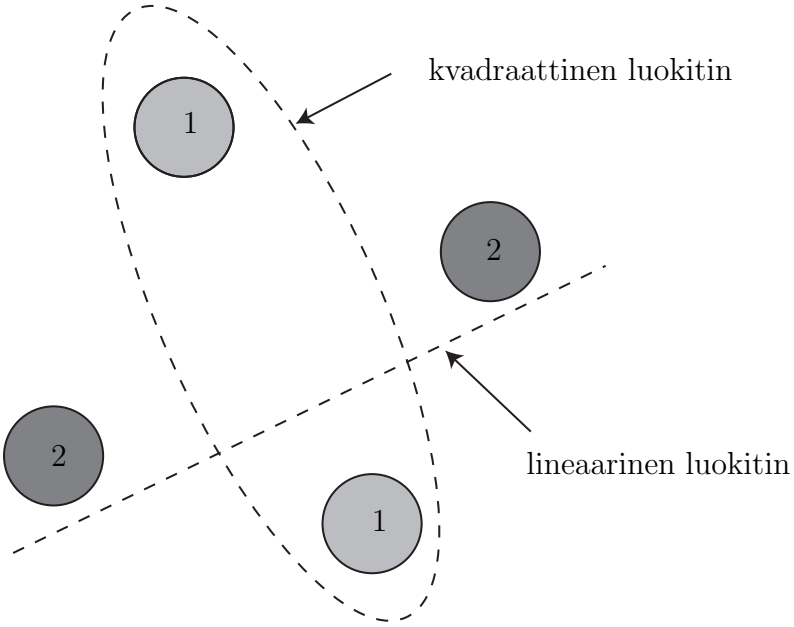
missä $a = [a^{(1)}, \ldots, a^{(d)}]^T$ ja $\alpha \in \mathbb{R}$ yritetään sitten valita optimaalisesti. Motivaationa tälle lähestymistavalle voi olla lineaarisen erotinfunktion t yksinkertaisuus, joka saattaa mahdollistaa tehokkaan luokittimen parametrien estimoinnin sekä kevyen laskennan paljon luokittelua vaativissa sovelluksissa. Idean toimivuudesta ei tietenkään kuitenkaan ole mitään takeita, jos luokat eivät separoidu hyvin hypertasolla (kuva 4.5).

Määritellään erotinfunktion (4.8) ehdollinen odotusarvo ja varianssi,

$$\begin{aligned}
\eta_j &= \mathbb{E}\big(t(X)|J=j\big) \\
\sigma_j^2 &= \mathbb{E}\big((t(X)-\eta_j)^2|J=j\big), \quad j=1,2.
\end{aligned}$$



Kuva 4.4: Mahalanobiksen etäisyyteen perustuva etäisyysluokitin kun $P_1 = P_2$ ja luokkatiheysfunktiot ovat muotoa $\mathrm{N}(\mu_j, \Sigma)$, jolloin se on optimaalinen Bayesin luokitin.



Kuva 4.5: Tilanne, jossa lineaarinen luokitin ei voi toimia optimaalisesti.

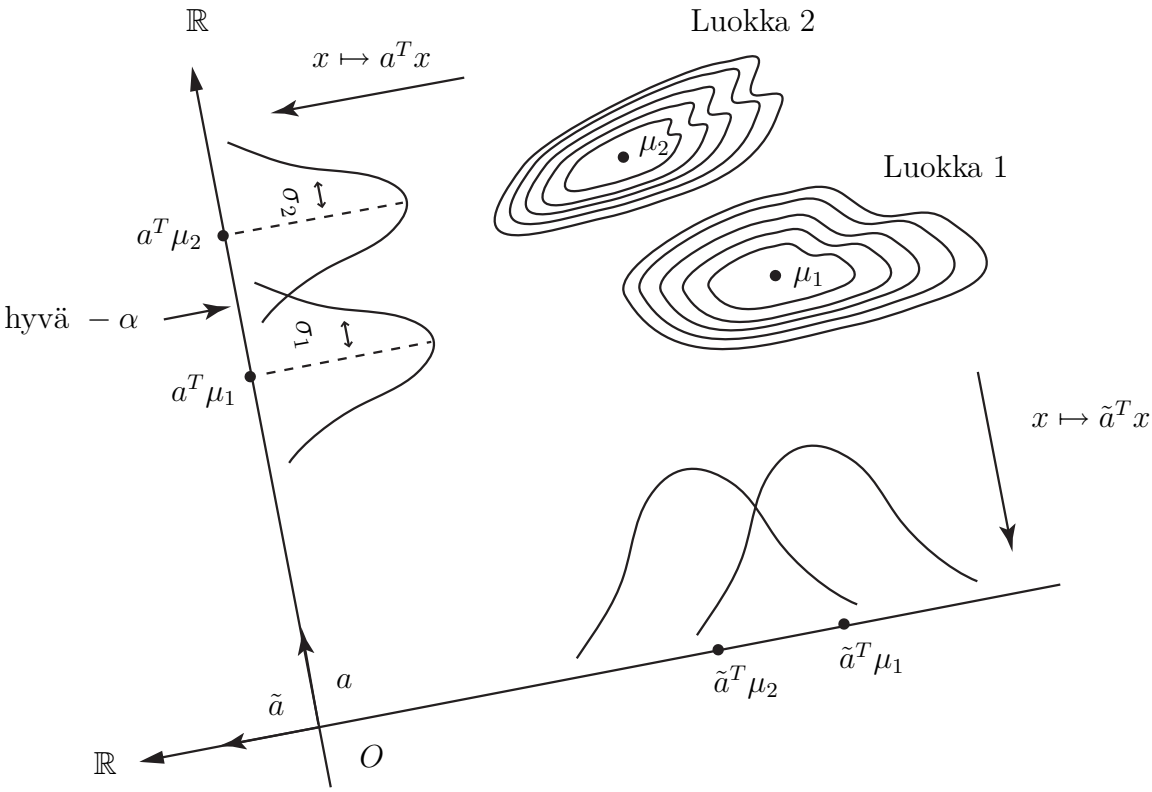
Siis,

$$\begin{aligned}\eta_j &= \mathbb{E}\left(a^T X + \alpha | J = j\right) \\ &= a^T \mathbb{E}(X|J=j) + \alpha \\ &= a^T \mu_j + \alpha\end{aligned}$$

ja

$$\begin{aligned}\sigma_j^2 &= \mathbb{E}\left[(a^T X + \alpha - a^T\mu_j - \alpha)^2 | J = j\right] \\ &= \mathbb{E}\left[\left(a^T(X-\mu_j)\right)^2 | J = j\right] \\ &= \mathbb{E}\left[a^T(X-\mu_j)\, a^T(X-\mu_j) | J = j\right] \\ &= \mathbb{E}\left[a^T(X-\mu_j)(X-\mu_j)^T a | J = j\right] \\ &= a^T\, \mathbb{E}\left[(X-\mu_j)(X-\mu_j)^T | J = j\right] a \\ &= a^T \Sigma_j\, a.\end{aligned}$$

Yllä siis μ_j ja Σ_j ovat X:n ehdollinen odotusarvo ja kovarianssi, jotka on määritelty jo aikaisemmin.



Kuva 4.6: Lineaarisen luokittimen $a^T x \underset{2}{\overset{1}{\lessgtr}} -\alpha$ optimointi. Kuvassa on oletettu, että $P_1 = P_2$ ja että $\|a\| = \|\tilde{a}\| = 1$. Jälkimmäinen ehto saadaan aina voimaan skaalaamalla erotinfunktio sopivasti. Kuvan tilanteessa a on selvästi parempi kuin $\tilde{a}$.

Nyt parametrien a ja α optimointi voidaan yrittää perustaa tunnuslukuihin η_1, η_2, σ_1^2 ja σ_2^2. Esimerkiksi voidaan yrittää asettaa tavoitteeksi suuri $|\eta_1 - \eta_2|$ ja pienet σ_1^2 ja σ_2^2. Tätä on havainnollisetettu kuvassa 4.6.

Maksimoidaan esimerkkinä *Fisherin kriteeri*

$$F(\eta_1, \eta_2, \sigma_1, \sigma_2) = \frac{(\eta_1 - \eta_2)^2}{\sigma_1^2 + \sigma_2^2}$$

parametrien a ja α suhteen. Tässä siis $\eta_j = a^T\mu_j + \alpha$ ja $\sigma_j^2 = a^T\Sigma_j\, a$, $j = 1, 2$,

joten tulee maksimoida

$$\begin{aligned} M(a,\alpha) &= F\big(\eta_1(a,\alpha),\ldots,\sigma_2(a,\alpha)\big) \\ &= \frac{(a^T\mu_1 + \alpha - a^T\mu_2 - \alpha)^2}{a^T\Sigma_1\, a + a^T\Sigma_2\, a} \\ &= \frac{\big(a^T(\mu_1-\mu_2)\big)^2}{a^T(\Sigma_1+\Sigma_2)a} \\ &= \frac{a^T\left[(\mu_1-\mu_2)(\mu_1-\mu_2)^T\right]a}{a^TCa}, \end{aligned}$$

missä $a \neq 0$ ja $C = \Sigma_1 + \Sigma_2$. Oletamme, että Σ_1 ja Σ_2 ovat positiivisesti definiittejä, jolloin C on positiivisesti definiitti ja $a^TCa > 0$. Lisäksi oletetamme, että $\mu_1 \neq \mu_2$.

$M(a,\alpha)$ ei riipu lainkaan α:sta ja voimme siis maksimoida funktion $a \mapsto M(a,\alpha) \equiv M(a)$, kun $a \in \mathbb{R}^d \setminus \{0\}$. Jos $\lambda \neq 0$, on

$$\begin{aligned} M(\lambda a) &= \frac{(\lambda a)^T\left[(\mu_1-\mu_2)(\mu_1-\mu_2)^T\right](\lambda a)}{(\lambda a)^TC(\lambda a)} \\ &= M(a), \end{aligned}$$

joten M saa vakioarvon origon kautta kulkevilla suorilla. Siksi optimoinnissa voitaisiin rajoittua vaikka kompaktiin joukkoon

$$S = \{a \in \mathbb{R}^d \mid a^TCa = 1\},$$

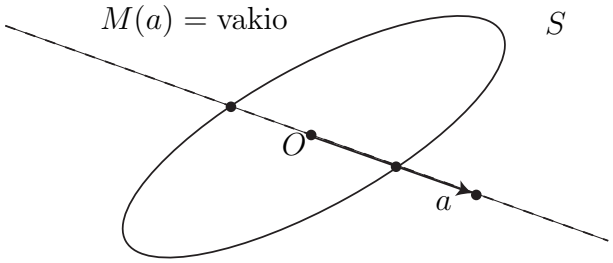
joka on origokeskinen ellipsoidi (kuva 4.7). Globaali maksimi on siten varmasti olemassa, koska jatkuva funktio M kompaktissa joukossa C saavuttaa siellä suurimman arvonsa.

Ehto maksimoivalle a:lle on

$$\nabla_a M(a) = 0$$

eli

$$\nabla_a F\big(\eta_1(a,\alpha),\ldots,\sigma_2(a,\alpha)\big) = 0$$



Kuva 4.7: Fisherin kriteerin optimointi.

mikä ketjusäännön nojalla voidaan kirjoittaa muotoon

$$\frac{\partial F}{\partial \eta_1}\nabla_a \eta_1(a,\alpha)+\frac{\partial F}{\partial \eta_2}\nabla_a \eta_2(a,\alpha)+\frac{\partial F}{\partial \sigma_1}\nabla_a \sigma_1(a,\alpha)+\frac{\partial F}{\partial \sigma_2}\nabla_a \sigma_2(a,\alpha) = 0. \quad (4.9)$$

Tässä

$$\frac{\partial F}{\partial \eta_1} = \frac{\partial F\big(\eta_1(a,\alpha),\ldots,\sigma_2(a,\alpha)\big)}{\partial \eta_1}$$

ja samoin muille osittaisderivaatoille. Edelleen,

$$\frac{\partial F}{\partial \eta_1} = \frac{\partial}{\partial \eta_1}\frac{(\eta_1-\eta_2)^2}{\sigma_1^2+\sigma_2^2} = \frac{2(\eta_1-\eta_2)}{\sigma_1^2+\sigma_2^2} = 2\frac{a^T(\mu_1-\mu_2)}{a^TCa},$$

$$\frac{\partial F}{\partial \eta_2} = \cdots = -\frac{2(\eta_1-\eta_2)}{\sigma_1^2+\sigma_2^2} = -2\frac{a^T(\mu_1-\mu_2)}{a^TCa},$$

$$\frac{\partial F}{\partial \sigma_1} = \frac{\partial}{\partial \sigma_1}\frac{(\eta_1-\eta_2)^2}{\sigma_1^2+\sigma_2^2} = \frac{(\eta_1-\eta_2)^2(-1)\cdot 2\sigma_1}{(\sigma_1^2+\sigma_2^2)^2} = -2\frac{[a^T(\mu_1-\mu_2)]^2\sqrt{a^T\Sigma_1 a}}{(a^TCa)^2},$$

$$\frac{\partial F}{\partial \sigma_2} = \cdots = \frac{(\eta_1-\eta_2)^2(-1)\cdot 2\sigma_2}{(\sigma_1^2+\sigma_2^2)^2} = -2\frac{[a^T(\mu_1-\mu_2)]^2\sqrt{a^T\Sigma_2 a}}{(a^TCa)^2},$$

$$\nabla_a \eta_j(a,\alpha) = \nabla_a(a^T\mu_j+\alpha) = \mu_j,$$

ja

$$\begin{aligned}\nabla_a \sigma_j(a,\alpha) &= \nabla_a\sqrt{a^T\Sigma_j\, a}\\ &= \frac{1}{2}\cdot\frac{1}{\sqrt{a^T\Sigma_j\, a}}\nabla_a(a^T\Sigma_j\, a)\\ &= \frac{1}{2}\cdot\frac{1}{\sqrt{a^T\Sigma_j\, a}}\cdot 2\Sigma_j\, a = \frac{\Sigma_j\, a}{\sqrt{a^T\Sigma_j\, a}}.\end{aligned}$$

Sijoittamalla nämä lausekkeet kaavaan (4.9) saadaan eräiden vaiheiden jälkeen yhtälö

$$\left[\frac{a^T(\mu_1-\mu_2)}{a^TCa}\Sigma_1+\frac{a^T(\mu_1-\mu_2)}{a^TCa}\Sigma_2\right]a=\mu_1-\mu_2. \tag{4.10}$$

Edelleen, (4.10):n nojalla

$$\left(\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2\right)\left(-\frac{2a^T(\mu_1-\mu_2)}{a^TCa}\right)a=\mu_2-\mu_1.$$

Koska a:n ollessa maksimoija on myös λa sitä, kun $\lambda\neq 0$, todetaan valitsemalla

$$\lambda=-\frac{2a^T(\mu_1-\mu_2)}{a^TCa},$$

että löytyy myös maksimoija a, jolle

$$\left(\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2\right)a=\mu_2-\mu_1. \tag{4.11}$$

Huomaa, että $\lambda\neq 0$, koska muuten $M(a)=0$ eikä a voisi olla maksimoija. Nyt siis (4.11):n nojalla

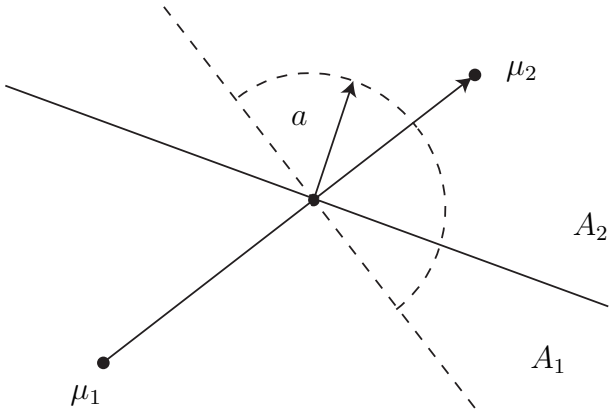
$$a=\left(\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2\right)^{-1}(\mu_2-\mu_1). \tag{4.12}$$

Näin saatava lineaarinen luokitin on *Fisherin lineaarinen luokitin*. Parametri α on määrättävä kokeellisesti.

Huomautus. Koska matriisi $\left(\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2\right)^{-1}$ on positiivisesti definiitti, on kaavan (4.12) nojalla $(\mu_2-\mu_1)^Ta>0$, mikä on sopusoinnussa sen kanssa, että sääntö $a^Tx+\alpha\underset{2}{\overset{1}{\lessgtr}}0$ luokittelee luokan 2 positiivisen a:n puolelle (kuva 4.8). Periaatteessa $-a$ olisi tietysti myös optimi mutta säännön $a^Tx+\alpha\underset{2}{\overset{1}{\lessgtr}}0$ mukainen oikea valinta on juuri $a=(\frac{1}{2}\Sigma_1+\frac{1}{2}\Sigma_2)^{-1}(\mu_2-\mu_1)$.

Jos itseasiassa $(X|J=j)\sim \mathrm{N}(\mu_j,\Sigma_j)$, $j=1,2$, voidaan osoittaa, että

$$\big(t(X)|J=j\big)\sim \mathrm{N}(\eta_j,\sigma_j^2),$$



Kuva 4.8: Fisherin luokittimen toiminta.

missä η_j ja σ_j^2 määriteltiin edellä. Silloin voidaan yrittää minimoida itse luokitteluvirhettä

$$e(\eta_1, \eta_2, \sigma_1, \sigma_2) = P_1 \int_0^\infty f_{t(X)}(u|J=1)du + P_2 \int_{-\infty}^0 f_{t(X)}(u|J=2)du,$$

missä siis $f_{t(X)}(\,\cdot|J=j)$ on $t(X)$:n ehdollinen tiheysfunktio ja t määriteltiin kaavassa (4.8). Osoittautuu, että tällöin optimaalisuusehto on muotoa

$$\big(w\Sigma_1 + (1-w)\Sigma_2\big)a = \mu_2 - \mu_1,$$

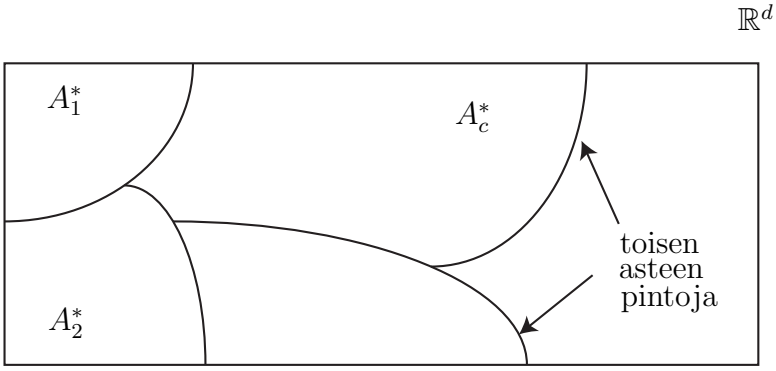
missä $w \in]0,1[$. Nyt paras w ja α on käytännössä määrättävä kokeellisesti. Sen jälkeen paras a saadaan tämän yhtälön ratkaisuna kun w:lle on löydetty optimaalinen arvo.

4.2 Monen luokan tapaus

Olkoon luokkien lukumäärä $c > 2$. Oletetaan lisäksi , että

$$(X|J=j) \sim \mathrm{N}(\mu_j, \Sigma_j),$$

$j = 1, \ldots, c$.



Kuva 4.9: Kvadraattisen luokittimen päätöspinnat ovat paloittain kvadraattisia.

Aikaisemman mukaan Bayesin luokitin $g^* : \mathbb{R}^d \to \{1, \ldots, c\}$ voidaan määritellä kaavalla

$$g^*(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}} P_j f_j(x),$$

missä P_j on luokan j prioritodennäköisyys ja f_j sen luokkatiheysfunktio. Edelleen, kuten on nähty, yhtä hyvin voidaan maksimoida

$$\begin{aligned}
&\log[P_j\, f_j(x)] \\
&\quad = \log P_j + \log \left[\frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_j)}}\, e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)} \right] \\
&\quad = \log P_j - \log(2\pi)^{d/2} - \log\sqrt{\det(\Sigma_j)} - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j).
\end{aligned}$$

Pudottamalla pois $-\log(2\pi)^{d/2}$, joka ei riipu j:stä nähdään, että

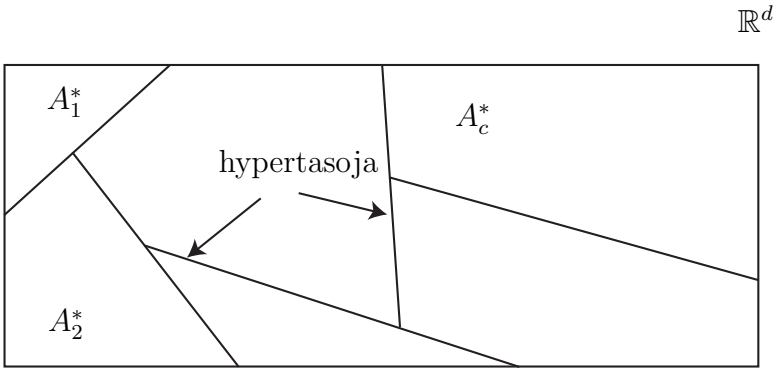
$$g^*(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}}\, t_j(x),$$

missä erotinfunktio t_j on

$$t_j(x) = -\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) - \frac{1}{2}\log\big(\det(\Sigma_j)\big) + \log P_j.$$

Saatu luokitin on *paloittain kvadraattinen luokitin*. Sen päätösalueita

$$A_j^* = \{x \in \mathbb{R}^d \mid g^*(x) = j\}$$



Kuva 4.10: Lineaarisen luokittimen päätöspinnat ovat paloittain lineaarisia, eli ne koostuvat hypertasojen osista.

rajaavat päätöspinnat ovat paloittain kvadraattisia (kuva 4.9).

Jos erityisesti $\Sigma_j = \Sigma$ kaikilla $j = 1, \ldots, c$, voi erotinfunktiosta t_j jättää pois termin $-\frac{1}{2}\log\big(\det(\Sigma_j)\big) = -\frac{1}{2}\log\big(\det(\Sigma)\big)$ sekä kvadraattisen termin

$$-\frac{1}{2}x^T\Sigma_j^{-1}x = -\frac{1}{2}x^T\Sigma\, x, \qquad j = 1, \ldots, c$$

ilman, että g^* muuttuu. Silloin saadaan erotinfunktioksi

$$\begin{aligned} t_j(x) &= \frac{1}{2}x^T\Sigma^{-1}\mu_j + \frac{1}{2}\mu_j^T\Sigma^{-1}x - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log P_j \\ &= \mu_j^T\Sigma^{-1}x - \frac{1}{2}\,\mu_j^T\Sigma^{-1}\mu_j + \log P_j. \end{aligned}$$

Saatu luokitin on *paloittain lineaarinen luokitin*. Sen päätöspinnat ovat paloittain lineaarisia (kuva 4.10).

Huomautus. Käytännössä kaikki tämän luvun normaalisuuteen perustuvat luokittimet vaativat suureiden P_j, μ_j ja Σ_j estimointia. Tätä on käsitelty lyhyesti luvussa 3.5. Todelliset sovellettavat lineaariset ja kvadraattiset luokittelijat saadaan korvaamalla edellä esitetyissä kaavoissa nämä suureet estimaateillaan $\widehat{P}_j$, $\hat{\mu}_j$ ja $\widehat{\Sigma}_j$. Samat luokittimimet voitaisiin johtaa myös ns. parametrisen tiheysfunktion estimoinnin kautta (ks. luku 6.1).

Luku 5

Regression käyttöön perustuvia parametrisia luokittimia

5.1 Regression teoriaa – kahden luokan tapaus

Tarkastellaan erotinfunktiota $t(\,\cdot\,;\theta) : \mathbb{R}^d \to \mathbb{R}$, missä $\theta \in \mathbb{R}^p$ on parametri(vektori).

Esimerkki 5.1.

(a) $t(x;\theta) = a^T x + \alpha$, missä $\theta = [a^T, \alpha]^T = [a^{(1)}, \ldots, a^{(d)}, \alpha]^T$. Tämä erotinfunktiotyyppi määrittelee yleisen lineaarisen luokittimen.

(b) $t(x;\theta) = x^T A x + a^T x + \alpha$, missä $\theta = [a_{11}, a_{12}, \ldots, a_{dd}, a^{(1)}, \ldots, a^{(d)}, \alpha]^T$, $A = [a_{ij}]$ ja $a = [a^{(1)}, \ldots, a^{(d)}]^T$. Erotinfunktio määrittelee nyt yleisen kvadraattista tyyppiä olevan luokittimen,

Idea on nyt toimia seuraavasti:

1. Valitaan sellainen θ, että mahdollisimman tarkkaan pätee

$$t(x;\theta) < 0, \text{ kun } x \text{ on luokasta 1,}$$
$$t(x;\theta) > 0, \text{ kun } x \text{ on luokasta 2.}$$

2. Otetaan sitten luokittimeksi

$$t(x,\theta) \underset{2}{\overset{1}{\lessgtr}} 0. \tag{5.1}$$

Eräs (heuristinen) tapa parametrin θ valitsemiseksi on yrittää määrätä se siten, että

$$t(x;\theta) \approx -1, \text{ kun } x \text{ on luokasta 1}$$
$$t(x;\theta) \approx 1, \text{ kun } x \text{ on luokasta 2.}$$

Tämä voidaan tehdä *regression* avulla.

Määritellään ensin $\gamma : \{1,2\} \to \{-1,1\}$,

$$\gamma(j) = \begin{cases} -1,\ j = 1, \text{ (luokka 1)} \\ +1,\ j = 2. \text{ (luokka 2)} \end{cases}$$

Nyt voidaan yrittää minimoida θ:n suhteen "keskimääräinen neliöllinen virhe"

$$R(\theta) = \mathbb{E}\left[\big(t(X;\theta) - \gamma(J)\big)^2\right]. \tag{5.2}$$

Näytämme seuraavaksi, että tätä heuristista ideaa voi itseasiassa perustella Bayesin luokittimen avulla. Perustelussa tarvitsemme seuraavaa tulosta: jos $G(X,J)$ on X:n ja J:n funktio, niin odotusarvo $\mathbb{E}[G(X,J)]$ voidaan laskea kahdessa vaiheessa, *iteroidusti*,

$$\mathbb{E}\left[G(X,J)\right] = \mathbb{E}_X\big[\mathbb{E}\big[G(X,J)|X\big]\big], \tag{5.3}$$

missä $\mathbb{E}_X$ on odotusarvo X:n suhteen ja $\mathbb{E}\big[G(X,J)|X\big]$ on odotusarvo ehdolla X, eli X:n ollessa kiinnitetty ja J:n varioidessa.

Tämä tulos on helppo osoittaa kun X ja J ovat kummatkin diskreettejä tai kummatkin ovat jatkuvia (esimerkiksi jatkuva tapaus mainitaan kirjan Tuominen: Todennäköisyyslaskenta I, huomautuksessa 5.3.12). Se pätee kuitenkin myös nyt, kun X on jatkuva ja J diskreetti – perustelua ei kuitenkaan esitetä tässä.

Saadaan

$$\begin{aligned}
R(\theta) &= \mathbb{E}\left[\big(t(X;\theta) - \gamma(J)\big)^2\right] \\
&= \mathbb{E}\left[\big(t(X;\theta) - \mathbb{E}(\gamma(J)|X) + \mathbb{E}(\gamma(J)|X) - \gamma(J)\big)^2\right] \\
&= \mathbb{E}\left[\big(t(X;\theta) - \mathbb{E}(\gamma(J)|X)\big)^2\right] + \mathbb{E}\left[\big(\mathbb{E}(\gamma(J)|X) - \gamma(J)\big)^2\right] \\
&\quad + 2\,\mathbb{E}\big\{\,[t(X;\theta) - \mathbb{E}(\gamma(J)|X)] \cdot [\mathbb{E}(\gamma(J)|X) - \gamma(J)]\,\big\}.
\end{aligned}$$

Tässä viimeinen termi häviää,

$$\begin{aligned}
&\mathbb{E}\big\{\,[t(X;\theta) - \mathbb{E}(\gamma(J)|X)] \cdot [\mathbb{E}(\gamma(J)|X) - \gamma(J)]\,\big\} \\
&\quad = \mathbb{E}_X\Big\{\mathbb{E}\Big\{\,[t(X;\theta) - \mathbb{E}(\gamma(J)|X)] \cdot [\mathbb{E}(\gamma(J)|X) - \gamma(J)]\,\big|X\Big\}\Big\} \\
&\quad = \mathbb{E}_X\big\{\,[t(X;\theta) - \mathbb{E}(\gamma(J)|X)]\ \mathbb{E}\,[\mathbb{E}(\gamma(J)|X) - \gamma(J)|X]\,\big\} \\
&\quad = \mathbb{E}_X\big\{\,[t(X;\theta) - \mathbb{E}(\gamma(J)|X)]\ \big[\mathbb{E}(\gamma(J)|X) - \mathbb{E}[\gamma(J)|X\big]\,\big\} \\
&\quad = \mathbb{E}_X\big\{t(X;\theta) - \mathbb{E}(\gamma(J)|X)\big\} \cdot 0 \\
&\quad = 0,
\end{aligned}$$

missä ensimmäinen yhtäsuuruus seuraa kaavasta (5.3). Keskimmäinen termi puolestaan ei riipu θ:sta, joten $R(\theta)$ minimoituu, kun

$$\mathbb{E}\big[\big(t(X;\theta) - \mathbb{E}(\gamma(J)|X)\big)^2\big]$$

minimoituu eli kun

$$\int_{\mathbb{R}^d} \big[t(x;\theta) - \mathbb{E}(\gamma(J)|X = x)\big]^2 f(x)dx \tag{5.4}$$

minimoituu, missä f on X:n tiheysfunktio. Tässä

$$\begin{aligned}\mathbb{E}\big(\gamma(J)|X=x\big) &= \gamma(1)\,P(1|x) + \gamma(2)\,P(2|x)\\ &= -P(1|x) + P(2|x).\end{aligned}$$

Näin ollen θ tulee valittua siten, että

$$t(x;\theta) \approx P(2|x) - P(1|x) \tag{5.5}$$

ja luokitin (5.1) on likimain

$$P(2|x) - P(1|x) \underset{2}{\overset{1}{\lessgtr}} 0$$

eli toisin sanoen

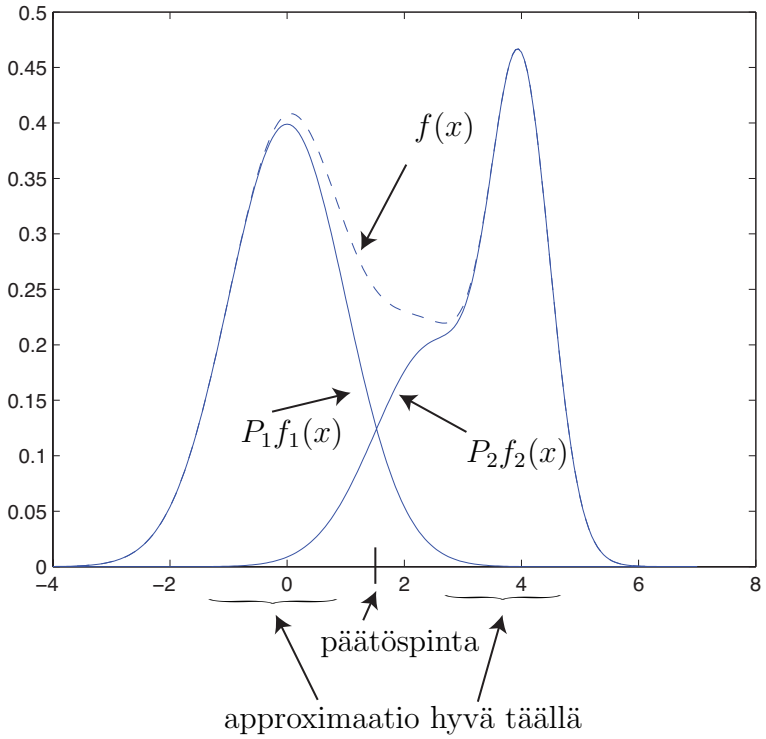
$$P(1|x) \underset{2}{\overset{1}{\gtrless}} P(2|x)$$

mikä on optimaalinen Bayesin luokitin. Siten $R(\theta)$:n minimointi on kuin onkin melko hyvin perusteltu ajatus.

Huonona puolena edellä esitetyssä lähestymistavassa on se, että (5.4):n nojalla approksimaation (5.5) tarkkuus tulee olla hyvä erityisesti silloin, kun $f(x)$ on suuri, tyypillisesti luokkien jakaumien keskellä. Kuitenkin Bayes-päätössäännön approksimaation *pitäisi* olla hyvä erityisesti *päätöspintojen* lähellä, ja siellä $f(x)$ taas usein on pieni (kuva 5.1).

5.2 Lineaarinen luokitin

Olkoon nyt $t(x;\theta) = a^T x + \alpha$, $\theta = [a^T, \alpha]^T$. Haetaan θ, joka minimoi suureen $R(\theta)$ (kaava (5.2)) eli minimoidaan (5.4) θ:n suhteen. Lauseke (5.4) voidaan



Kuva 5.1: Keskimääräisen neliöllisen virheen minimointiin perustuva luokitin approksimoi luokkien posterioritodennäköisyyksiä parhaiten siellä, missä luokkien tiheysfunktiot ovat suurimmillaan eikä välttämättä optimaalisen päätöspinnan lähellä, mikä olisi toivottavampaa.

kirjoittaa muodossa

$$\begin{aligned}
&\int_{\mathbb{R}^d} \big(t(x;\theta) + P(1|x) - P(2|x)\big)^2 f(x)dx \\
&= \int_{\mathbb{R}^d} t(x;\theta)^2 f(x)dx + 2\int_{\mathbb{R}^d} t(x;\theta)\big(P(1|x) - P(2|x)\big) f(x)dx \\
&\quad + \int_{\mathbb{R}^d} \big(P(1|x) - P(2|x)\big)^2 f(x)dx.
\end{aligned}$$

Tässä viimeinen termi ei riipu θ:sta ja voidaan siis unohtaa. Edelleen,

$$\begin{aligned}
\int_{\mathbb{R}^d} t(x;\theta)^2 f(x)dx &= \int_{\mathbb{R}^d} t(x;\theta)^2 [P_1 f_1(x) + P_2 f_2(x)]dx \\
&= P_1 \int_{\mathbb{R}^d} t(x;\theta)^2 f_1(x)dx + P_2 \int_{\mathbb{R}^d} t(x;\theta)^2 f_2(x)dx \\
&= P_1 \mathbb{E}\big[t(X;\theta)^2 | J = 1\big] + P_2 \mathbb{E}\big[t(X;\theta)^2 | J = 2\big] \\
&= P_1(\sigma_1^2 + \eta_1^2) + P_2(\sigma_2^2 + \eta_2^2),
\end{aligned}$$

missä η_j ja σ_j^2 ovat $t(X;\theta)$:n ehdollinen odotusarvo ja varianssi luokassa j (ks. luku 4.1.2). Tässä hyödynnettiin mielivaltaiselle satunnaismuuttujalle Y tunnetusti pätevää tulosta $\mathbb{E}(Y^2) = D^2(Y) + \mathbb{E}(Y)^2$ ja sitä, että tämä hajotelma on voimassa myös ehdolliselle odotusarvolle ja varianssille.

Toiseksi,

$$\begin{aligned}
&\int_{\mathbb{R}^d} t(x;\theta)\big(P(1|x) - P(2|x)\big) f(x)dx \\
&\qquad = \int_{\mathbb{R}^d} t(x;\theta)\big(P_1 f_1(x) - P_2 f_2(x)\big)dx \\
&\qquad = P_1 \int_{\mathbb{R}^d} t(x;\theta) f_1(x)dx - P_2 \int_{\mathbb{R}^d} t(x;\theta) f_2(x)dx \\
&\qquad = P_1 \mathbb{E}\big(t(X;\theta)|J = 1\big) - P_2 \mathbb{E}\big(t(X;\theta)|J = 2\big) \\
&\qquad = P_1\eta_1 - P_2\eta_2.
\end{aligned}$$

Ottaen huomioon, että $\eta_j = a^T\mu_j + \alpha$ ja $\sigma_j^2 = a^T\Sigma_j\, a$, on siis minimoitava muuttujien a ja α suhteen funktio

$$\begin{aligned} M(\theta) &= M(a,\alpha) \\ &= P_1\big(a^T\Sigma_1\, a + (a^T\mu_1+\alpha)^2\big) + P_2\big(a^T\Sigma_2\, a + (a^T\mu_2+\alpha)^2\big) \\ &\quad + 2\big[P_1(a^T\mu_1+\alpha) - P_2(a^T\mu_2+\alpha)\big]. \end{aligned}$$

Pienten laskujen jälkeen saadaan, että

$$M(\theta) = (\theta-\bar\theta)^T B(\theta-\bar\theta) - \bar\theta^T B\bar\theta,$$

missä B on positiivisesti definiitti,

$$B = \begin{bmatrix} S & \mu \\ \mu^T & 1 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)},$$

$$S = \mathbb{E}(XX^T), \qquad \mu = \mathbb{E}(X) = P_1\mu_1 + P_2\mu_2,$$

$$\bar\theta = -B^{-1}c, \qquad c = \begin{bmatrix} P_1\mu_1 - P_2\mu_2 \\ P_1 - P_2 \end{bmatrix} \in \mathbb{R}^{d+1}.$$

Koska siis B on positiivisesti definiitti, on M:llä on yksikäsitteinen globaali minimi kohdassa $\theta = \bar\theta = -B^{-1}c \equiv [\bar a^T, \bar\alpha]^T$, eli kun $a = \bar a$ ja $\alpha = \bar\alpha$. Regressiosta saatava luokitin on siis

$$\bar a^T x + \bar\alpha \underset{2}{\overset{1}{\lessgtr}} 0. \tag{5.6}$$

Voidaan osoittaa, että jos $P_1 = P_2$ ja

$$(X|J=j) \sim \mathrm{N}(\mu_j, \Sigma), \qquad (j=1,2),$$

on saatu luokitin itseasiassa Bayesin luokitin.

Yleisesti voidaan myös osoittaa, että jos $\bar{g}$ on luokitin (5.6), pätee sen virhetodennäköisyydelle

$$e(\bar{g}) \leq \frac{R(\bar{\theta})}{1 - R(\bar{\theta})}$$

(ks. P. A. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach, Prentice-Hall, Inc., 1982, s.160).

Lineaarinen luokitin käytännössä

Todellisuudessa meillä on käytössä vain luokitellut opetusvektorit (X_1, J_1), $\ldots, (X_n, J_n)$ ja $R(\theta)$ on korvattava estimaatilla

$$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \big(t(X_i; \theta) - \gamma(J_i)\big)^2, \tag{5.7}$$

joka minimoidaan sitten parametrin $\theta = [a^T, \alpha]^T$ suhteen.

Merkitään

$$D_n = \begin{bmatrix} X_1^T & 1 \\ \vdots & \vdots \\ X_n^T & 1 \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}, \quad \Gamma = \begin{bmatrix} \gamma(J_1) \\ \vdots \\ \gamma(J_n) \end{bmatrix} \in \{-1, 1\}^n.$$

Tällöin

$$t(X_i; \theta) = a^T X_i + \alpha = [X_i^T \ 1] \begin{bmatrix} a \\ \alpha \end{bmatrix},$$

ja

$$\widehat{R}(\theta) = \frac{1}{n} \left\| D_n \, \theta - \Gamma \right\|^2.$$

Siten $\widehat{R}(\theta)$:n minimoi yhtälöryhmän $D_n \, \theta = \Gamma$ yleistetty ratkaisu,

$$\tilde{\theta} = D_n^+ \, \Gamma.$$

Huomautus. (5.7) on sopiva $R(\theta)$:n estimaatti nimenomaan sekoiteotannan tilanteessa. Erillisessä otannassa pitää käyttää estimaattia

$$\widehat{R}(\theta) = \sum_{j=1}^{2} P_j \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} \big(t(X_{ji}; \theta) - \gamma(j)\big)^2,$$

kun luokasta j on käytössä opetusvektorit $X_{j1}, \ldots X_{jn_j}$ $(j = 1, 2)$.

5.3 Monen luokan tapaus

Jos luokkia on $c > 2$ kappaletta, voidaan kullekin luokalle määrätä oma erotinfunktio $t_j(\,\cdot\,;\theta)$ ja θ siten, että

$$t_j(x;\theta) \approx \begin{cases} 1, & \text{kun } x \text{ on luokasta } j, \\ 0, & \text{kun } x \text{ ei ole luokasta } j. \end{cases} \tag{5.8}$$

Luokitin $g : \mathbb{R}^d \to \{1, \ldots, c\}$ määritellään sitten kaavalla

$$g(x) = \operatorname*{argmax}_{j=1,\ldots,c} t_j(x, \theta). \tag{5.9}$$

Olkoon $t(\,\cdot\,;\theta) : \mathbb{R}^d \to \mathbb{R}^c$ kuvaus

$$t(x;\theta) = \big[t_1(x;\theta), \ldots, t_c(x;\theta)\big]^T, \quad x \in \mathbb{R}^d$$

ja $e_j = [0, \ldots, 0, \overset{j}{1}, 0, \ldots, 0]^T$ $\mathbb{R}^c$:n standardikannan j:s kantavektori, sekä $\gamma(j) = e_j$, $j = 1, \ldots, c$. Ehtoon (5.8) voidaan pyrkiä minimoimalla nyt

$$R(\theta) = \mathbb{E}\big(\|t(X;\theta) - \gamma(J)\|^2 \big) \tag{5.10}$$

parametrin θ suhteen.

Analogisesti tapauksen $c = 2$ kanssa voidaan osoittaa, että jos $R(\bar{\theta})$ on pieni, niin

$$t(x;\bar{\theta}) = \begin{bmatrix} t_1(x;\bar{\theta}) \\ \vdots \\ t_c(x;\bar{\theta}) \end{bmatrix} \approx \begin{bmatrix} P(1|x) \\ \vdots \\ P(c|x) \end{bmatrix}.$$

Siten (5.9) on likimain Bayesin luokitin, kun $\theta = \bar{\theta}$.

Käytännössä (5.10) korvataan opetusvektoreiden $(X_1, J_1), \ldots, (X_n, J_n)$ avulla saatavalla estimaatilla

$$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \|t(X_i; \theta) - \gamma(J_i)\|^2 . \tag{5.11}$$

ja erillisen otannan tilanteessa tätä hieman modifioidaan.

Lineaarisessa tapauksessa optimointi hoituu taas pseudoinversiolla. Olkoon nimittäin

$$t_j(x; \theta) = a_j^T x + \alpha_j, \qquad j = 1, \ldots, c,$$

$$A = [a_1, \ldots, a_c] \in \mathbb{R}^{d \times c},$$

$$\alpha = [\alpha_1, \ldots, \alpha_c]^T \in \mathbb{R}^d,$$

$$\theta = \begin{bmatrix} A \\ \alpha^T \end{bmatrix} \in \mathbb{R}^{(d+1) \times c},$$

$$D_n = \begin{bmatrix} X_1^T & 1 \\ \vdots & \vdots \\ X_n^T & 1 \end{bmatrix} \in \mathbb{R}^{n \times (d+1)},$$

$$\Gamma = \begin{bmatrix} \gamma(J_1)^T \\ \vdots \\ \gamma(J_n)^T \end{bmatrix} = \begin{bmatrix} e_{J_1}^T \\ \vdots \\ e_{J_n}^T \end{bmatrix}.$$

Silloin

$$\widehat{R}(\theta) = \frac{1}{n} \|D_n \, \theta - \Gamma\|^2 ,$$

missä $\|\cdot\|$ on Frobeniuksen normi matriisille, $\|B\|^2 = \sum b_{ij}^2$, kun $B = [b_{ij}]$.

$\widehat{R}(\theta)$:n minimoi nyt (HT)

$$\tilde{\theta} = \begin{bmatrix} \tilde{A} \\ \tilde{\alpha} \end{bmatrix} = \begin{bmatrix} \tilde{a}_1 & \cdots & \tilde{a}_c \\ \tilde{\alpha}_1 & \cdots & \tilde{\alpha}_c \end{bmatrix} = D_n^+ \, \Gamma.$$

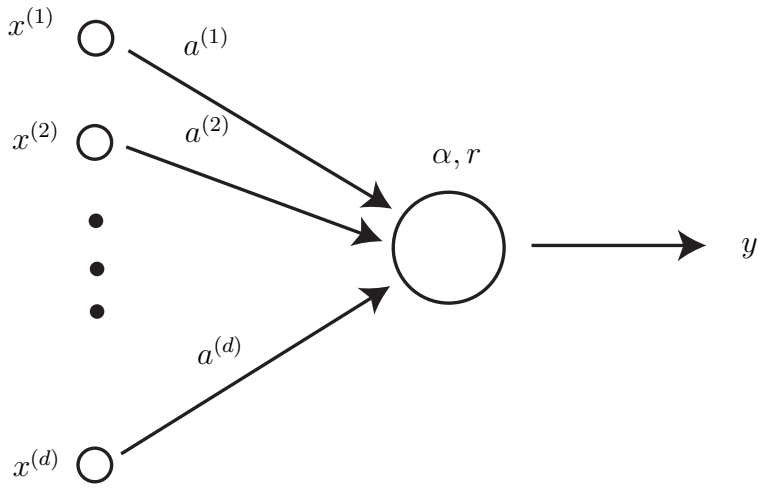
5.4 Neuroverkko

Tosiasia on, että monissa sovelluksissa ei automaattisilla hahmontunnistumenetelmillä kymmenien vuosien kehitystyöstä huolimatta edelleenkään päästä lähellekään ihmisen suorituskykyä. Jo 1960-luvulta alkaen onkin pyritty kehittämään hahmontunnistusmenetelmiä, jotka jollain tasolla jäljittelevät biologisten aistijärjestelmien toimintaa. Itseasiassa koko modernin tietojenkäsittelyn teoriankin voi katsoa syntyneen 1940-luvulla tällaisten ideoiden pohjalta.

Biologisen aistijärjestelmän perustana on hermoverkko, joukko toisiinsa kytkettyjä hermosoluja, *neuroneja*. Tästä äärimmäinen esimerkki ovat ihmisen aivot, joissa on noin 10^{11} neuronia, yksi neuroni tyypillisesti kytkettynä jopa 10^4:ään toiseen neuroniin. Tämän massiivisen kytkentöjen määrän ajatellaan olevan tärkeä syy aivojen huikeaan suorituskykyyn.

Neuronit kytkeytyvät toisiina *synapsien* avulla ja kullakin synapsilla on siihen kytkevään neuroniin joko kiihdyttävä tai estävä vaikutus. Neuroni laukoo sähköimpulsseja pitkin *aksonia*, kun siihen tulevat syötteet ylittävät tietyn kynnysarvon. Keinotekoisessa hermoverkossa, *neuroverkossa*, yksittäistä hermosolua voidaan mallintaa ns. *formaalilla neuronilla*, jonka idean esittivät Warren McCullogh ja Walter Pitts vuonna 1943 julkaistussa tutkimuksessaan (kuva 5.2). Tätä tutkimusta pidetään myös teoreettisen tietojenkäsittelytieteen lähtölaukauksena.

Formaalissa neuronissa olevilla suureilla on seuraavat tulkinnat:



Kuva 5.2: McCulloghin ja Pittsin formaali neuroni.

$x = [x^{(1)}, \ldots, x^{(d)}]^T$	syöte
$a = [a^{(1)}, \ldots, a^{(d)}]^T$	synapsien kytkentävoimakkuudet, ns. painot
α	neuronin kynnysarvo
$y = r(a^T x + \alpha)$	neuronin ulostulo
$r : \mathbb{R} \to \mathbb{R}$	aktivaatiofunktio

Esimerkki 5.2. Eräs yksinkertainen aktivaatiofunktio on

$$r(u) = \operatorname{sgn}(u) = \begin{cases} -1, & \text{kun } u < 0, \\ +1, & \text{kun } u \geq 0 \end{cases}$$

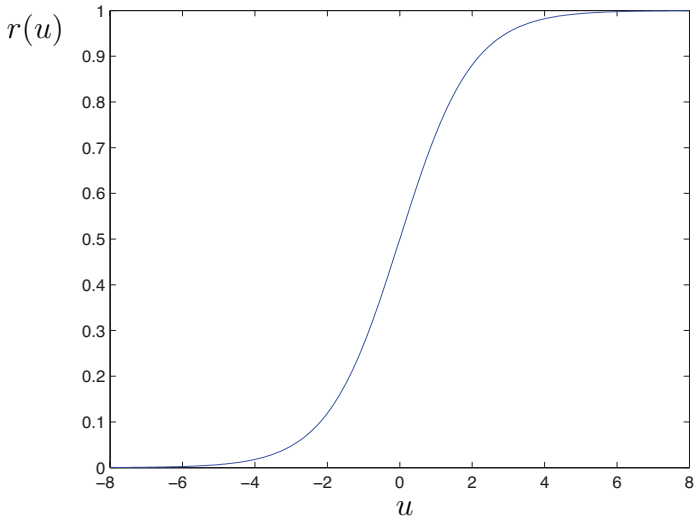
Tällöin neuronin ulostulo on

$$y = r(a^T x + \alpha) = \begin{cases} -1, & \text{kun } a^T x + \alpha < 0, \\ +1, & \text{kun } a^T x + \alpha \geq 0. \end{cases}$$

Tällainen formaali neuroni toimii siis itse asiassa lineaarisena luokittimena, kun luokkia 1 ja 2 vastaavat ulostulon arvot -1 ja +1. ‖

Esimerkki 5.3. (Sigmoidi) Aktivaatiofunktiolta r usein vaaditaan, että se on kasvava ja että $\lim_{u \to -\infty} r(u) = 0$, $\lim_{u \to \infty} r(u) = 1$. Erittäin suosittu aktivaatiofunktio on logistinen funktio, eli *sigmoidi* (kuva 5.3),

$$r(u) = \frac{1}{1 + e^{-u}}.$$



Kuva 5.3: Sigmoidi.

‖

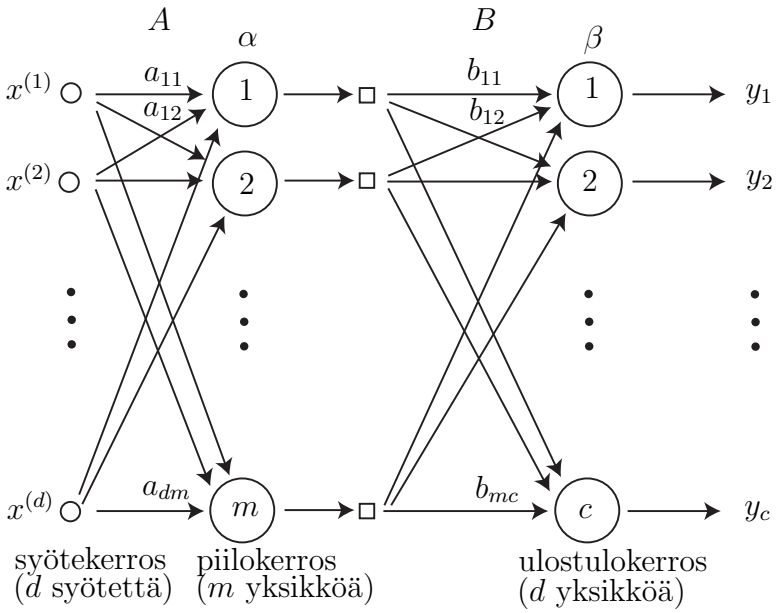
Kytkemällä formaaleja neuroneja verkoiksi saadaan keinotekoinen hermoverkko eli neuroverkko. Suosituin malli on ns. kaksikerroksinen yksisuuntainen verkko, jossa aktivaatiofunktiona on sigmoidi (kuva 5.4).

Tässä mallissa

$A = [a_{ik}]$	a_{ik} on syötteen $x^{(i)}$ paino yksikköön k
$\alpha = [\alpha_1, \ldots, \alpha_m]^T$	piilokerroksen yksiköiden kynnykset
$B = [b_{jk}]$	b_{jk} on piilokerroksen yksikön k ulostulon paino yksikköön j ulostulokerroksessa
$\beta = [\beta_1, \ldots, \beta_c]^T$	ulostulokerroksen kynnykset

Kootaan nyt kaikki painot ja kynnykset parametrivektoriin θ. Silloin verkko määrittelee kuvauksen $\mathbb{R}^d \to \mathbb{R}^c$,

$$x \mapsto t(x; \theta) = y.$$



Kuva 5.4: Kaksikerroksinen yksisuuntainen keinotekoinen hermoverkko.

Verkko voidaan saada toimimaan luokittimena minimoimalla θ:n suhteen lauseke (5.10),

$$R(\theta) = \mathbb{E}\big(\|t(X;\theta) - \gamma(J)\|^2 \big)$$

tai käytännössä tämän estimaatti (5.11),

$$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \|t(X_i;\theta) - \gamma(J_i)\|^2 .$$

Tällaisen neuroverkon tekee erityisen kiinnostavaksi kaksi seikkaa:

1. Voidaan osoittaa, että funktioilla $t(\,\cdot\,;\theta)$ voidaan approksimoida hyvin monenlaisia funktioita mielivaltaisella tarkkuudella.

2. θ:n optimointia varten löytyy kätevä algoritmi, jolla $\widehat{R}(\theta)$:n osittaisderivaatat θ:n komponenttien suhteen ja siten gradientti $\nabla_\theta \widehat{R}(\theta)$) voidaan laskea.

Kuvatunlainen neuroverkko onkin varsin suosittu monissa hahmontunnistussovelluksissa.

Luku 6

Tiheysfunktion estimointiin perustuvia parametrittomia luokittimia

6.1 Tiheysfunktion estimoinnista

Tarkastellaan Bayesin luokitinta c:n luokan tilanteessa,

$$g^*(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}} P_j f_j(x),$$

(vrt. (3.8)). Kuten jo aikaisemmin todettiin, käytännössä P_j ja f_j pitää estimoida opetusaineiston $(X_1, J_1), \ldots, (X_n, J_n)$ avulla. Erityisen haastavaa on luokkatiheysfunktion f_j estimointi.

Olkoon yleisesti $f : \mathbb{R}^d \to [0, \infty[$ jonkun satunnaisvektorin X tiheysfunktio ja $X_1, \ldots, X_n \sim f$ eli $X_1, \ldots, X_n$ on satunnaisotos X:n jakaumasta. Niin sanotussa *parametrisessa* tiheysfunktion estimoinnissa oletamme, että $f = f(\,\cdot\,; \theta)$, jollain θ, missä $f(\,\cdot\,; \theta)$ on parametrivektorin θ avulla määritelty

tiheysfunktio. Sitten θ estimoidaan otoksesta.

Esimerkki 6.1. Olkoon $\theta = [\mu, \Sigma]$ (sopivasti vektoriksi kirjoitettuna) ja

$$f(x;\theta) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}}\, e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^d.$$

Nyt siis $f = f(\,\cdot\,;\theta)$ tarkoittaa, että oletamme $X \sim N(\mu, \Sigma)$. Voimme ottaa (ks. luku 2.3),

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T, \qquad \hat{\theta} = [\hat{\mu}, \widehat{\Sigma}],$$

jolloin toivon mukaan $f(\,\cdot\,;\hat{\theta}) \approx f$. ‖

Jos tätä ideaa nyt sovelletaan Bayesin luokittimeen (vrt. luku 3.5), saadaan ensin luokkatiheysfunktiolle f_j estimaatti

$$\hat{f}_j = f(\,\cdot\,;\hat{\theta}_j), \qquad \hat{\theta}_j = [\hat{\mu}_j, \widehat{\Sigma}_j],$$
$$\hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ji}, \qquad \widehat{\Sigma}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(X_{ji} - \hat{\mu}_j)(X_{ji} - \hat{\mu}_j)^T,$$

missä $X_{j1}, \ldots, X_{jn_j}$ ovat opetusvektorit luokasta j. Ja jos sitten prioritodennäköisyydet P_j on estimoitu vaikka suureilla

$$\widehat{P}_j = \frac{n_j}{n},$$

saa esimerkiksi estimoitu kvadraattinen luokitin tapauksessa $c = 2$ muodon

$$\frac{1}{2}(x-\hat{\mu}_1)^T\widehat{\Sigma}_1^{-1}(x-\hat{\mu}_1) - \frac{1}{2}(x-\hat{\mu}_2)^T\widehat{\Sigma}_2^{-1}(x-\hat{\mu}_2)$$
$$- \frac{1}{2}\log\left(\frac{\det(\widehat{\Sigma}_2)}{\det(\widehat{\Sigma}_1)}\right) - \log\left(\frac{\widehat{P}_1}{\widehat{P}_2}\right) \underset{2}{\overset{1}{\lessgtr}} 0$$

(vertaa (4.1)).

Vastaavasti estimoidaan muut luvun 4 normaalijakaumaan perustuvat luokittimet tai ylipäänsä parametrien μ_j ja Σ_j käyttöön perustuvat luokittimet, kuten esimerkiksi Fisherin luokitin.

Tässä luvussa kuitenkin tarkastelemme niin sanottuja *parametrittomia* tiheysfunktion estimointimenetelmiä ja niihin perustuvia luokittimia.

Kuva 6.1: Histogrammin määrittelyssä käytettyjä reaaliakselin jakopisteitä.

6.1.1 Histogrammi

Tavallisin ja tunnetuin tiheysfunktion parametriton estimaattori on histogrammi. Olkoon ensin $d = 1$ ja tarkastellaan reaaliakselin pisteistöä $u_k \in \mathbb{R}$, $k \in \mathbb{Z}$, $\cdots < u_k < u_{k+1} < \cdots$ (kuva 6.1). Usein käytetään itseasiassa tasavälistä jakoa, $u_{k+1} - u_k = h$, $k \in \mathbb{Z}$. Olkoon $X_1, \ldots, X_n \sim f$ satunnaisotos. Merkitään

$$N_k = \big|\{i \mid X_i \in [u_k, u_{k+1}[\}\big|, \qquad k \in \mathbb{Z}.$$

Tässä merkintä $|S|$ tarkoittaa (äärellisen) joukon S alkioiden lukumäärää. Selvästikin vain äärellisen monella k on $N_k \neq 0$ ja lisäksi $\sum_k N_k = n$. Nyt f:n *histogrammiestimaattori* on

$$\hat{f}_n(x) = \sum_{k \in \mathbb{Z}} \frac{N_k/n}{u_{k+1} - u_k} 1_{[u_k, u_{k+1}[}(x), \quad x \in \mathbb{R},$$

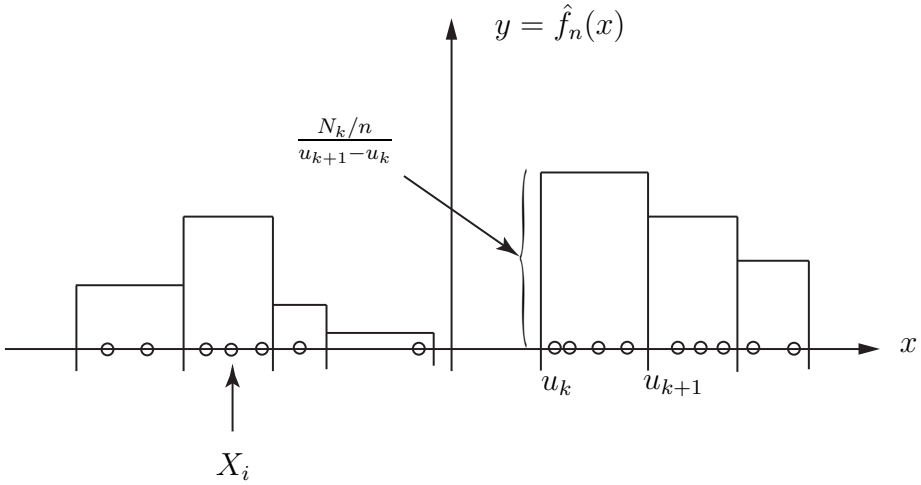
missä $1_{[u_k, u_{k+1}[}$ on välin $[u_k, u_{k+1}[$ karakteristinen funktio. Siis

$$f_n(x) = \frac{N_k/n}{u_{k+1} - u_k},$$

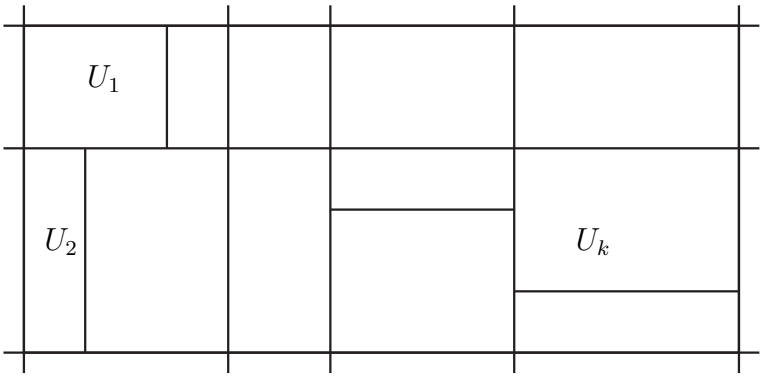
kun $x \in [u_k, u_{k+1}[$ (kuva 6.2).

Histogrammin ideaa voi perustella olettamalla, että $|u_{k+1} - u_k|$ pieni ja n suuri, jolloin pisteessä $x \in [u_k, u_{k+1}[$ saadaan

$$\begin{aligned} \hat{f}_n(x) = \frac{N_k/n}{u_{k+1} - u_k} &\approx \frac{\mathbb{P}\big(X \in [u_k, u_{k+1}[\big)}{u_{k+1} - u_k} \\ &= \frac{\int_{u_k}^{u_{k+1}} f(u)\, du}{u_{k+1} - u_k} \approx \frac{f(x) \cdot (u_{k+1} - u_k)}{u_{k+1} - u_k} = f(x), \end{aligned}$$



Kuva 6.2: Tiheysfunktion histogrammiestimaattori.



Kuva 6.3: Eräs avaruuden osiinjako moniulotteista histogrammia varten.

missä ensimmäinen approksimaatio seuraa todennäköisyyslaskennan suurten lukujen laista. Siten todellakin $\hat{f}_n(x) \approx f(x)$.

Kun $d > 1$, välit $[u_k, u_{k+1}[$ korvataan pistevierailla joukoilla $U_k \subset \mathbb{R}^d$, esimerkiksi suorakulmaisilla särmiöillä (kuva 6.3) ja asetetaan

$$\hat{f}_n(x) = \sum_k \frac{N_k/n}{m(U_k)} 1_{U_k}(x), \qquad x \in \mathbb{R}^d,$$

missä $m(U_k)$ on joukon U_k tilavuus (tai Lebesguen mitta).

Histogrammiestimaattorin käytössä on kuitenkin eräitä ongelmia:

- Tarvittavien laatikoiden määrä kasvaa räjähdysmäisesti d:n mukana.
- Jopa tasavälisessä histogrammissa joudutaan välin pituuden h lisäksi valitsemaan myös hilan $\cdots < u_k < u_{k+1} < \cdots$ paikka.
- Estimaattori $\hat{f}_n$ on epäjatkuva.
- Histogrammin eräät teoreettiset estimointiominaisuudet eivät ole kovin hyvät.

6.1.2 Ydinestimaattori

Olkoon $d = 1$, f X:n tiheysfunktio ja $X_1, \ldots, X_n \sim f$ satunnaisotos. Olkoon edelleen

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

X:n kertymäfunktio.

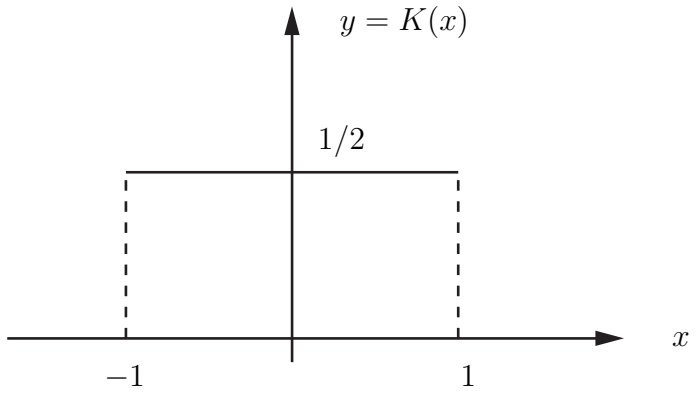
F:ää approksimoi ns. empiirinen kertymäfunktio,

$$\widehat{F}_n(x) = \frac{1}{n} \left| \{i \mid X_i \leq x\} \right| .$$

Valitaan nyt $h > 0$ ja approksimoidaan edelleen seuraavasti:

$$\begin{aligned} f(x) = F'(x) &\approx \frac{1}{2} \left\{ \frac{\widehat{F}_n(x+h) - \widehat{F}_n(x)}{h} + \frac{\widehat{F}_n(x) - \widehat{F}_n(x-h)}{h} \right\} \\ &= \frac{1}{2h} \left[\widehat{F}_n(x+h) - \widehat{F}(x-h) \right] \\ &= \frac{1}{2nh} \left| \{i \mid x - h < X_i \leq x + h\} \right| \\ &= \frac{1}{2nh} \left| \{i \mid -h < X_i - x \leq h\} \right| \\ &\equiv \hat{f}_n(x). \end{aligned}$$

Tässä $\hat{f}_n$ tunnetaan f:n *naiivina estimaattorina*.



Kuva 6.4: Naiivin ydinestimaattorin määrittelyssä käytettävä funktio K.

Olkoon sitten $K = \frac{1}{2}\,1_{[-1,1[}$ (kuva 6.4). Silloin

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{2nh}\Big|\{i \mid -h < X_i - x \le h\}\Big| \\
&= \frac{1}{2nh}\left|\left\{i \mid -1 \le \frac{x - X_i}{h} < 1\right\}\right| \\
&= \frac{1}{2nh}\sum_{i=1}^{n} 1_{[-1,1[}\left(\frac{x - X_i}{h}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \qquad x \in \mathbb{R}. \qquad (6.1)
\end{aligned}$$

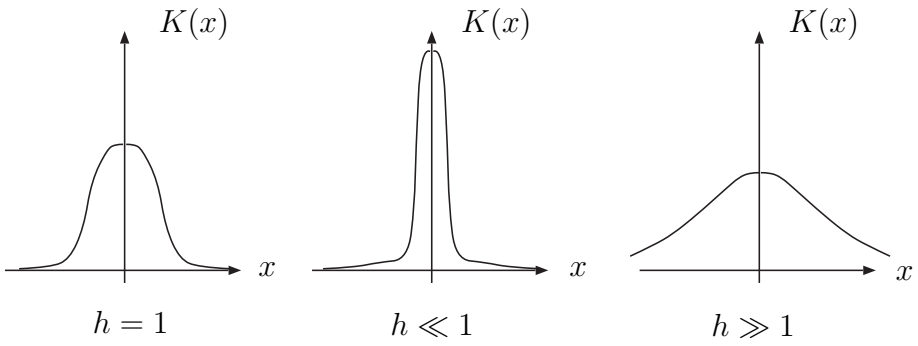
Kaava (6.1) motivoi seuraavan määritelmän.

Määritelmä 6.2. Olkoon $K : \mathbb{R}^d \to \mathbb{R}$ integroituva, $\int\limits_{\mathbb{R}^d} K(x)\,dx = 1$, ja $h > 0$. Silloin f:n *ydinestimaattori* on

$$\hat{f}_n(x; h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right), \qquad x \in \mathbb{R}^d. \qquad (6.2)$$

Tässä K on estimaattorin *ydin* ja h on *silotusparametri*. Edelleen, $\frac{x-X_i}{h}$ tarkoittaa vektoria $h^{-1}(x - X_i)$. Merkitään

$$K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right), \qquad x \in \mathbb{R}^d.$$



Kuva 6.5: Ytimen K muodon riippuvuus silotusparametrista h.

Silloin

$$\hat{f}_n(x,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i). \tag{6.3}$$

Nyt pätee

$$\begin{aligned}\int_{\mathbb{R}^d} K_h(x)dx &= \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x}{h}\right) dx = \int_{\mathbb{R}^d} \frac{1}{h^d}\, K(y)h^d dy \\ &= \int_{\mathbb{R}^d} K(y)dy = 1.\end{aligned}$$

missä toisessa yhtälössä tehtiin sijoitus $y = x/h$. Tästä seuraa edelleen helposti (HT), että

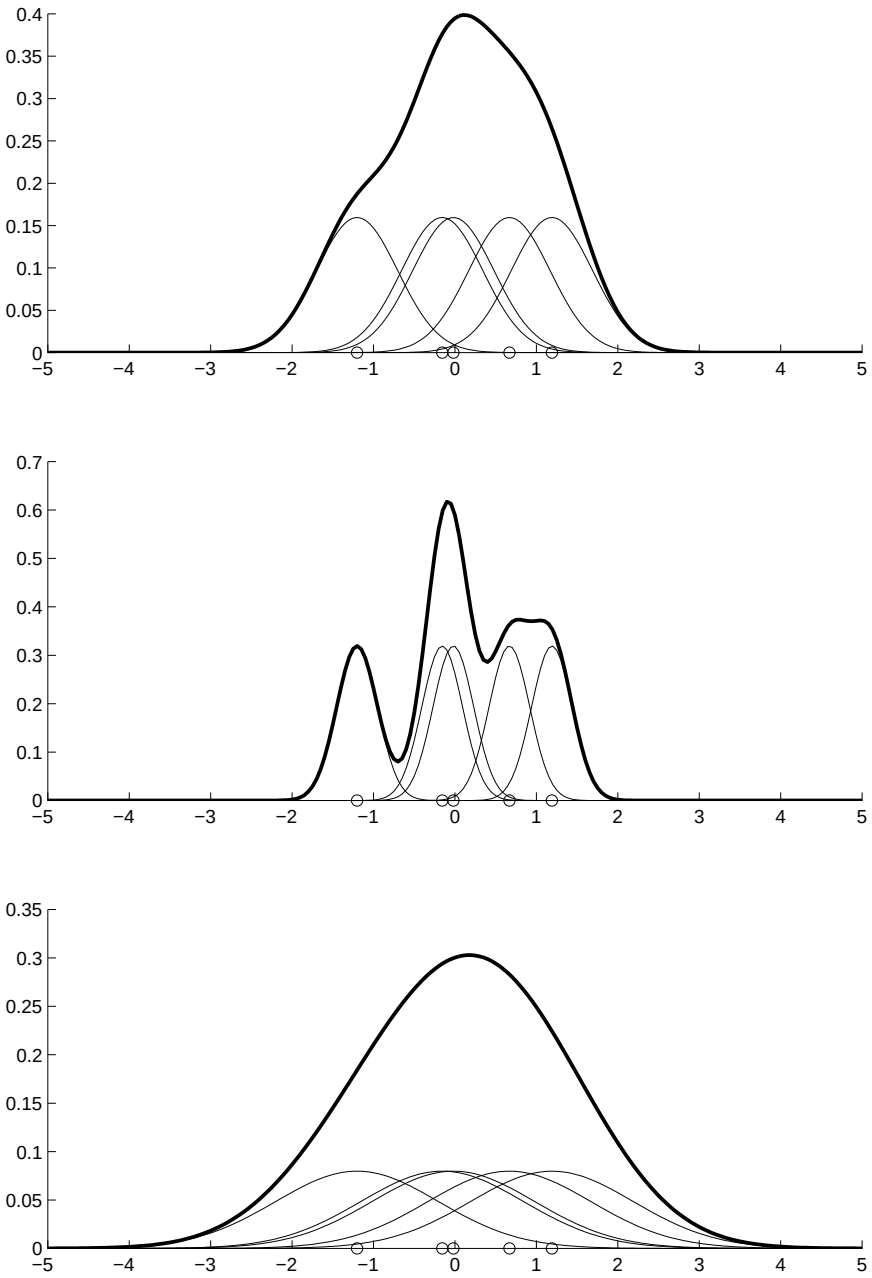
$$\int_{\mathbb{R}^d} \hat{f}(x;h)dx = 1.$$

Usein ytimeksi K valitaan itseasiassa jokin tiheysfunktio, esimerkiksi

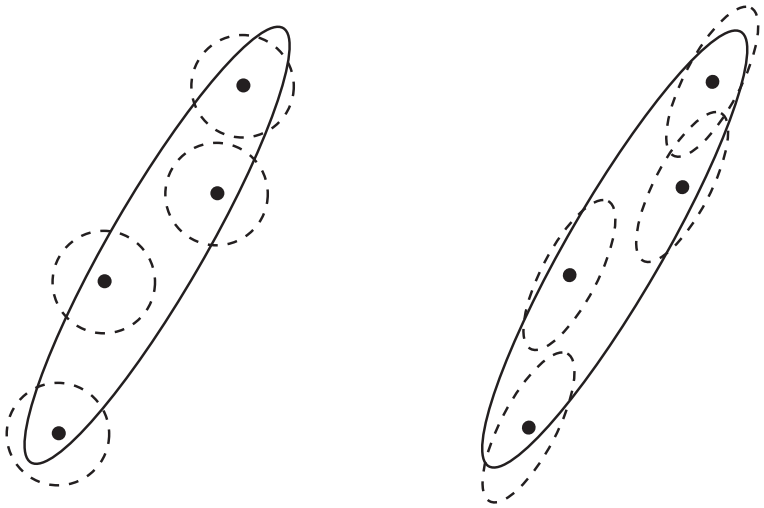
$$K(x) = \frac{1}{(2\pi)^{d/2}}\, e^{-\frac{1}{2}\|x\|^2}, \qquad x \in \mathbb{R}^d, \tag{6.4}$$

jota kutsutaan tässä yhteydessä *Gaussin ytimeksi*.

Kuvassa 6.5 on esitetty miten silotusparamteri h vaikuttaa K_h:n muotoon: pieni h vastaa kapeaa ja korkeaa ydintä kun taas surella h ydin on laakea ja matala. Ydinestimaattorissa $\hat{f}_n(\cdot;h)$ pisteisiin X_i sijoitetaan skaalatut ytimet $\frac{1}{n}K_h$ ja sitten ne summataan yhteen. Tätä on havainnollistettu kuvassa 6.6.



Kuva 6.6: Tiheysfunktion ydinestimaatti (paksu viiva) muodostettuna kolmella silotusparametrin arvolla. Liian pieni h (keskimmäinen kuva) johtaa rosoiseen estimaattiin ja liian suuri h (alin kuva) puolestaan liian sileään estimaattiin. Otospisteet on saatu jakaumasta $\mathrm{N}(0, 1)$. Ne ovat kaikissa kuvissa samat ja merkitty pienillä ympyröillä vaaka-akselille.



Kuva 6.7: Tiheysfunktion estimaatin riippuvuus käytetyn ytimen muodosta. Vasemmalla on ytimenä jakauman $\mathrm{N}(0, I)$ tiheysfunktio ja oikealla jakauman $\mathrm{N}(0, \Sigma)$ tiheysfunktio. Estimoitavan tiheysfunktion muotoa kuvaa yhtenäisellä viivalla piirretty tasa-arvokäyrä, jonka ajatellaan määräytyvän jakauman kovarianssimatriisista Σ.

Jos f:n "muoto" (tasa-arvopinnat) poikkeaa voimakkaasti pallosymmetriasta, voi esimerkiksi ytimen (6.4) sijaan olla parempi käyttää muunnettua ydintä

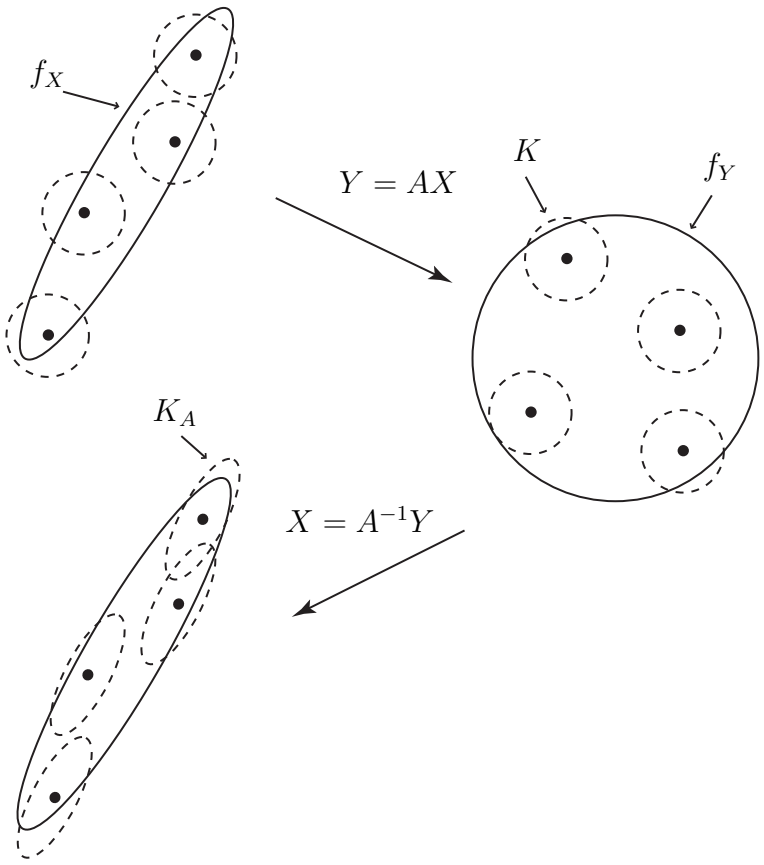
$$\det(\Sigma^{-1/2})K(\Sigma^{-1/2}\,x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}, \quad x \in \mathbb{R}^d,$$

eli jakauman $\mathrm{N}(0, \Sigma)$ tiheysfunktiota, missä Σ on X:n kovarianssimatriisi (kuva 6.7).

Toinen mahdollisuus on käyttää valkaisua (luku 2.5):

(i) Olkoon $\Sigma_X = U\Lambda U^T$, jolloin valkaisun toteuttaa matriisi $A = \Lambda^{-1/2}U^T$. Olkoon edelleen valkaistu satunnaismuuttuja $Y = AX$ ja $Y_i = AX_i$, $i = 1 \ldots, n$. Nyt $\Sigma_Y = I_d$.

(ii) Muodostetaan Y:n tiheysfunktiolle f_Y ydinestimaattori



Kuva 6.8: Valkaisemalla saatava tiheysfunktion ydinestimaattori.

$\hat{f}_{Y,n}(\cdot\,; h)$ jollain pallosymmetrisellä ytimellä K käyttäen muunnettua otosta $Y_1, \ldots, Y_n$.

(*iii*) Muodostetaan X:n tiheysfunktiolle f_X estimaattori

$$\hat{f}_{X,n}(x; h) = |\det A| \hat{f}_{Y,n}(Ax; h), \quad x \in \mathbb{R}^d.$$

Voidaan osoittaa (HT), että $\hat{f}_{X,n}$ on silloin ydinestimaattori muunnetulla ytimellä $K_A(x) = |\det A| K(Ax)$ (kuva 6.8). Käytännössä kovarianssimatriisin Σ paikalla joudutaan käyttämään sen estimaattia.

Ytimen muotoa paljon tärkeämpi kysymys on itseasiassa silotusparametrin $h > 0$ valinta. Palaamme myöhemmin siihen miten h voidaan käytännössä

valita luokitinta konstruoitaessa. Seuraavassa olemme kuitenkin kiinnostuneita ensisijaisesti siitä, että $\hat{f}_n(\cdot; h)$ estimoi mahdollisimman hyvin f:ää.

6.1.3 Ydinestimaattorin tarkentuvuus

Olkoon ensin $x \in \mathbb{R}^d$ kiinteä. Koska $X_1, \ldots, X_n \sim f$ ovat satunnaisvektoreita, on

$$\hat{f}_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

satunnaismuuttuja. Siis $\hat{f}_n(\cdot; h)$ on itseasiassa *satunnaisfunktio* (tai *stokastinen prosessi*).

Estimointivirhe pisteessä x on

$$\hat{f}_n(x; h) - f(x).$$

Tämäkin on satunnaismuuttuja ja siksi onkin helpompaa tarkastella esimerkiksi keskimääräistä neliöllistä virhettä

$$\mathbb{E}\big(\hat{f}_n(x; h) - f(x)\big)^2,$$

joka on reaaliluku.

Virhe koko avaruudessa $\mathbb{R}^d$ määritellään esimerkiksi integraalina

$$\int_{\mathbb{R}^d} \mathbb{E}\big(\hat{f}_n(x; h) - f(x)\big)^2 dx.$$

Voidaan osoittaa, että

$$\int_{\mathbb{R}^d} \mathbb{E}\big(\hat{f}_n(x; h) - f(x)\big)^2 dx = \mathbb{E} \int_{\mathbb{R}^d} \big(\hat{f}_n(x; h) - f(x)\big)^2 dx.$$

Tätä kutsutaan *keskimääräiseksi integroiduksi neliölliseksi virheeksi*. Analysoidaan tätä virhettä seuraavassa yksinkertaisuuden vuoksi tapauksessa

$d = 1$. Olkoon $\|g\|_2$ funktion $g : \mathbb{R} \to \mathbb{R}$ niin sanottu L^2-normi,

$$\|g\|_2 = \sqrt{\int_{-\infty}^{\infty} g(x)^2 dx}.$$

Lause 6.3. *Olkoon $f : \mathbb{R} \to [0, \infty[$ kaksi kertaa jatkuvasti derivoituva tiheysfunktio, jolle $\|f\|_2$, $\|f'\|_2$, $\|f''\|_2 < \infty$. Olkoon ytimelle K voimassa $K(x) = K(-x)$ kaikilla $x \in \mathbb{R}$ (symmetria), $\|K\|_2 < \infty$ ja $\sigma_K^2 = \int_{-\infty}^{\infty} x^2 K(x) dx < \infty$. Silloin*

$$\mathbb{E} \int_{-\infty}^{\infty} \big(\hat{f}_n(x;h) - f(x)\big)^2 dx = \frac{1}{4} h^4 \sigma_K^4 \|f''\|_2^2 + \frac{\|K\|_2^2}{nh} + \varepsilon(h) \left(h^4 + \frac{1}{nh}\right), \quad (6.5)$$

missä $\varepsilon(h) \to 0$, kun $h \to 0+$.

Todistus. Olkoon $x \in \mathbb{R}$ kiinteä ja merkitään

$$Y_i = K_h(x - X_i), \qquad i = 1, \ldots, n.$$

Silloin $Y_1, \ldots, Y_n$ ovat riippumattomia satunnaismuuttujia (koska X_i:t ovat sitä) ja

$$\hat{f}_n(x;h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Olkoon edelleen

$$Z = \hat{f}_n(x;h) - f(x).$$

Silloin

$$\begin{aligned} \mathbb{E}\big(\hat{f}_n(x;h) - f(x)\big)^2 &= \mathbb{E}Z^2 = (\mathbb{E}Z)^2 + D^2(Z) \\ &= \Big(\mathbb{E}\hat{f}_n(x;h) - f(x)\Big)^2 + D^2\big(\hat{f}_n(x;h)\big), \end{aligned} \quad (6.6)$$

missä $\mathbb{E}\hat{f}_n(x;h) - f(x)$ ja ja $D^2\big(\hat{f}_n(x;h)\big)$ ovat vastaavasti estimaattorin $\hat{f}_n(\cdot;h)$ *harha* ja *varianssi*.

Nyt

$$\mathbb{E}\hat{f}_n(x;h) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Y_i.$$

Tässä

$$\begin{aligned}\mathbb{E}Y_i = \mathbb{E}K_h(x - X_i) &= \int_{-\infty}^{\infty} K_h(x-y)f(y)dy \\ &= \int_{-\infty}^{\infty} f(x-y)K_h(y)dy \equiv f * K_h(x),\end{aligned}$$

missä toinen yhtäsuuruus seurasi siitä, että $X_i \sim f$ ja kolmannessa otettiin uudeksi integoimismuuttujaksi $x - y$. Lauseke $f * K_h$ on f:n ja K_h:n *konvoluutio*.

Siten

$$\mathbb{E}\hat{f}_n(x;h) = \frac{1}{n}\sum_{i=1}^{n}(f * K_h)(x) = f * K_h(x)$$

ja (6.6):n harhan neliö saadaan muotoon

$$\big(\mathbb{E}\hat{f}_n(x;h) - f(x)\big)^2 = \big(f * K_h(x) - f(x)\big)^2. \tag{6.7}$$

Varianssille puolestaan saadaan

$$D^2\big(\hat{f}_n(x;h)\big) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2 Y_i,$$

koska $Y_1, \ldots, Y_n$ ovat riippumattomia. Edelleen,

$$\begin{aligned}D^2\, Y_i &= \mathbb{E}(Y_i^2) - (\mathbb{E}Y_i)^2 \\ &= \mathbb{E}\left[K_h(x - X_i)^2\right] - \left[\mathbb{E}K_h(x - X_i)\right]^2 \\ &= \int_{-\infty}^{\infty}\big(K_h(x-y)\big)^2 f(y)dy - \left(\int_{-\infty}^{\infty} K_h(x-y)f(y)dy\right)^2 \\ &= f * (K_h)^2(x) - \big(f * K_h(x)\big)^2,\end{aligned}$$

joten

$$D^2\big(\hat{f}_n(x;h)\big) = \frac{1}{n}\left[f * (K_h)^2(x) - \big(f * K_h(x)\big)^2\right]. \tag{6.8}$$

Tiheysfunktion f toisen kertaluvun Taylorin kehitelmä pisteessä x on

$$\begin{aligned} f(x+y) &= f(x) + f'(x)y + \frac{1}{2}f''(\theta)y^2 \\ &= f(x) + f'(x)y + \frac{1}{2}f''(x)y^2 + \frac{1}{2}\left[f''(\theta) - f''(x)\right]y^2, \end{aligned}$$

missä θ on pisteiden x ja $x+y$ välissä. Merkitään

$$\frac{1}{2}\left[f''(\theta) - f''(x)\right] = \delta(x,y).$$

Siten kaavassa (6.7)

$$\begin{aligned} f * K_h(x) - f(x) &= \int_{-\infty}^{\infty} f(x-y)K_h(y)dy - f(x) \\ &= \int_{-\infty}^{\infty} \big[f(x-y) - f(x)\big]K_h(y)dy \\ &= \int_{-\infty}^{\infty} \big[f(x+hz) - f(x)\big]K(z)dz \\ &= \int_{-\infty}^{\infty} \left[f'(x)hz + \frac{1}{2}f''(x)(hz)^2 + \delta(x,hz)(hz)^2\right]K(z)dz \\ &= f'(x)h\int_{-\infty}^{\infty} zK(z)dz + \frac{1}{2}h^2 f''(x)\int_{-\infty}^{\infty} z^2K(z)dz \\ &\quad + h^2\int_{-\infty}^{\infty} \delta(x,hz)z^2K(z)dz, \end{aligned}$$

missä toinen yhtäsuuruus seuraa ehdosta $\int_{\mathbb{R}^d} K_h(y)dy = 1$ ja kolmannessa yhtälössä vaihdettiin integroimimuuttujaksi $z = -y/h$ sekä hyödynnettin ytimen K symmetrisyyttä. Tässä kehitelmässä

- $\int_{-\infty}^{\infty} zK(z)dz = 0$, koska K on symmetrinen,
- $\int_{-\infty}^{\infty} z^2K(z)dz = \sigma_K^2$,
- $\delta(x, hz) = f''(\theta) - f''(x)$ missä θ on pisteiden x ja $x + hz$ välissä, siis pieni, kun $h \to 0+$.

Tästä seuraa helosti, että integroidulle harhalle pätee

$$\int_{-\infty}^{\infty} \big(f * K_h(x) - f(x)\big)^2 dx = \frac{1}{4}\, h^4 \sigma_K^4 \, \|f''\|_2^2 + h^4 \varepsilon_1(h), \tag{6.9}$$

missä $\varepsilon_1(h) \to 0$, kun $h \to 0+$.

Vastaavasti integroidulle varianssille saadaan tuloksesta (6.8)

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{1}{n} \Big\{ f * (K_h)^2(x) - \big(f * K_h(x)\big)^2 \Big\}\, dx \\ &\qquad = \frac{1}{nh}\big(\|K\|_2^2 + \varepsilon_2(h)\big) - \frac{1}{n}\big(\|f\|_2^2 + \varepsilon_3(h)\big) \\ &\qquad = \frac{1}{nh}\, \|K\|_2^2 + \frac{1}{nh}\varepsilon_4(h), \end{aligned} \tag{6.10}$$

missä $\varepsilon_2(h)$, $\varepsilon_3(h)$, $\varepsilon_4(h) \to 0$, kun $h \to 0+$.

Yhdistämällä tulokset (6.9) ja (6.10) saadaan lopulta väite. □

Korollaari 6.4. *Olkoon lauseen 6.3 oletukset voimassa ja (h_n) jono silotusparametreja, joille pätee*

$$\lim_{n\to\infty} h_n = 0 \quad \textit{ja} \quad \lim_{n\to\infty} n\, h_n = \infty. \tag{6.11}$$

Silloin $\hat{f}_n(\cdot\,; h)$ on tiheysfunktion f tarkentuva *estimaattori, eli*

$$\lim_{n\to\infty} \mathbb{E} \int_{-\infty}^{\infty} \big(\hat{f}_n(x; h_n) - f(x)\big)^2 dx = 0.$$

Tällöin siis estimaattori $\hat{f}_n(\cdot\,; h)$ suppenee kohti f:ää keskimääräisen integroidun neliöllisen virheen mielessä.

Todistettu lause ja sen korollaari osoittavat, että ydinestimaattorilla voidaan estimoida hyvin monenlaisia tiheysfunktioita mielivaltaisen tarkasti kunhan vaan käytettävissä on riittävän suuri otos $X_1, \ldots, X_n \sim f$.

Lause 6.3 pätee myös $d-$ulotteisessa tilanteessa ja tällöin ehdot (6.11) saavat muodon

$$\lim_{n\to\infty} h_n = 0 \quad \text{ja} \quad \lim_{n\to\infty} nh_n^d = \infty. \tag{6.12}$$

Voitaisiin myös tarkastella virhettä

$$\mathbb{E} \int_{\mathbb{R}^d} \big|\hat{f}_n(x; h) - f(x)\big| dx,$$

jolloin *ilman mitään erityisoletuksia* f:stä ja K:sta saadaan

$$\lim_{n\to\infty} \mathbb{E} \int_{\mathbb{R}^d} \big|\hat{f}_n(x; h_n) - f(x)\big| dx = 0,$$

kun ehdot (6.12) ovat voimassa. Todistus on kuitenkin varsin monimutkainen. Vieläpä pätee, että

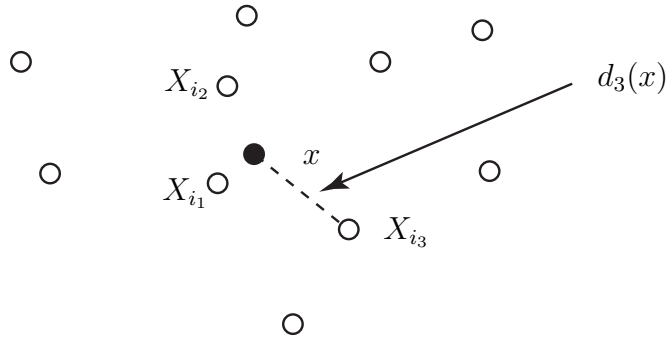
$$\lim_{n\to\infty} \int_{\mathbb{R}^d} \big|\hat{f}_n(x; h_n) - f(x)\big| dx = 0$$

todennäköisyydellä 1.

6.1.4 Lähinaapuriestimaattori

Olkoon $X_1, \ldots, X_n \sim f$ otos ja $x \in \mathbb{R}^d$. Permutoidaan etäisyydet $\|x - X_i\|$ kasvavaan järjestykseen,

$$\|x - X_{i_1}\| \leq \|x - X_{i_2}\| \leq \cdots \leq \|x - X_{i_n}\|$$



Kuva 6.9: Piste x ja sen kuusi lähinaapuria.

ja määritellään x:n etäisyys sen k:nteen lähinaapuriin (kuva 6.9),

$$d_k(x) = \|x - X_{i_k}\|,$$

$k \in \{1, \ldots, n\}$.

Olkoon $r > 0$ ja

$$B(x, r) = \{y \in \mathbb{R}^d \mid \|y - x\| \leq r\}$$

x-keskinen r-säteinen pallo, sekä $c_d = m\big(B(0,1)\big)$ yksikköpallon tilavuus. Silloin, voidaan osoittaa (HT), että

$$m\big(B(x,r)\big) = r^d c_d.$$

Määritelmä 6.5. Tiheysfunktion f *k-lähinaapuriestimaattori* on

$$\hat{f}_n(x;k) = \frac{k/n}{m\big(B(x, d_k(x))\big)} = \frac{k/n}{c_d[d_k(x)]^d}. \tag{6.13}$$

Tätä estimaattoria voidaan motivoida seuraavasti. Olkoon $n \gg 1$ ja $k \ll n$. Silloin, jos $X \sim f$, on

$$\begin{aligned} \frac{k}{n} &\approx \mathbb{P}\big(X \in B(x, d_k(x))\big) \\ &= \int\limits_{B(x,d_k(x))} f(y)dy \approx f(x)\, m\big(B(x, d_k(x))\big), \end{aligned}$$

missä ensimmäinen approksimaatio seuraa suurten lukujen laista ($n \gg 1$) ja toinen ehdosta $k \ll n$, jolloin pallo $B(x, d_k(x))$ on pieni. Siten todellakin

$$f(x) \approx \frac{k/n}{m\big(B(x, d_k(x))\big)} \tag{6.14}$$

(vrt. (6.13)). Silotusparametrin rooli on nyt k:lla sillä kun k on pieni, on estimaatti $\hat{f}_n(\cdot\,; k)$ rosoinen ja kun k on suuri, on estimaatti sileä.

Vastineena korollaarille 6.4 voidaan todistaa, että jos (k_n) on jono sellaisia luonnollisia lukuja, että

$$\lim_{n\to\infty} k_n = \infty \quad \text{ja} \quad \lim_{n\to\infty} \frac{k_n}{n} = 0,$$

niin $\hat{f}_n(x; k_n) \to f(x)$ stokastisesti (melkein kaikilla x).

Kuten ydinestimaattorin tapauksessa, jos tiheysfunktion f muoto poikkeaa voimakkaasti pallosymmetrisyydestä, voi olla hyvä käyttää valkaisua, $Y = AX$. Olkoon siis $Y_i = AX_i$, $i = 1, \ldots, n$ ja

$$\hat{f}_{Y,n}(y; k) = \frac{k/n}{c_d[d_k(y)]^d}$$

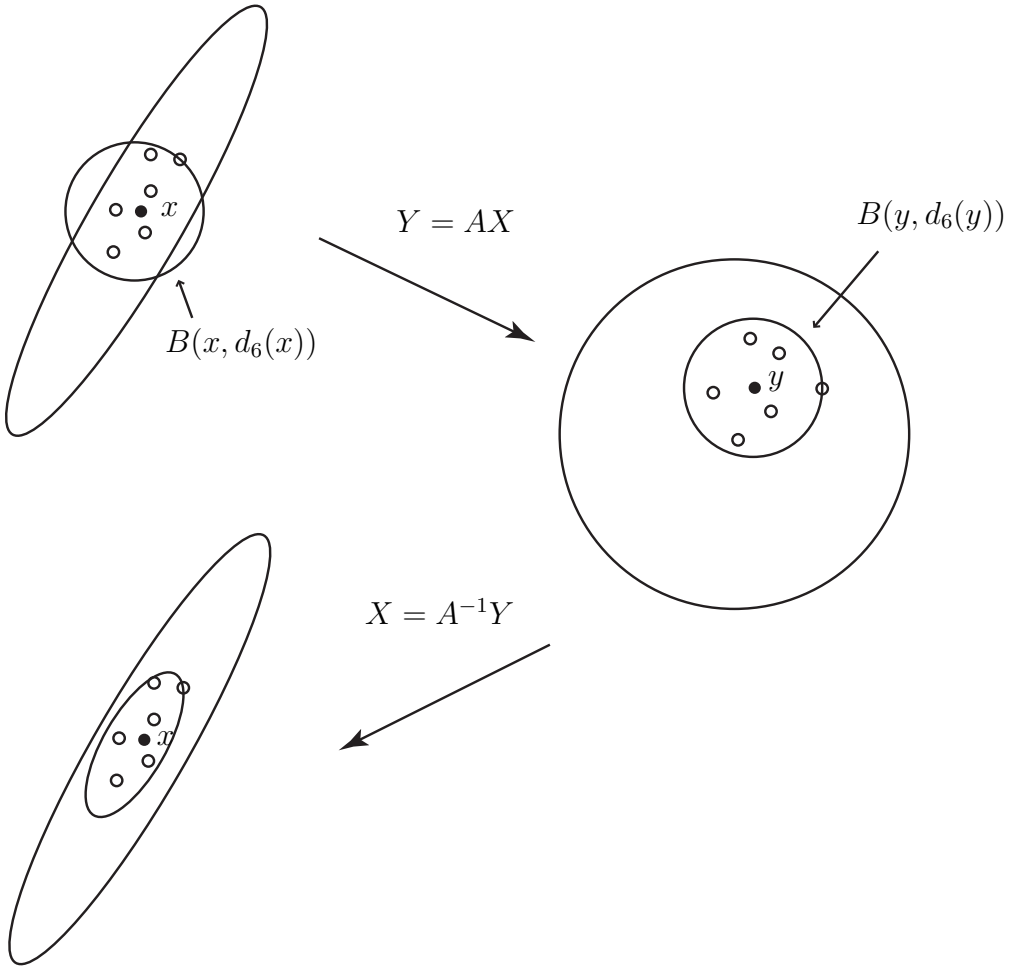
Y:n k-lähinaapuriestimaattori, joka perustuu otokseen $Y_1, \ldots, Y_n$. Silloin tiheysfunktiolle f saadaan estimaattori

$$\hat{f}_{X,n}(x; k) = |\det A| \hat{f}_{Y,n}(Ax; k), \quad x \in \mathbb{R}^d.$$

Voidaan osoittaa (HT), että $\hat{f}_{X,n}$ on itseasiassa k-lähinaapuriestimaattori otoksesta $X_1, \ldots, X_n$, kun $\mathbb{R}^d$:n normina on $\|\cdot\|_{\Sigma_X^{-1}}$ (kuva 6.10). Ideana on, että approksimaatio (6.14),

$$\int\limits_{B(x,d_6(x))} f(u)du \approx f(x)\, m\big(B\big(x, d_6(x)\big)\big)$$

luultavasti pätee paremmin valkaisun jälkeen.



Kuva 6.10: Valkaisuun perustuva tiheysfunktion lähinaapuriestimaattori.

6.2 Parametrittomat luokittimet

Olkoon g^* Bayesin luokitin,

$$g^*(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}}\, P_j f_j(x), \qquad x \in \mathbb{R}^d.$$

Olkoon edelleen $(X_1, J_1), \ldots, (X_n, J_n)$ opetusaineisto ja $X_{j1}, \ldots, X_{jn_j}$ opetusvektorit luokasta j, $j = 1, \ldots, c$. Siis

$$\{(X_1, J_1), \ldots, (X_n, J_n)\} = \bigcup_{j=1}^{c} \left\{(X_{j1}, J_{j1}), \ldots, (X_{jn_j}, J_{jn_j})\right\}.$$

Estimoimalla prioritodennäköisyydet P_j sopivilla estimaattoreilla $\widehat{P}_{j,n}$ ja luokkatiheysfunktiot f_j joillain parametrittomilla estimaattoreilla $\hat{f}_{j,n}$ saadaan *parametriton luokitin* $g_n : \mathbb{R}^d \to \{1, \ldots, c\}$,

$$g_n(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}}\, \widehat{P}_{j,n}\, \hat{f}_{j,n}(x), \qquad x \in \mathbb{R}^d. \tag{6.15}$$

Sekoiteotannan tapauksessa voidaan ottaa

$$\widehat{P}_{j,n} = \frac{n_j}{n},$$

jolloin suurten lukujen laista seuraa, että

$$\widehat{P}_{j,n} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{J_i = j\}} \xrightarrow[n\to\infty]{} \mathbb{E}1(J = j) = \mathbb{P}(J = j) = P_j$$

stokastisesti (tai jopa todennäköisyydellä 1).

Esimerkiksi Neymanin-Pearsonin luokittimessa tarvitsee kuitenkin estimoida vain luokkatiheysfunktiot f_j.

Sovellettaessa ydinestimointia, on

$$\hat{f}_{j,n}(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} K_h(x - X_{ji}), \qquad x \in \mathbb{R}^d.$$

Voidaan myös valita kullekin luokalle oma K sekä oma h.

Käytettäessä k-lähinaapuriestimointia on

$$\hat{f}_{j,n}(x) = \frac{k/n_j}{c_d[d_{j,k}(x)]^d}, \qquad x \in \mathbb{R}^d,$$

missä $d_{j,k}(x)$ on x:n etäisyys sen k:nteen lähinaapuriin joukossa $\{X_{j1}, \ldots, X_{jn_j}\}$. Jos nyt $\widehat{P}_{j,n} = n_j/n$, saadaan tästä *k-lähinaapuriluokitin*

$$g_n(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}} \left\{ \frac{n_j}{n} \cdot \frac{k/n_j}{c_d[d_{j,k}(x)]^d} \right\} = \underset{j=1,\ldots,c}{\operatorname{argmin}} d_{j,k}(x),$$

eli x luokitellaan siihen luokkaan, josta määrätty k:s lähinaapuri on sitä lähinnä.

Erikoistapauksessa $k = 1$ saadaan 1-lähinaapuriluokitin eli lyhyesti *lähinaapuriluokitin*,

$$g_n(x) = x\text{:ää lähimmän opetusvektorin } X_i \text{ luokka.}$$

Tavallisesti k-lähinaapuriluokitin kuitenkin määritellään hieman toisin. Olkoon $d_k(x)$ kuten luvussa 6.1.4,

$$\|x - X_{i_1}\| \leq \|x - X_{i_2}\| \leq \cdots \leq \|x - X_{i_k}\| = d_k(x)$$

ja määritellään

$$m_j = \big|\{l \mid J_{i_l} = j,\ l = 1, \ldots, k\}\big|,$$

eli m_j on luokasta j peräisin olevien opetusvektoreiden lukumäärä joukossa $\{X_{i_1}, \ldots, X_{i_k}\}$. Otetaan nyt luokkatiheysfunktion estimaattoriksi

$$\hat{f}_{j,n}(x) = \frac{m_j/n_j}{m\big(B(x, d_k(x))\big)} = \frac{m_j/n_j}{c_d[d_k(x)]^d}, \quad x \in \mathbb{R}^d,$$

jota voidaan motivoida aivan kuten estimaattoria $\hat{f}_n(\cdot; k)$ luvussa 6.1.4. Ottamalla vielä $\widehat{P}_{j,n} = n_j/n$ saadaan silloin luokitin

$$g_n(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}} \left\{ \frac{n_j}{n} \cdot \frac{m_j/n_j}{c_d[d_k(x)]^d} \right\} = \underset{j=1,\ldots,c}{\operatorname{argmax}} m_j,$$

joka luokittelee x:n siihen luokkaan, jonka edustajia on eniten pallossa $B\big(x, d_k(x)\big)$. Tämä on ns. *äänestys-k-lähinaapuriluokitin*. Tasapelitilanteet on määriteltävä erikseen. Voidaan esimerkiksi sopia, että tasapeliluokista voittaa se, jonka m_j:s vektori on lähinnä x:ää.

6.3 Parametrittomien luokittimien teoreettisia ominaisuuksia

Tarkastellaan yleistä parametritonta luokitinta (6.15),

$$g_n(x) = \operatorname*{argmax}_{j=1,\ldots,c} \widehat{P}_{j,n}\hat{f}_{j,n}(x), \qquad x \in \mathbb{R}^d.$$

Luokitin g_n itseasiassa riippuu opetusaineistosta $(X_1, J_1), \ldots, (X_n, J_n)$, joten se on satunnaisfunktio. Siten luokittimen g_n ehdollinen virhetodennäköisyys

$$e(g_n) = \mathbb{P}\big(g_n(X) \neq J | (X_1, J_1), \ldots, (X_n, J_n)\big)$$

on satunnaismuuttuja, joka riippuu opetusaineistosta. Ottamalla odotusarvo opetusaineiston suhteen saadaan tällaisen satunnaisen luokittimen virhetodennäköisyys,

$$\mathbb{E}e(g_n) = \mathbb{P}\big(g_n(X) \neq J\big).$$

Voidaan osoittaa, että hyvin yleisillä oletuksilla pätee sekoiteotantaan perustuville ydin- ja k-lähinaapuriluokittimelle (sopivilla (h_n) tai (k_n)), että

$$\begin{aligned} &e(g_n) \xrightarrow[n\to\infty]{} e(g^*) \quad \text{todennäköisyydellä 1,} \\ &\mathbb{E}e(g_n) \xrightarrow[n\to\infty]{} e(g^*), \end{aligned}$$

missä $e(g^*)$ on Bayesin luokittimen virhetodennäköisyys.

Todistetaan ensin seuraava aputulos.

Lemma 6.6. *Luokittimelle g_n pätee todennäköisyydellä 1, että*

$$0 \leq e(g_n) - e(g^*) \leq \sum_{j=1}^{c} \int_{\mathbb{R}^d} \left| P_j \, f_j(x) - \widehat{P}_{j,n} \, \hat{f}_{j,n}(x) \right| dx.$$

Todistus. (Hahmotelma)

Aivan kuten Lauseessa 3.10, voidaan nytkin osoittaa, että

$$e(g_n) = 1 - \int_{\mathbb{R}^d} P_{g_n(x)} f_{g_n(x)}(x) dx$$

(vrt. (3.10)). Ero on siinä, että nyt ovat kyseessä satunnaismuuttujat ja yllä oleva kaava itseasiassa pätee (vain) todennäköisyydellä 1. Edelleen, kuten lauseessa 3.10,

$$e(g^*) = 1 - \int_{\mathbb{R}^d} P_{g^*(x)} f_{g^*(x)}(x) dx.$$

Siten

$$e(g_n) - e(g^*) = \int_{\mathbb{R}^d} \big(P_{g^*(x)} f_{g^*(x)}(x) - P_{g_n(x)} f_{g_n(x)}(x) \big) dx \tag{6.16}$$

(todennäköisyydellä 1). Koska $j = g^*(x)$ maksimoi tulon $P_j f_j(x)$, on yllä oleva integroitava ≥ 0, joten

$$e(g_n) - e(g^*) \geq 0 \qquad \text{(todennäköisyydellä 1).}$$

Jos $g^*(x) = g_n(x)$, niin integrandi häviää. Jos taas $g^*(x) \neq g_n(x)$, saadaan

$$\begin{aligned}
&P_{g^*(x)} f_{g^*(x)}(x) - P_{g_n(x)} f_{g_n(x)}(x) \\
&= \big(P_{g^*(x)} f_{g^*(x)}(x) - \widehat{P}_{g_n(x),n} \hat{f}_{g_n(x),n}(x) \big) \\
&\qquad + \big(\widehat{P}_{g_n(x),n} \hat{f}_{g_n(x),n}(x) - P_{g_n(x)} f_{g_n(x)}(x) \big) \\
&\leq \big(P_{g^*(x)} f_{g^*(x)}(x) - \widehat{P}_{g^*(x),n} \hat{f}_{g^*(x),n}(x) \big) \\
&\qquad + \big(\widehat{P}_{g_n(x),n} \hat{f}_{g_n(x),n}(x) - P_{g_n(x)} f_{g_n(x)}(x) \big),
\end{aligned}$$

sillä $j = g_n(x)$ maksimoi tulon $\widehat{P}_{j,n}\hat{f}_{j,n}(x)$. Siten kaavan (6.16) integrandille saadaan yläarvio

$$\begin{aligned}
&P_{g^*(x)}f_{g^*(x)}(x) - P_{g_n(x)}f_{g_n(x)}(x) \\
&\leq \left|P_{g^*(x)}f_{g^*(x)}(x) - \widehat{P}_{g^*(x),n}\hat{f}_{g^*(x),n}(x)\right| \\
&\qquad\qquad + \left|\widehat{P}_{g_n(x),n}\hat{f}_{g_n(x),n}(x) - P_{g_n(x)}f_{g_n(x)}(x)\right| \\
&\leq \sum_{j=1}^{c} \left|P_j f_j(x) - \widehat{P}_{j,n}\hat{f}_{j,n}(x)\right|,
\end{aligned}$$

(todennäköisyydellä 1) missä viimeinen epäyhtälö seuraa ehdosta $g^*(x) \neq g_n(x)$. Tästä seuraa väite. □

Olkoon nyt $\hat{f}_{j,n}$ ydinestimaattori sellaisella siloitusparametrilla h_n, että $h_n \to 0$ ja $nh_n^d \to \infty$, kun $n \to \infty$. Luvun 6.1.2 mukaan on tällöin todennäköisyydellä 1

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} \left|\hat{f}_{j,n}(x) - f_j(x)\right| dx = 0$$

ja

$$\lim_{n\to\infty} \mathbb{E} \int_{\mathbb{R}^d} \left|\hat{f}_{j,n}(x) - f_j(x)\right| dx = 0.$$

Jos $\widehat{P}_{j,n} \to P_j$ todennäköisyydellä 1 (esimerkiksi $\widehat{P}_{j,n} = n_j/n$ sekoiteotannassa), on lemman 6.6 nojalla tällöin

$$
\begin{aligned}
0 \leq e(g_n) - e(g^*) &\leq \sum_{j=1}^{c} \int_{\mathbb{R}^d} \left| P_j f_j(x) - \widehat{P}_{j,n} \hat{f}_{j,n}(x) \right| dx \\
&\leq \sum_{j=1}^{c} \int_{\mathbb{R}^d} \left| P_j f_j(x) - P_j \hat{f}_{j,n}(x) \right| dx \\
&\qquad + \sum_{j=1}^{c} \int_{\mathbb{R}^d} \left| P_j \hat{f}_{j,n}(x) - \widehat{P}_{j,n} \hat{f}_{j,n}(x) \right| dx \\
&= \sum_{j=1}^{c} P_j \int_{\mathbb{R}^d} \left| f_j(x) - \hat{f}_{j,n}(x) \right| dx + \sum_{j=1}^{c} \left| P_j - \widehat{P}_{j,n} \right| \int_{\mathbb{R}^d} \hat{f}_{j,n}(x) dx \\
&\leq \sum_{j=1}^{c} \int_{\mathbb{R}^d} \left| f_j(x) - \hat{f}_{j,n}(x) \right| dx + \sum_{j=1}^{c} \left| P_j - \widehat{P}_{j,n} \right| \to 0
\end{aligned}
$$

todennäköisyydellä 1, kun $n \to \infty$. Siten, todennäköisyydellä 1,

$$
e(g_n) \xrightarrow[n\to\infty]{} e(g^*)
$$

ja tästä voidaan osoittaa seuraavan, että myös

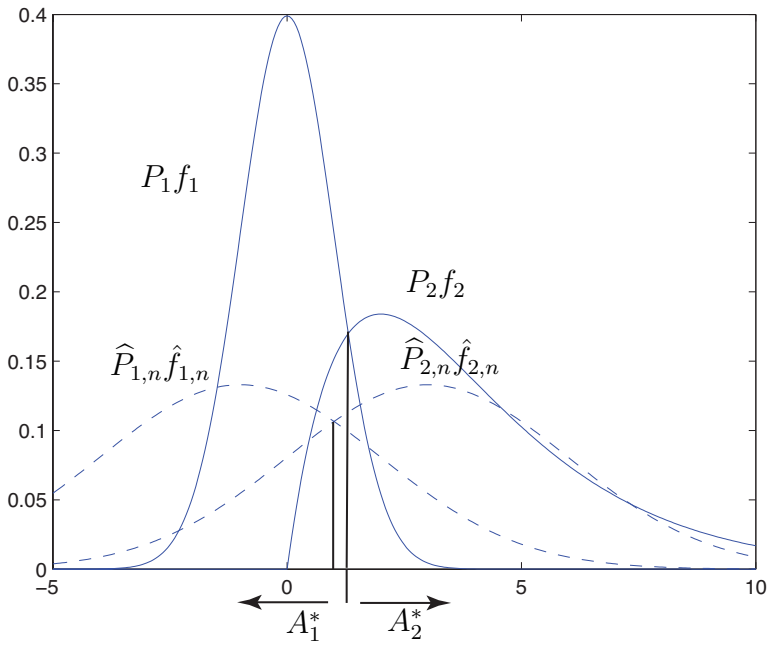
$$
\mathbb{E}e(g_n) \xrightarrow[n\to\infty]{} e(g^*).
$$

Vastaavat tulokset pätevät äänestys-k-lähinaapuriluokittimelle, jos $k_n \to \infty$, $k_n/n \to 0$, kun $n \to \infty$.

Nämä tulokset ovat tietysti varsin luontevia: jos pätee suurella tarkkuudella, että

$$
\hat{f}_{j,n} \approx f_j \quad \text{ja} \quad \widehat{P}_{j,n} \approx P_j, \tag{6.17}
$$

niin tietysti $g_n \approx g^*$ ja siis $e(g_n) \approx e(g^*)$. On kuitenkin syytä huomauttaa, että pienen virheen $e(g_n)$ saamiseksi ei kuitenkaan ole *välttämätöntä*, että (6.17) pätee sillä ratkaisevaa on vain päätöspinnan oikea paikka (kuva 6.11).



Kuva 6.11: Myös harhaiset tiheysfunktion estimaatit voivat johtaa hyvään luokittimeen, koska ratkaisevaa on vain päätöspinnan oikea paikka.

Luku 7

Regression käyttöön perustuvia parametrittomia luokittimia

Edellisessä luvussa estimoitiin Bayesin luokitinta

$$g^*(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}}\ P_j f_j(x), \qquad x \in \mathbb{R}^d,$$

estimoimalla P_j:t ja f_j:t. Bayesin luokitin voitiin määritellä kuitenkin myös suoraan posterioritodennäköisyyksien avulla

$$g^*(x) = \underset{j=1,\ldots,c}{\operatorname{argmax}}\ P(j|x), \qquad x \in \mathbb{R}^d.$$

Tässä luvussa selvitämmekin lyhyesti miten todennäköisyydet $P(j|x)$ voidaan estimoida parametrittomasti.

Olkoon tavalliseen tapaan (X, J) luokkainformaation sisältävä piirrevektori ja merkitään luokan j indikaattoria $Y_j = 1_{\{J=j\}}$, $j = 1, \ldots, c$,

$$Y_j = \begin{cases} 1, & \text{kun } J = j, \\ 0, & \text{muulloin.} \end{cases}$$

Silloin

$$\mathbb{E}(Y_j|X = x) = 1 \cdot P(J = j|x) + 0 \cdot P(J \neq j|x) = P(j|x)$$

eli siis

$$P(j|x) = \mathbb{E}(Y_j|X = x).$$

Tässä $\mathbb{E}(Y_j|X = x) = P(j|x)$ on Y_j:n *regressiofunktio* X:n suhteen ja luvussa 5.3 näimme, että sitä voi yrittää estimoida pienimmän neliösumman menetelmällä. Nyt kuitenkin lähestymme tätä estimointitehtävää suoremmin, *lokaalin keskiarvoistamisen* kautta. Olkoon $(X_1, J_1), \ldots, (X_n, J_n)$ opetusaineisto ja

$$Y_{ji} = 1_{\{J_i=j\}} = \begin{cases} 1, & \text{kun } J_i = j, \\ 0, & \text{muulloin.} \end{cases}$$

Tällöin $Y_{j1}, \ldots, Y_{jn} \sim Y_j$ ja

$$\mathbb{E}(Y_j|X = x) \approx \sum_{i=1}^{n} W(x - X_i; h) Y_{ji}, \tag{7.1}$$

jos painot $W(x - X_i; h) \geq 0$ toteuttavat ehdot

$$\sum_{i=1}^{n} W(x - X_i; h) = 1 \tag{7.2}$$

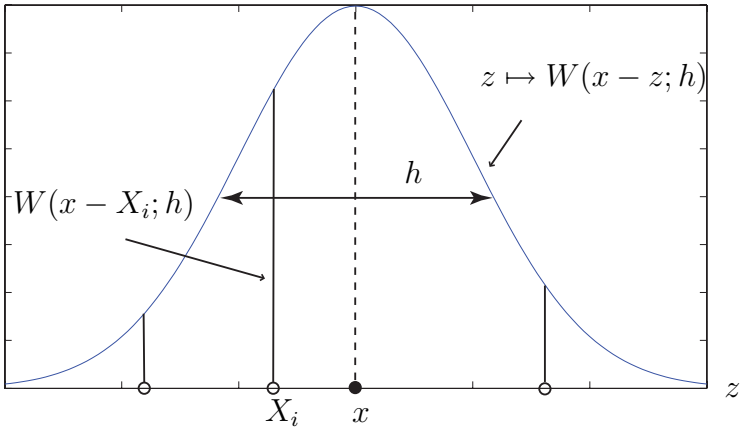
ja

$$W(x - X_i; h) \approx 0, \tag{7.3}$$

kun X_i on kaukana x pisteestä x. Tässä h on painofunktion leveyttä säätelevä parametri (kuva 7.1). Ajatus menetelmässä on seuraavanlainen. Odotusarvo $\mathbb{E}(Y_j|X = x)$ voidaan periaatteessa estimoida laskemalla keskiarvo suuresta otoksesta satunnaismuuttujan Y_j arvoja, joita vastaava piirrevektorin arvo on täsmälleen x. Nyt kuitenkaan pisteeseen x ei luultavasti satu yhtäkään opetusvektoria X_i, joten lasketaan sen sijaan keskiarvo x:n lähelle sattuvien piirrevektorien avulla. Näin siis (7.1) estimoi odotusarvoa $\mathbb{E}(Y_j|X = x)$ *painotetulla lokaalilla keskiarvolla*.

Selvästi (7.1):ssä on

$$\sum_{i=1}^{n} W(x - X_i; h) Y_{ji} = \sum_{i=1}^{n} W(x - X_i; h) 1_{\{J_i=j\}} = \sum_{i=1}^{n_j} W(x - X_{ji}; h),$$



Kuva 7.1: Parametriton regressio voidaan tulkita lokaalina keskiarvona painofunktiolla W.

kun $X_{j1}, \ldots, X_{jn_j}$ ovat luokasta j peräisin olevat opetusvektorit. Jos nyt K on luvun 6.1.2 tyyppinen ydinfunktio, esimerkiksi Gaussin ydin, ja

$$W(x - X_i; h) = \frac{K_h(x - X_i)}{\sum_{l=1}^{n} K_h(x - X_l)},$$

niin silloin (7.2) ja (7.3) pätevät ja estimaattorimme posterioritodennäköisyyksille $P(j|x)$ on

$$\begin{aligned}
\widehat{P}(j|x) &= \sum_{i=1}^{n_j} W(x - X_{ji}; h) \\
&= \sum_{i=1}^{n_j} \left[\frac{K_h(x - X_{ji})}{\sum_{l=1}^{n} K_h(x - X_l)} \right] \\
&= \frac{\sum_{i=1}^{n_j} K_h(x - X_{ji})}{\sum_{l=1}^{n} K_h(x - X_l)}.
\end{aligned}$$

Siten

$$\widehat{P}(j|x) = \frac{\frac{n_j}{n} \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} K_h(x - X_{ji})}{\frac{1}{n} \sum_{l=1}^{n} K_h(x - X_l)} = \frac{\widehat{P}_{j,n} \hat{f}_{j,n}(x)}{\hat{f}_n(x)},$$

eli kaavassa (vertaa (3.5))

$$P(j|x) = \frac{P_j f_j(x)}{f(x)}$$

on f_j ja f estimoitu ydinestimaattoreilla. Näin saamme itseasiassa luvun 6.2 ydinestimaattoriin perustuvan luokittimen.

Voidaan osoittaa (HT), että $\widehat{P}(j|x)$ saadaan myös lukuna $\bar{a}_j = \bar{a}_j(x)$, joka minimoi a_j:n suhteen neliösumman

$$\sum_{i=1}^{n} (a_j - Y_{ji})^2 K_h(x - X_i),$$

missä $Y_{ji} = 1_{\{J_i=j\}}$ kuten edellä. Tämän yleistys on korvata vakio a_j ensimmäisen asteen polynomilla $z \mapsto a_j + b_j^T(x-z)$ ja etsiä $\bar{a}_j = \bar{a}_j(x)$ ja $\bar{b}_j = \bar{b}_j(x)$, jotka minimoivat (a_j, b_j):n suhteen neliösumman

$$\sum_{i=1}^{n} (a_j + b_j^T(x - X_i) - Y_{ji})^2 K_h(x - X_i)$$

ja ottaa sitten $\widehat{P}(j|x) = \bar{a}_j(x)$. Tätä sanotaan *lokaaliseksi lineaariseksi regressioksi* ja voidaan osoittaa, että myös tällöin $\widehat{P}(j|x)$ on muotoa

$$\widehat{P}(j|x) = \bar{a}_j(x) = \sum_{i=1}^{n_j} W(x - X_{ji}; h),$$

missä

$$\sum_{i=1}^{n} W(x - X_i; h) = 1,$$

mutta nyt ei välttämättä ole $W(x - X_i; h) \geq 0$, $i = 1, \ldots, n$.

Huomautus. Luvuissa 6 ja 7 keskityttiin erilaisiin ydinestimointimenetelmiin mutta on olemassa myös monia muita parametrittomia tiheysfunktion ja regressiofunktion estimointimenetelmiä. Esimerkkejä ovat silottavat splinit ja ortogonaalisarjakehitelmät, erikoistapauksena aallokkeet (waveletit).

Luku 8

Luokitteluvirheen estimointi

8.1 Yleistä

Olkoon opetusaineisto $(X_1, J_1), \ldots, (X_n, J_n)$ ja g_n jokin siihen perustuva luokitin. Oikeastaan siis g_n on satunnaisfunktio (vertaa luku 6.2) mutta pidämme tässä luvussa opetusaineistoa *kiinteänä*, jolloin g_n on tavallinen (deterministinen) luokitin. Matemaattisesti tämä tarkoittaa, että erilaiset todennäköisyydet lasketaan *ehdolla* $(X_1, J_1), \ldots, (X_n, J_n)$. Esimerkiksi g_n:n virhetodennäköisyys on itseasiassa ehdollinen todennäköisyys

$$e(g_n) = \mathbb{P}\big(g_n(X) \neq J \mid (X_1, J_1), \ldots, (X_n, J_n)\big).$$

Jätämme opetusaineiston pois merkinnöistä ja merkitsemme myös $g_n = g$.

Oletetaan ensin, että meillä on opetusaineiston lisäksi käytössä myös otos

$$(X_{n+i}, J_{n+i}), \qquad i = 1, 2, \ldots$$

(X, J):n jakaumasta. Silloin voidaan määritellä luokitteluvirheen estimaattori

$$\hat{e}_m(g) = \frac{1}{m} \sum_{i=1}^{m} 1\{g(X_{n+i}) \neq J_{n+i}\}.$$

Tässä merkitään typografisista syistä indikaattoria

$$1_{\{g(X_{n+i}) \neq J_{n+i}\}} = 1\{g(X_{n+i}) \neq J_{n+i}\}.$$

Siis $1\{g(X_{n+i}) \neq J_{n+i}\}$ on satunnaismuuttuja, jolle

$$1\{g(X_{n+i}) \neq J_{n+i}\} = \begin{cases} 1, & \text{kun } g(X_{n+i}) \neq J_{n+i}, \\ 0, & \text{kun } g(X_{n+i}) = J_{n+i}. \end{cases}$$

Koska $(X_{n+i}, J_{n+i}) \sim (X, J)$, on nyt

$$\mathbb{E}\big[1\{g(X_{n+i}) \neq J_{n+i}\}\big] = \mathbb{E}\big[1\{g(X) \neq J\}\big] = \mathbb{P}\big(g(X) \neq J\big) = e(g).$$

Siten $\hat{e}_m(g)$ on virhetodennäköisyyden $e(g)$ *harhaton* estimaattori, eli

$$\mathbb{E}\big[\hat{e}_m(g)\big] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[1\{g(X_{n+i}) \neq J_{n+i}\}\right] = \frac{1}{m} m e(g) = e(g).$$

Toisaalta, suurten lukujen lain nojalla

$$\lim_{m \to \infty} \hat{e}_m(g) = e(g)$$

todennäköisyydellä 1.

Siten $\hat{e}_m$ olisi oikein sopiva $e(g)$:n estimaattori, mutta käytännössä meillä kuitenkin on käytössä ainoastaan *äärellinen* opetusaineisto

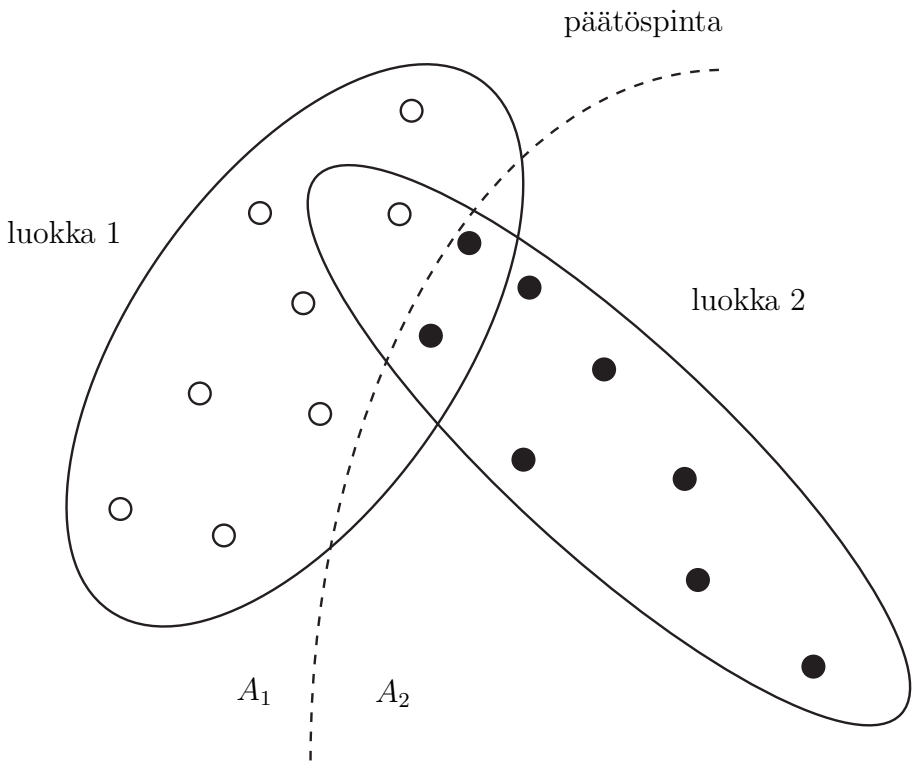
$$(X_1, J_1), \ldots, (X_N, J_N),$$

jonka avulla tulee voida *sekä* muodostaa luokitin g *että* estimoida sen virhetodennäköisyys.

Seuraavissa luvuissa tarkastellaan eri tapoja suoriutua tästä tehtävästä.

8.2 Takaisinsijoitusestimaattori

Luvun otsikon ”takaisinsijoitus” on suomennos englanninkielen termistä ”resubstitution”, jonka lyhennämme seuraavssa ”RS”. Tässä menetelmässä koko



Kuva 8.1: Kuvan esimerkissä luokan 1 ja 2 opetusvektoreita on merkitty vastaavasti symboleilla ∘ ja •. Nyt luokitteluvirheen takaisinsijoitusestimaatti $\hat{e}_N^{\mathrm{RS}}(g) = 0$ vaikka selvästi todellinen luokittelvirhe $e(g) > 0$.

opetusaineisto käytetään luokittimen konstruointiin ($N = n$) ja virhe estimoidaan myös koko aineiston avulla,

$$\hat{e}_N^{\mathrm{RS}}(g) = \frac{1}{N} \sum_{i=1}^{N} 1\{g(X_i) \neq J_i\}.$$

Tämä estimaatti yleensä antaa *liian optimistisen* tuloksen (liian pienen arvion $e(g)$:lle) ja sitä *ei* itse asiassa pitäisi käyttää ollenkaan. Selitys estimaattorin ylioptimistisuudelle on se, että luokitin voi mukautua liian hyvin opetusaineistoon ja siksi toimia huonosti uusille (X_{N+i}, J_{N+i}), $i = 1, 2, \ldots$ (kuva 8.1).

8.3 Pidätysestimaattori

Jaetaan nyt koko opetusaineisto erilliseen *opetus-* ja *testijoukkoon*,

opetus:	$(X_1, J_1), \ldots, (X_n, J_n)$
testi:	$(X_{n+1}, J_{n+1}), \ldots, (X_N, J_N)$

Sitten g konstruoidaan opetusjoukon avulla ja luokitteluvirhe estimoidaan estimaattorilla (pidätys = ”hold out”, HO)

$$\hat{e}_m^{\mathrm{HO}}(g) = \frac{1}{m} \sum_{i=1}^{m} 1\{g(X_{n+i}) \neq J_{n+i}\},$$

missä $m = N - n$. Idea on se, että testauksessa käytetty aineisto on silloin riippumaton luokittimen konstruoinnissa käytetystä opetusaineistosta ja näin vältetään takaisinsijoitusestimaattorin ylioptimistisuus. Selvästi

$$\mathbb{E}\left[\hat{e}_m^{\mathrm{HO}}(g)\right] = e(g), \tag{8.1}$$

koska $\mathbb{E}[1\{g(X_{n+i}) \neq J_{n+i}\}] = \mathbb{P}\big(g(X) \neq J\big)$. Kuten jo luvun alussa todettiin, tässä opetusaineisto ajatellaan kiinnitetyksi.

8.4 Kierrätysestimaattori

Pidätysestimaattorin huono puoli on se, että osaa koko aineistosta ei päästä käyttämään luokittimen konstruoinnissa. Oletetaan siis, että luokitin g on konstruoitu koko aineiston

$$(X_1, J_1), \ldots, (X_N, J_N)$$

avulla. Kierrätysestimaattorissa (”rotation”, RO) g:n virhetodennäköisyyttä estimoidaan jakamalla koko aineisto r:ään likimain yhtäsuureen osaan osit-

tamalla $\{1, \ldots, N\}$,

$$\{1, \ldots, N\} = \bigcup_{k=1}^{r} I_k, \qquad I_k \cap I_l = \emptyset, \text{ kun } k \neq l.$$

Sitten kullakin $k \in \{1, \ldots, r\}$:

(i) Konstruoidaan luokitin g_k opetusjoukon

$$\left\{(X_i, J_i) \mid i \in \bigcup_{l \neq k} I_l\right\}$$

avulla.

(ii) Estimoidaan g_k:n virhetodennäköisyys pidätysestimaattorilla

$$\hat{e}^{\mathrm{HO}}_{|I_k|}(g_k) = \frac{1}{|I_k|} \sum_{i \in I_k} 1\{g_k(X_i) \neq J_i\}.$$

(iii) Otetaan $e(g)$:n estimaattoriksi

$$\hat{e}^{\mathrm{RO}}_r = \frac{1}{r} \sum_{k=1}^{r} \hat{e}^{\mathrm{HO}}_{|I_k|}(g_k).$$

Ideana on, että jos $|I_k| \ll N$ kaikilla k, niin luultavasti $g_k \approx g$, jolloin

$$\hat{e}^{\mathrm{HO}}_{|I_k|}(g_k) \approx e(g).$$

Ja lisäksi pidätysestimaattorissa $\hat{e}^{\mathrm{HO}}_{|I_k|}(g_k)$ on *erilliset* opetus- ja testijoukot.

Äärimmäinen ja usein käytetty kierrätysestimaattori on *leave-one-out* menetelmä, jolle

$$I_k = \{k\}, \qquad k = 1, \ldots, N,$$

ja jossa siis yksi pari (X_i, J_i) kerrallaan jätetään pois. Kierrätysestimaattoreita kutsutaan usein myös *ristiinvalidointiestimaattoreiksi* (cross-validation).

8.5 Virherajoista

Tarkastellaan pidätysestimaattoria. Nyt

$$m\hat{e}_m^{\mathrm{HO}}(g) = \sum_{i=1}^{m} 1\{g(X_{n+i}) \neq J_{n+i}\}.$$

Tässä $1\{g(X_{n+i}) \neq J_{n+i}\} = 1$ todennäköisyydellä

$$\mathbb{P}\big(g(X_{n+i}) \neq J_{n+i}\big) = \mathbb{P}\big(g(X) \neq J\big) = e(g).$$

Siten

$$m\hat{e}_m^{\mathrm{HO}}(g) \sim \mathrm{Bin}\big(m, e(g)\big).$$

Näin ollen

$$\begin{aligned} m\mathbb{E}\left[\hat{e}_m^{\mathrm{HO}}(g)\right] &= me(g), \\ m^2 D^2\left[\hat{e}_m^{\mathrm{HO}}(g)\right] &= me(g)\big(1 - e(g)\big), \end{aligned}$$

josta edelleen saadaan

$$\begin{aligned} \mathbb{E}\left[\hat{e}_m^{\mathrm{HO}}(g)\right] &= e(g), \\ D^2\left[\hat{e}_m^{\mathrm{HO}}(g)\right] &= e(g)(1 - e(g))/m \end{aligned}$$

(vrt. (8.1)). Kun m on suuri, on keskeisen raja-arvolauseen nojalla likimain

$$Z_m = \frac{\hat{e}_m^{\mathrm{HO}}(g) - e(g)}{\sqrt{e(g)(1 - e(g))/m}} \sim \mathrm{N}(0, 1),$$

joten likimain pätee

$$\mathbb{P}(-1.96 \leq Z_m \leq 1.96) = 0.95.$$

Siten likimain (pienen laskun jälkeen)

$$\mathbb{P}\left(\hat{e} - 1.96\sqrt{\frac{\hat{e}(1-\hat{e})}{m}} \leq e(g) \leq \hat{e} + 1.96\sqrt{\frac{\hat{e}(1-\hat{e})}{m}}\right) = 0.95,$$

missä $\hat{e} = \hat{e}_m^{\mathrm{HO}}(g)$. Näin $e(g)$:lle saadaan 95% *luottamusväli*

$$\left[\hat{e} - 1.96\sqrt{\frac{\hat{e}(1-\hat{e})}{m}}, \hat{e} + 1.96\sqrt{\frac{\hat{e}(1-\hat{e})}{m}}\right].$$

8.6 Kahden luokittimen vertailu

Oletetaan, että luokittimet g_1 ja g_2 on konstruoitu opetusjoukosta $(X_1, J_1), \ldots, (X_n, J_n)$ ja vertaillaan niiden suorituskykyä testijoukon $(X_{n+1}, J_{n+1}, \ldots, X_{n+m}, J_{n+m})$ avulla. Onko $e(g_1) > e(g_2)$ tai $e(g_1) < e(g_2)$?

Määritellään

$$U_i = 1\{g_1(X_{n+i}) \neq J_{n+i}\} - 1\{g_2(X_{n+i}) \neq J_{n+i}\},$$

$i = 1, \ldots, m$. Silloin

$$\mathbb{E}U_i = e(g_1) - e(g_2) \qquad i = 1, \ldots, m.$$

Kuten aikaisemminkin, tässä opetusjoukko ajatellaan kiinnitetyksi. Jos siis

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} U_i,$$

niin myös pätee

$$\mathbb{E}\bar{U} = e(g_1) - e(g_2).$$

Olkoon $D^2(U_i) = \sigma^2$, jolloin

$$D^2(\bar{U}) = \frac{1}{m^2} \cdot m\sigma^2 = \frac{\sigma^2}{m}.$$

Siten, kun m on suuri, on keskeisen raja-arvolauseen nojalla likimäärin

$$\frac{\bar{U} - [e(g_1) - e(g_2)]}{\sigma/\sqrt{m}} \sim \mathrm{N}(0, 1).$$

On siis likimäärin voimassa

$$\mathbb{P}\left(-1.96 \leq \frac{\bar{U} - [e(g_1) - e(g_2)]}{\sigma/\sqrt{m}} \leq 1.96\right) = 0.95.$$

Korvataan vielä tuntematon σ^2 otosestimaatillaan

$$s^2 = \frac{1}{m} \sum_{i=1}^{m} (U_i - \bar{U})^2$$

jolloin likimain

$$\mathbb{P}\left(\bar{U} - 1.96\frac{s}{\sqrt{m}} \leq e(g_1) - e(g_2) \leq \bar{U} + 1.96\frac{s}{\sqrt{m}}\right) = 0.95.$$

Siten:

- Jos $\bar{U} + 1.96\frac{s}{\sqrt{m}} < 0$, niin luultavasti $e(g_1) < e(g_2)$.
- Jos $\bar{U} - 1.96\frac{s}{\sqrt{m}} > 0$, niin luultavasti $e(g_1) > e(g_2)$.
- Jos $U - 1.96\frac{s}{\sqrt{m}} \leq 0 \leq U + 1.96\frac{s}{\sqrt{m}}$, niin g_1:n ja g_2:n suorituskykyjen välillä ei luultavasti ole eroa – ainakaan mahdollisuutta $g_1 = g_2$ ei voi hylätä tällä luottamustasolla.

Nämä päätelmät pätevät 95% luottamustasolla.

8.7 Silotusparametrin valinta luokittimessa

Ydinestimointia käyttävässä luokittimessa on valittava silotusparametri h. Lähinaapuriluokittimessa puolestaan on kiinnitettävä tarkasteltavien naapurien lukumäärä k. Näiden valintojen teko on osa luokittimen konstruointia ja siten se on perustettava käytettävissä olevaan opetusaineistoon. Ydinestimoinnin tapauksessa voidaan esimerkiksi menetellä seuraavasti.

(*i*) Valitaan kokeiltavat arvot $\{h_1, \ldots, h_s\}$.

(*ii*) Kullakin h_l muodostetaan luokitin g_{h_l} ja lasketaan virhe-estimaatti $\hat{e}(g_{h_l})$ esimerkiksi opetusjoukon kierrätysestimaattorilla.

(*iii*) Valitaan silotusparametriksi h se h_l, jolla $\hat{e}(g_{h_l})$ on pienin.

Aivan vastaavasti voidaan tietysti lähinaapuriluokittimen tapauksessa valita kokeiltavista arvoista $\{k_1, \ldots, k_s\}$ paras.

Luku 9

Piirteiden irrotus

9.1 Yleistä

Hahmontunnistus *ei* ole pelkkää luokittelua. Ennen kuin luokinta päästään soveltamaan, joudutaan tehtyä alkuperäistä, ”raakaa” mittausta tavallisesti esikäsittelemään jollain tavalla, esimerkiksi häiriöitä (kohinaa) poistetaan sopivalla menetelmällä. Sen jälkeen esikäsitellystä mittauksesta muodostetaan varsinainen luokiteltava piirrevektori soveltamalla jotain *piirteenirrotusmenetelmää*. Tilanne on siis alla olevan kaavion kaltainen, jossa z on aluperäinen mittaus, $y \in \mathbb{R}^D$ on esikäsitelty mittaus, $x = \varphi(y) \in \mathbb{R}^d$ piirrevektori ja j luokka.

Piirteiden irrotukseen on useita syitä. Siinä missä luokittimena useimmiten käytetään jotain edellisissä luvuissa esitettyä standardimentelmää, voidaan

piirteenirrottimella esimerkiksi päästä hyödyntämään sovelluskohtaista tietoa mittausten alkuperästä. Hyvä piirteenirrotin onkin usein tärkeämpää kuin tietyn luokittimen käyttö. Tavallisesti piirteenirrottimessa $\varphi : \mathbb{R}^D \to \mathbb{R}^d$ on myös $d \ll D$, jolla helpotetaan luokittimen estimointitehtävää. Estimointi nimittäin on helpompaa matalaulotteisessa avaruudessa, kun käytössä kuitenkin on vain äärellinen opetusjoukko. Dimension pienenemisestä saatava hyöty estimoinnissa on usein suurempi kuin tällaisessa piirteenirrotuksessa mahdollisesti menetettävä informaatio. Joskus on myös mahdollista hyvällä piirteenirrottajalla päästä tilanteeseen, jossa yksinkertainenkin luokitin toimii hyvin (kuva 9.1). Bayesin luokittimen virhetodennäköisyys *ei* voi kuitenkaan pienentyä piirteenirrotuksessa, kuten seuraava lause osoittaa.

Lause 9.1. *Olkoon $\varphi : \mathbb{R}^D \to \mathbb{R}^d$ piirteenirrotin, $X = \varphi(Y)$, G^* satunnaisvektoriin Y ja g^* satunnaisvektoriin X liittyvä Bayesin luokitin. Tällöin*

$$e(g^*) \geq e(G^*).$$

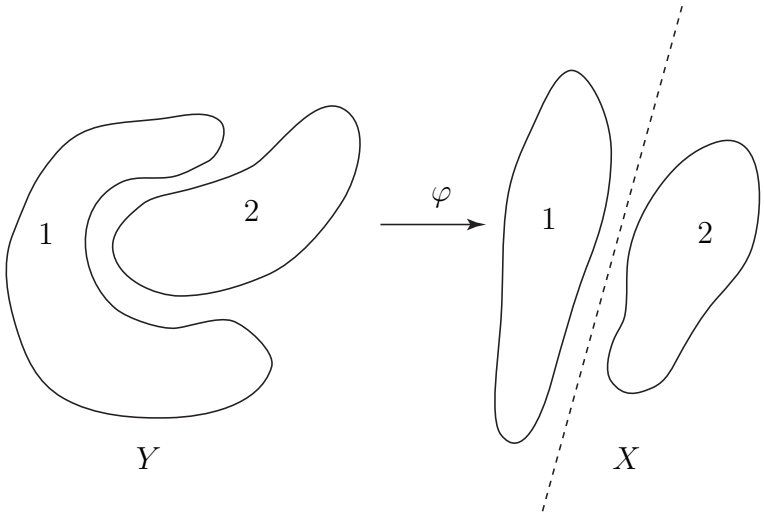
Todistus. Koska $g^*(X) = g^*(\varphi(Y)) = (g^* \circ \varphi)(Y)$, on

$$\begin{aligned} e(g^*) &= \mathbb{P}\big(g^*(X) \neq J\big) = \mathbb{P}\big((g^* \circ \varphi)(Y) \neq J\big) \\ &\geq \mathbb{P}\big(G^*(Y) \neq J\big) = e(G^*), \end{aligned}$$

sillä $g^* \circ \varphi$ on luokitin (Y, J):lle. □

Todistuksessa hyödynnettiin sitä, että Bayesin minimivirheluokitin voidaan määritellä vaikka tarkasteltavilla jakaumilla ei olisikaan tiheysfunktioita. Voidaan nimittäin aina käyttää kaavaa (3.6).

Ihannetapauksessa luokitin g ja piirteenirrotin φ tulisi optimoida *yhdessä*. Käytännössä tämä on usein liian raskasta ja piirteenirrottajaksi φ joudutaan valitsemaan jokin ”yleispätevä” menetelmä, usein sopivaan sovelluskohtaiseen muunnokseen yhdistettynä. Seuraavassa esitellään joukko tällaisia yleispäteviä menetelmiä. Ne perustuvat sellaisiin jo aikaisemmin esiintyneisiin suureisiin kuin jakauman odotusarvo (μ, μ_j), kovarianssi (Σ, Σ_j) ja luokkien



Kuva 9.1: Sopivalla piirteenirrottimella voi hankalankin tilanteen transformoida sellaiseksi, että esimerkiksi tehokas lineaarinen luokittelu tulee mahdolliseksi.

prioritodennäköisyydet P_j. Käytännössä nämä suureet joudutaan tietysti estimoimaan opetusaineiston avulla.

9.2 Separoituvuuskriteereitä

Olkoon J_φ piirteenirrottajasta φ riippuva mitta, ns. *separoitusvuuskriteeri*, piirrevektoria $X = \varphi(Y)$ vastaavien luokkien erottumiselle toisistaan. Ideana on pyrkiä valitsemaan $\varphi = \bar{\varphi}$, jolle

$$J_{\bar{\varphi}} = \max_{\varphi \in \mathcal{F}} J_\varphi, \tag{9.1}$$

eli φ, joka maksimoi kriteerin J_φ, kun $\mathcal{F}$ on jokin joukko piirteeirrotimiksi sopivia funktioita. Seuraavassa annetaan esimerkkejä eräistä mahdollisista separoituvuuskriteereistä.

9.2.1 Bayesin virhetodennäköisyys

Eräs luonteva idea on perustaa separoituvuuskriteeri on Bayesin virhetodennäköisyyteen,

$$J_\varphi = -e(g^*) = -\mathbb{P}\big((g^* \circ \varphi)(Y) \neq J\big).$$

Tämä on tietyssä mielessä ”paras” mitta luokkien erottumiselle toisistaan, mutta sitä on useinmiten mahdoton käyttää, koska J_φ:n arvoa ei osata laskea (tai estimoida).

9.2.2 Luokkien välinen etäisyys

Olkoon $X = \varphi(Y)$ ja olkoot

$$\begin{aligned}\mu_j &= \mathbb{E}(X|J=j),\\ \Sigma_j &= \mathbb{E}\left[(X-\mu_j)(X-\mu_j)^T|J=j\right]\end{aligned}$$

luokan j odotusarvo ja kovarianssimatriisi. Merkitään

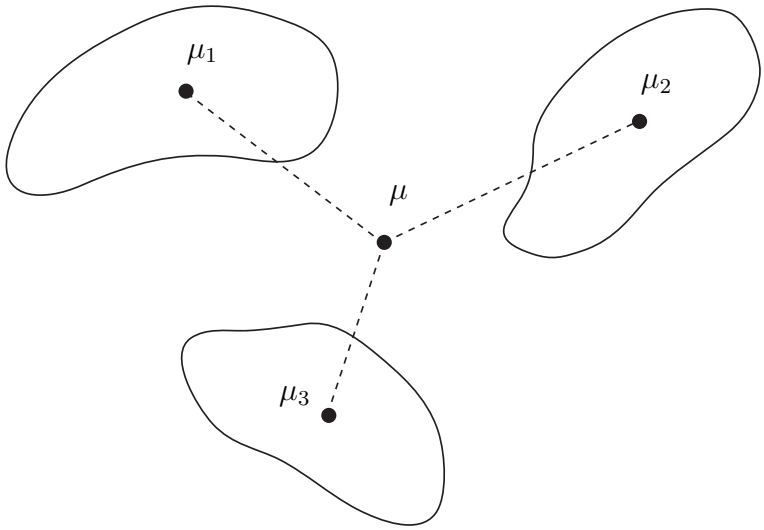
$$S_w = \sum_{j=1}^{c} P_j \Sigma_j, \tag{9.2}$$

$$S_b = \sum_{j=1}^{c} P_j (\mu_j - \mu)(\mu_j - \mu)^T. \tag{9.3}$$

Tässä

$$\mu = \sum_{j=1}^{c} P_j\, \mu_j = \mathbb{E}(X)$$

on koko X:n jakauman odotusarvo (kuva 9.2). Suure S_w on luokkien *sisäinen* (”*w*ithin”) variaatio ja S_b on vastaavasti luokkien *välinen* (”*b*etween”) variaatio.



Kuva 9.2: Satunnaisvektorin X odotusarvo μ on luokkien odotusarvojen μ_j prioritodennäköisyyksillä P_j painotettu keskiarvo.

Nyt

$$\begin{aligned}\mathrm{tr}(S_b) &= \sum_{j=1}^{c} P_j \mathrm{tr}\left[(\mu_j - \mu)(\mu_j - \mu)^T\right] \\ &= \sum_{j=1}^{c} P_j \mathrm{tr}\left[(\mu_j - \mu)^T(\mu_j - \mu)\right] = \sum_{j=1}^{c} P_j \left\|\mu_j - \mu\right\|^2,\end{aligned}$$

ja vastaavasti

$$\mathrm{tr}(S_w) = \sum_{j=1}^{c} P_j \, \mathbb{E}\big(\|X - \mu_j\|^2 \,|J = j\big).$$

Näemme, että tietyssä mielessä matriisin S_b jälki $\mathrm{tr}(S_b)$ mittaa luokkien välistä etäisyyttä ja $\mathrm{tr}(S_w)$ puolestaan mittaa luokkien sisäistä hajontaa. Siksi eräs luonteva valinta maksimoitavaksi separoituvuuskriteeriksi on

$$J_\varphi = \frac{\mathrm{tr}(S_b)}{\mathrm{tr}(S_w)}. \tag{9.4}$$

Huomaa, että tässä $X = \varphi(Y)$, joten oikea puoli todella riippuu φ:stä. Vastaavanlaisia suureisiin S_b ja S_w perustuvia kriteereitä on muitakin mutta emme käsittele niitä tässä lähemmin.

9.2.3 Estimoitu luokitteluvirhe

Olkoon piirrevektoriin $X = \varphi(Y)$ perustuva luokitin g ja asetetaan

$$J_\varphi = -\hat{e}(g),$$

missä $\hat{e}(g)$ on esimerkiksi opetusjoukkoon perustuva luokitteluvirheen kierrätysestimaatti. Tällaisella kriteerillä voidaan hakea nimenomaan luokittimeen g sopivaa piirteenirrottajaa.

9.2.4 Luokkatiheysfunktioiden välinen etäisyys

Kahden luokan tapauksessa luokkatiheysfunktioden Chernoffin etäisyyttä (harjoitus 13, tehtävä 3) voidaan käyttää myös separoituvuuskriteerinä asettamalla

$$J_\varphi = -\log \int_{\mathbb{R}^d} f_1^u \, f_2^{1-u}.$$

Multinormaalissa tapauksessa saadaan tälle myös eksplisiittinen lauseke (harjoitus 13, tehtävä 3). Tämä kriteeri soveltuu erityisesti tilanteeseen, jossa $P_1 = P_2$, koska luokkien prioritodennäköisyydet eivät vaikuta siihen. On toki myös monia muita tiheysfunktioden ”etäisyyksiä”, joita voidaan käyttä separoituvuuskriteereinä.

9.3 Piirteiden valinta

Siirrymme sitten tarkastelemaan mahdollisia yleiskäyttöisiä piirteenirrotinkuvauksia φ eli sopivia avaruuksia $\mathcal{F}$ kaavassa (9.1). Olkoon ensin $1 \leq d \leq D$ ja $\varphi : \mathbb{R}^D \to \mathbb{R}^d$ kuvaus $\varphi(y) = [y^{(i_1)}, \ldots, y^{(i_d)}]^T$, missä indeksit $1 \leq i_1 < i_2 < \cdots < i_d \leq D$ on kiinnitetty. Siis

$$X = [X^{(1)}, \ldots, X^{(d)}]^T = [Y^{(i_1)}, \ldots, Y^{(i_d)}]^T$$

on Y:n osavektori, joka on muodostettu *valitsemalla* tietyt Y komponentit.

Merkitään $I = \{i_1, \ldots, i_d\}$, $\varphi = \varphi_I$ ja $J_\varphi = J_{\varphi_I} \equiv J(I)$, missä J_φ on jokin separoituvuuskriteeri (vrt. edellinen luku). Pyrkimyksenä on sitten maksimoida $J(I)$ joukon I suhteen. Käytännön ongelmana on kombinatorinen räjähdys: mahdollisten joukkojen I lukumäärä $\binom{D}{d}$ kasvaa helposti valtavaksi.

Esimerkki 9.2.

$$\binom{20}{10} = 184756, \quad \binom{100}{10} > 10^{13}.$$

‖

Eräs ratkaisu ongelmaan on käyttää suboptimaalisia hakumenetelmiä, joissa vain osa mahdollisista joukoista I käydään optimoinnissa läpi.

Esimerkki 9.3. Valitaan $I = \{i_1, \ldots, i_d\}$ siten, että

$$J(\{i_1\}) \geq J(\{i_2\}) \geq \cdots \geq J(\{i_d\}) \geq J(\{i\})$$

kaikilla $i \notin I$. ‖

Esimerkki 9.4. Eräs suosittu suboptimaalinen hakualgoritmi on Sequential Forward Selection (SFS):

(*i*) Valitaan $I_1 = \{i_1\}$ siten että $J(\{i_1\}) \geq J(\{i\})$ kaikilla $i \neq i_1$.

(*ii*) Kun $I_k = \{i_1, \ldots, i_k\}$ on valittu, etsitään sellainen $i_{k+1} \notin I_k$, että

$$J\big(I_k \cup \{i_{k+1}\}\big) \geq J\big(I_k \cup \{i\}\big) \quad \text{kaikilla } i \notin I_k.$$

Otetaan $I_{k+1} = I_k \cup \{i_{k+1}\}$.

(*iii*) Lopetetaan, kun $k + 1 = d$ ja otetaan $\varphi = \varphi_{I_d}$.

Vastaavalla tavalla toimii Sequential Backward Selection (SBS). ‖

Esitellyistä separoituvuuskriteereistä esimerkiksi kriteeri (9.4) ja luvun 9.2.3 menetelmä sopivat hyvin SFS:n tai SBS:n kanssa käytettäväksi.

9.4 Lineaarinen piirteiden irrotus

Tarkastellaa sitten *lineaarista* piirteiden irrotusta, eli kaavassa (9.1)

$$\mathcal{F} = \{\varphi \mid \varphi : \mathbb{R}^D \to \mathbb{R}^d \text{ on lineaarinen}\}$$

ja $1 \leq d \leq D$. Olkoot tässä luvussa edelleen μ, Σ, μ_j ja Σ_j satunnaisvektoriin Y liittyvät odotusarvot ja kovarianssit, eli $\mu = \mathbb{E}(Y)$ jne. Olkoot samoin S_w ja S_b satunnaisvektoriin Y liittyvät suureet (9.2) ja (9.3).

Esimerkki 9.5. Tarkastellaan kahden luokan tapausta, $c = 2$, ja yksiulotteisen piirrevektorin muodostamista, eli $d = 1$. Nyt siis $x = \varphi(y) = a^T y$, $a \in \mathbb{R}^D$, missä a on kiinteä kuvauksen φ määräävä vektori. Silloin

$$\mathbb{E}(X) = a^T \mathbb{E}(Y) = a^T \mu, \qquad \mathbb{E}(X|J = j) = a^T \mu_j,$$

missä jälkimmäinen kaava antaa X:n odotusarvon luokassa j. Edelleen, X:n varianssi luokassa j on

$$\begin{aligned}\mathbb{E}\big[(X - a^T\mu_j)^2 | J = j\big] &= \mathbb{E}\big[\big(a^T(Y - \mu_j)\big)^2 | J = j\big] \\ &= \mathbb{E}\big[\big(a^T(Y - \mu_j)(Y - \mu_j)^T a | J = j\big] = a^T \Sigma_j a,\end{aligned}$$

koska $X = a^T Y$. Näinollen on luontevaa maksimoida a:n suhteen kriteeri

$$J_\varphi = \frac{(a^T\mu_1 - a^T\mu_2)^2}{P_1 a^T \Sigma_1 a + P_2 a^T \Sigma_2 a} = \frac{\big[a^T(\mu_1 - \mu_2)\big]^2}{a^T(P_1\Sigma_1 + P_2\Sigma_2)a} = \frac{1}{P_1 P_2} \cdot \frac{a^T S_b a}{a^T S_w a}.$$

Tämä on oleellisesti vanha tuttu *Fisherin kriteeri* (Luku 4.1). ‖

Edellisen esimerkin idea voidaan yleistää lineaarikuvaukselle $\varphi : \mathbb{R}^D \to \mathbb{R}^d$, $1 \leq d \leq D$. Silloin haemme vektoreita $a_1, \ldots, a_d \in \mathbb{R}^D$, jotka sopivalla tavalla maksimoivat kriteerin

$$J_\varphi = \frac{a^T S_b a}{a^T S_w a}.$$

Oletetaan, että S_w on positiivisesti definiitti ja olkoon $b = S_w^{1/2} a$ eli $a = S_w^{-1/2} b$. Silloin

$$J_\varphi = \frac{b^T S_w^{-1/2} S_b S_w^{-1/2} b}{b^T S_w^{-1/2} S_w S_w^{-1/2} b} = \frac{b^T S_w^{-1/2} S_b S_w^{-1/2} b}{\|b\|^2}.$$

J_φ on selvästi invariantti b:n skaalauksen suhteen, eli b:n korvaaminen λb:llä ($\lambda \in \mathbb{R}$) ei muuta sen arvoa. Siten voimme maksimoida suureen

$$J_\varphi = b^T S_w^{-1/2} S_b S_w^{-1/2} b$$

ehdolla $\|b\| = 1$. Matriisi $S_w^{-1/2} S_b S_w^{-1/2}$ on selvästi symmetrinen. Olkoon

$$S_w^{-1/2} S_b S_w^{-1/2} = U \Lambda U^T$$

sen ominaisarvohajotelma, missä $U = [u_1, \ldots, u_D]$ ja $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$, $\lambda_1 \geq \cdots \geq \lambda_D$. Silloin

$$J_\varphi = \sum_{j=1}^{D} \lambda_j (b^T u_j)^2.$$

Tässä

$$\sum_{j=1}^{D} (b^T u_j)^2 = \|b\|^2 = 1,$$

joten

$$J_\varphi \leq \lambda_1 \sum_{j=1}^{D} (b^T u_j)^2 = \lambda_1$$

ja J_φ maksimoituu esimerkiksi kun $b = u_1$ eli $a \equiv a_1 = S_w^{-1/2} u_1$.

Seuraavaksi on luontevaa valita $b = u_2$ eli $a_2 = S_w^{-1/2} u_2$ ja niin edelleen. Tällöin

$$\begin{aligned} a_i^T S_w a_i &= u_i^T S_w^{-1/2} S_w S_w^{-1/2} u_i = u_i^T u_i = 1, \quad i = 1, \ldots, D, \\ a_i^T S_w a_k &= u_i^T S_w^{-1/2} S_w S_w^{-1/2} u_k = u_i^T u_k = 0, \quad k = 1, \ldots, i-1. \end{aligned}$$

Siten $\|a_i\|_{S_w} = 1$ kaikilla i ja $a_i \perp a_{i-1}, \ldots, a_1$ normia $\|\cdot\|_{S_w}$ vastaavan sisätulon $\langle a, a' \rangle_{S_w} = a^T S_w a'$ mielessä. On helppo nähdä, että näillä ehdoilla perättäiset valinnat $a = a_i$, $i = 1, \ldots, D$ itse asiassa juuri kasvattavat J_φ:n arvoa eniten.

Edelleen,

$$\sum_{j=1}^{c} P_j (\mu_j - \mu) = \sum_{j=1}^{c} P_j \mu_j - \Big(\sum_{j=1}^{c} P_j \Big) \mu = \mu - \mu = 0,$$

joten vektorit $\mu_1 - \mu, \ldots, \mu_c - \mu$ ovat lineaarisesti riippuvat. Siten matriisin

$$S_b = \sum_{j=1}^{c} P_j(\mu_j - \mu)(\mu_j - \mu)^T$$

aste on korkeintaan $c-1$, joten korkeintaan $c-1$ ominaisarvoista λ_j ovat nollasta eroavia (positiivisia). Siten

$$J_\varphi = \sum_{j=1}^{c-1} \lambda_j (b^T u_j)^2$$

ja valinnat $b_i = u_i$, $i > c-1$, eivät itse asiassa enää kasvata J_φ:tä lainkaan.

Olkoon siis $M = \min\{d, c-1\}$. Silloin edellisen perusteella saadaan matriisista $A = [a_1, \ldots, a_M]^T$ lupaavan tuntuinen piirteenirrotin $\varphi : \mathbb{R}^D \to \mathbb{R}^d$, kun asetetaan

$$\varphi(y) = Ay.$$

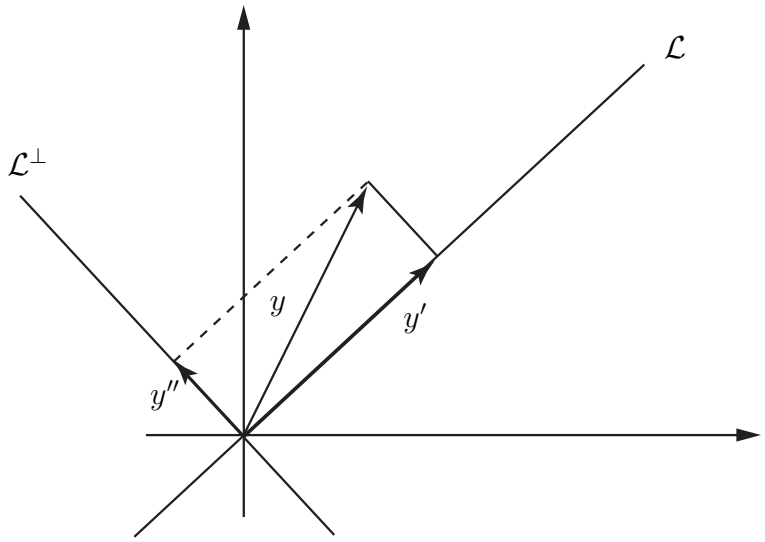
Huomataan erityisesti, että kun luokkia on vain kaksi ($c = 2$), saadaan täällä menetelmällä aikaiseksi vain yksiulotteinen piirrevektori, jonka ainoa komponentti $a_1^T y$ on idealtaan tuttu jo Fisherin luokittimesta.

9.5 Karhunen-Loève muunnos

Eräs keskeinen pyrkimys piirteenirrotuksessa on hahmovektorin dimension pienentäminen. Ehkäpä suosituin menetelmä on seuraavassa esiteltävä Karhunen-Loève muunnos.

Olkoon taas Y D-ulotteinen satunnaisvektori sekä $\mu = \mathbb{E}(Y)$ ja Σ vastaavasti Y:n odotusarvo ja kovarianssimatriisi. Olkoon edelleen $\{v_1, \ldots, v_D\} \subset \mathbb{R}^D$ ortonormaali joukko, $1 \leq d \leq D$, ja

$$\mathcal{L} = \mathrm{sp}\{v_1, \ldots, v_d\},$$
$$\mathcal{L}^\perp = \{y \in \mathbb{R}^D | y^T v_i = 0,\ i = 1, \ldots, d\}.$$



Kuva 9.3: Vektorin y jako avaruuden $\mathcal{L}$ suuntaiseen ja sitä vastaan kohtisuoraan komponenttiin.

Kun $y \in \mathbb{R}^D$, on tunnetusti olemassa sellaiset yksikäsitteiset $y' \in \mathcal{L}$, $y'' \in \mathcal{L}^\perp$, että

$$y = y' + y'',$$

missä

$$y' = \sum_{i=1}^{d} (v_i^T y) v_i$$

(kuva 9.3).

Sama hajotelma pätee silloin tietysti myös satunnaisvektoreille,

$$Y = Y' + Y'',$$

missä

$$Y' = \sum_{i=1}^{d} (v_i^T Y) v_i$$

ja $Y'' = Y - Y'$.

Merkitään

$$\mu' = \mathbb{E}(Y') = \sum_{i=1}^{d} (v_i^T \mu) v_i,$$
$$\mu'' = \mathbb{E}(Y'') = \mathbb{E}(Y) - \mathbb{E}(Y') = \mu - \mu'.$$

Tavoitteena on nyt korvata Y ortonormaalilla projektiollaan Y' avaruudelle $\mathcal{L}$ siten että tässä projektiossa menetetään mahdollismman vähän koko Y:ssä olevaa informaatiota. Onhan nimittäin niin, että projektiossa myös luokittelun kannalta tärkeää informaatiota voi hävitä. Miten $\mathcal{L}$ on valittava, jotta menetetty informaation määrä jossain mielessä minimoituu? Yksi melko luonteva idea on minimoida $\mathbb{E}\left(\|Y'' - \mu''\|^2\right)$. Jos nimittäin $\mathbb{E}\left(\|Y'' - \mu''\|^2\right) \approx 0$, niin $Y'' \approx \mu''$ on likimain vakio, jolloin Y saadaan Y':sta likimain yksinkertaisella siirrolla: $Y \approx Y' + \mu''$. Silloin voidaan ajatella, että Y:ssä olevaa informaatiota ei ole juurikaan hukata, jos se korvataan satunnaisvektorilla Y'.

Nyt suoraviivaisilla laskuilla saadaan

$$\begin{aligned}
\mathbb{E}\big(\, \|Y'' - \mu''\|^2 \,\big) &= \mathbb{E}\big(\, \|Y - Y' - (\mu - \mu')\|^2 \,\big) \\
&= \mathbb{E}\left(\left\| Y - \sum_{i=1}^{d} (v_i^T Y) v_i - \mu + \sum_{i=1}^{d} (v_i^T \mu) v_i \right\|^2 \right) \\
&= \cdots = \mathbb{E}\left(\|Y - \mu\|^2\right) - \sum_{i=1}^{d} \mathbb{E}\left[v_i^T (Y - \mu)\right]^2 .
\end{aligned}$$

Siten meidän tulee itse asiassa maksimoida

$$\sum_{i=1}^{d} \mathbb{E}\left[v_i^T (Y - \mu)\right]^2$$

ortonormaalin joukon $\{v_1, \ldots, v_d\}$ suhteen. Ratkaisun tähän optimointitehtävään antaa seuraava lause.

Lause 9.6. *Olkoot $u_1, \ldots, u_D$ matriisin Σ ortonormaalit ominaisvektorit, $\Sigma u_j = \lambda_j u_j$, $\lambda_1 \geq \cdots \geq \lambda_D$. Silloin*

$$\sum_{i=1}^{d} \mathbb{E}\left[v_i^T(Y-\mu)\right]^2 \leq \sum_{i=1}^{d} \mathbb{E}\left[u_i^T(Y-\mu)\right]^2$$

kaikilla ortonormaaleilla $v_1, \ldots, v_d$.

Todistus. Olkoon $a \in \mathbb{R}^D$. Silloin

$$\begin{aligned}\mathbb{E}\left[a^T(Y-\mu)\right]^2 &= \mathbb{E}\left[a^T(Y-\mu)(Y-\mu)^T a\right] \\ &= a^T \mathbb{E}\left[(Y-\mu)(Y-\mu)^T\right] a = a^T \Sigma\, a.\end{aligned}$$

Nyt $v_i = \sum_{j=1}^{D} (u_j^T v_i) u_j$, joten

$$\begin{aligned}\sum_{i=1}^{d} \mathbb{E}\left[v_i^T(Y-\mu)\right]^2 &= \sum_{i=1}^{d} v_i^T \Sigma v_i \\ &= \sum_{i=1}^{d} \Big[\sum_{j=1}^{D} (u_j^T v_i) u_j\Big]^T \Sigma \Big[\sum_{k=1}^{D} (u_k^T v_i) u_k\Big] \\ &= \sum_{i=1}^{d} \Big[\sum_{j=1}^{D} (u_j^T v_i) u_j\Big]^T \sum_{k=1}^{D} (u_k^T v_i) \lambda_k u_k \\ &= \sum_{i=1}^{d} \sum_{j,k=1}^{D} (u_j^T v_i)(u_k^T v_i) \lambda_k u_j^T u_k \\ &= \sum_{j=1}^{D} \lambda_j \sum_{i=1}^{d} (u_j^T v_i)^2,\end{aligned}$$

missä toiseksi viimeinen yhtälö seurasi vektorien u_j ortonormaalisuudesta. Olkoon yllä

$$\beta_j = \sum_{i=1}^{d} (u_j^T v_i)^2.$$

Silloin $\beta_j \leq \|u_j\|^2 = 1$ ja

$$\sum_{j=1}^{D} \beta_j = \sum_{i=1}^{d} \sum_{j=1}^{D} (u_j^T v_i)^2 = \sum_{i=1}^{d} \|v_i\|^2 = d, \tag{9.5}$$

koska $\|v_i\|^2 = 1$ kaikilla i. Nyt saadaan siis, että

$$\begin{aligned}
\sum_{i=1}^{d} \mathbb{E}\left[v_i^T(Y-\mu)\right]^2 = \sum_{j=1}^{D} \beta_j \lambda_j &= \sum_{j=1}^{d-1} \beta_j \lambda_j + \sum_{j=d}^{D} \beta_j \lambda_j \\
&\leq \sum_{j=1}^{d-1} \beta_j \lambda_j + \lambda_d \Big(d - \sum_{j=1}^{d-1} \beta_j \Big) \\
&= \sum_{j=1}^{d-1} \beta_j (\lambda_j - \lambda_d) + d\lambda_d \\
&\leq \sum_{j=1}^{d-1} (\lambda_j - \lambda_d) + d\lambda_d \\
&= \sum_{j=1}^{d-1} \lambda_j + (d-1)(-\lambda_d) + d\lambda_d \\
&= \sum_{j=1}^{d} \lambda_j,
\end{aligned}$$

missä ensimmäinen epäyhtälö seuraa siitä, että $\lambda_j \leq \lambda_d$, kun $j \geq d$ sekä kaavasta (9.5) ja toinen epäyhtälö siitä, että $\beta_j \leq 1$ ja $\lambda_j - \lambda_d \geq 0$ kaikilla $j < d$. Mutta toisaalta pätee, että

$$\sum_{i=1}^{d} \mathbb{E}\left[u_i^T(Y-\mu)\right]^2 = \sum_{i=1}^{d} u_i^T \Sigma u_i = \sum_{i=1}^{d} \lambda_i,$$

koska $\Sigma u_i = \lambda_i u_i$. Tästä seuraa väite. □

Satunnaisvektorin Y *Karhunen-Loève muunnos* (KL-muunnos) määritellään nyt kaavalla

$$\Psi(Y) = \left[u_1^T Y, \ldots, u_D^T Y\right]^T,$$

missä satunnaismuuttujat $u_i^T Y$, $i = 1, \ldots, D$ ovat Y:n *pääkomponentit*. Karhunen-Loève muunnosta voidaan ajatella kiertona, jossa avaruuden $\mathbb{R}^D$ standardikannan määräämä koordinaatisto kierretään kannan $\{u_1, \ldots, u_D\}$ määräämäksi koordinaatistoksi.

KL-muunnosta kutsutaan myös pääkomponenttimuunnokseksi. Sillä on mm. seuraavat ominaisuudet:

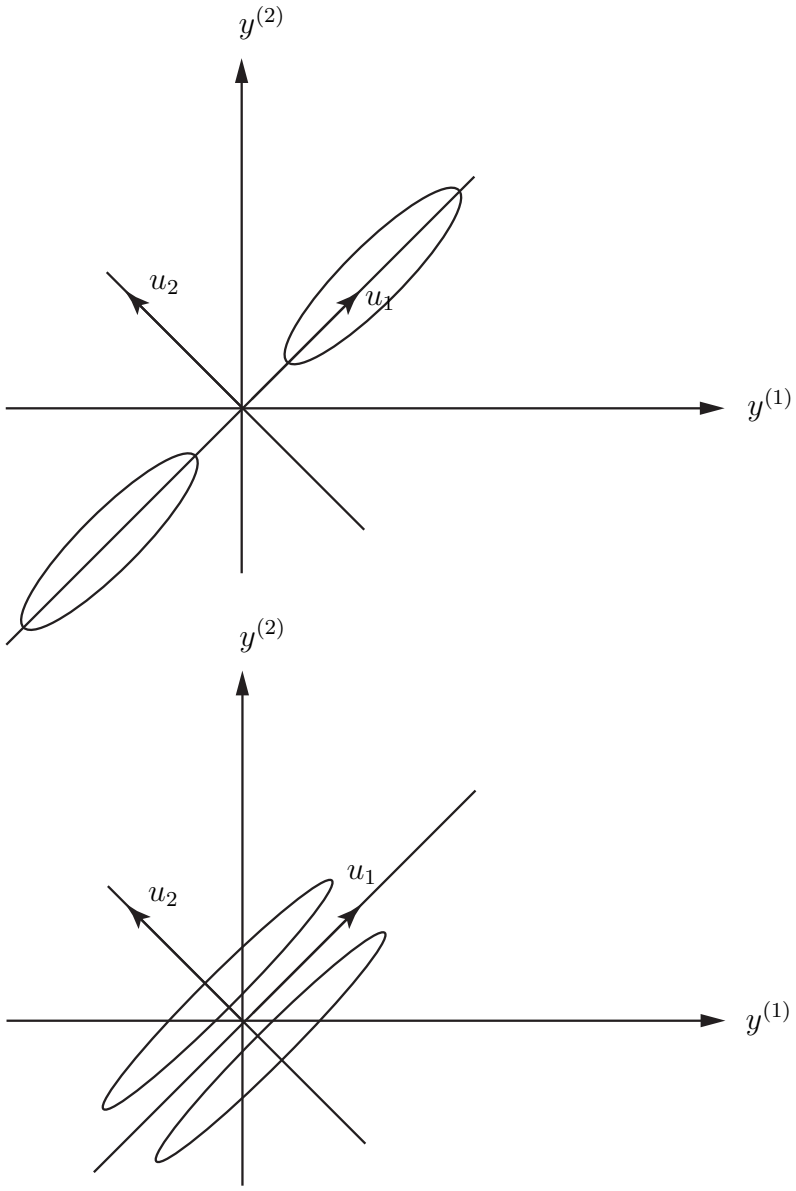
1. Satunnaismuuttujat $u_1^T Y, \ldots, u_D^T Y$ ovat korreloimattomia (HT).
2. $\{u_1, \ldots, u_D\}$ ratkaisee seuraavan optimointitehtävän:
   (*i*) Hae yksikkövektori $u = u_1$, joka maksimoi satunnaismuuttujan $u^T Y$ varianssin.
   (*ii*) Hae yksikkövektori $u = u_{i+1}$, joka maksimoi satunnaismuuttujan $u^T Y$ varianssin ehdolla, että $u^T Y$ ei korreloi satunnaismuuttujien $u_1^T Y, \ldots, u_i^T Y$ kanssa. (HT)

KL-muunnosta käytetään hahmovektorin dimension pienentämiseen määrittelemällä piirteenirrottaja $\varphi : \mathbb{R}^D \to \mathbb{R}^d$, $d < D$ kaavalla

$$\varphi(y) = \left[u_1^T y, \ldots, u_d^T y\right]^T,$$

eli ottamalla piirrevektoriksi vain osa hahmovektorin Y pääkomponenteista.

On kuitenkin syytä huomata, että vaikka KL-muunnos onkin varsin suosittu piirteenirrotin hahmontunnistuksessa, se ei mitenkään ota huomioon hahmovektoriin liittyvää luokkainformaatiota. Tämä saattaa joskus tehdä KL-muunoksen käytön ongelmalliseksi (kuva 9.4). KL-muunnoksesta onkin siksi kehitetty myös versioita, jotka yrittävät jollain tavoin huomioida luokkainformaation, silloin kun sitä on saatavilla.



Kuva 9.4: Yläkuvassa on tilanne, jossa luokittelu voidaan tehokkaasti perustaa ensimmäiseen pääkomponenttiin. Alemmassa kuvassa ensimmäinen pääkomponentti toimisi huonosti kun taas toinen toimisi hyvin. Tässä tilanteessa myös esimerkiksi Fisherin luokitin olisi hyvä valinta.