Luku 5. Estimointiteorian perusteita

T-61.2010 Datasta tietoon, syksy 2011

professori Erkki Oja

Tietojenkäsittelytieteen laitos, Aalto-yliopisto

10.11.2011

Tämän luvun sisältö

Luku käydään läpi kahdella luennolla.

1. Estimointiteorian perusteita
   - Perusjakaumat 1-ulotteisina
   - Yleistys vektoridatalle, d:n muuttujan normaalijakauma
   - Suurimman uskottavuuden periaate
   - Bayes-estimointi
   - Regressiosovitus
   - Esimerkki regressiosta: neuroverkko

Perusjakaumat 1-ulotteisina

- Kun datasta halutaan muodostaa malleja, ne ovat yleensä tilastollisia (esim. regressio, luokittelu, ryhmittely, ...)

  ▸ esimerkki regressiosta ▸ esimerkki luokittelusta ▸ esimerkki ryhmittelystä

- Siksi tarvitaan todennäköisyyslaskentaa
- Peruskäsite siellä on todennäköisyysjakauma (*kertymäfunktio*) skalaariarvoiselle satunnaismuuttujalle x:

$$F(x_0) = P(x \leq x_0) \tag{1}$$

  antaa todennnäköisyyden sille että $x \leq x_0$.
- Funktio $F(x_0)$ on ei-vähenevä, jatkuva ja $0 \leq F(x_0) \leq 1$.

Perusjakaumat 1-ulotteisina (2)

- *Tiheysfunktio* x:lle pisteessä $x = x_0$

$$p(x_0) = \left. \frac{dF(x)}{dx} \right|_{x=x_0} \tag{2}$$

  on funktion $F(x)$ derivaatta pisteessä $x = x_0$

- Vastaavasti tietenkin

$$F(x_0) = \int_{-\infty}^{x_0} p(x)dx$$

Perusjakaumat 1-ulotteisina (3)

- Tiheysfunktiolla on seuraavat ominaisuudet:

$$\begin{aligned} \int_{-\infty}^{\infty} p(x)dx &= 1 \\ \int_{-\infty}^{\infty} xp(x)dx &= \mu = \mathrm{E}\{x\} \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx &= \sigma^2 = \mathrm{Var}\{x\}. \end{aligned}$$

- Tässä μ ja σ ovat x:n odotusarvo ja keskihajonta.

Perusjakaumat 1-ulotteisina

Esimerkkejä

- Oletan tunnetuiksi jotkut yleisimmät tiheysfunktiot, kuten *normaalijakauma* (eli Gaussin jakauma)

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

eksponentiaalinen jakauma

$$p(x) = \lambda e^{-\lambda x}$$

ja *tasainen jakauma* (välillä [a,b])

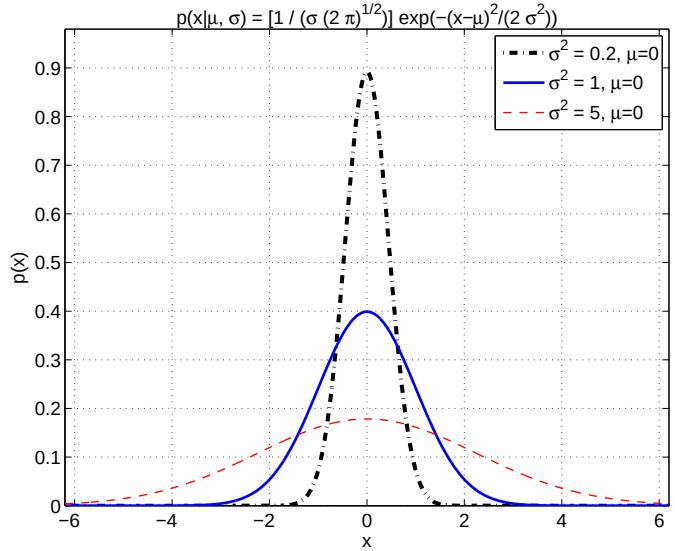
$$p(x) = 1/(b-a), \text{ kun } x \in [a,b], \text{ 0 muuten}$$

Perusjakaumat 1-ulotteisina

Normaalijakauman tiheysfunktio p(x)

Parametrit μ (keskiarvo) ja σ (keskihajonta, σ^2 varianssi)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



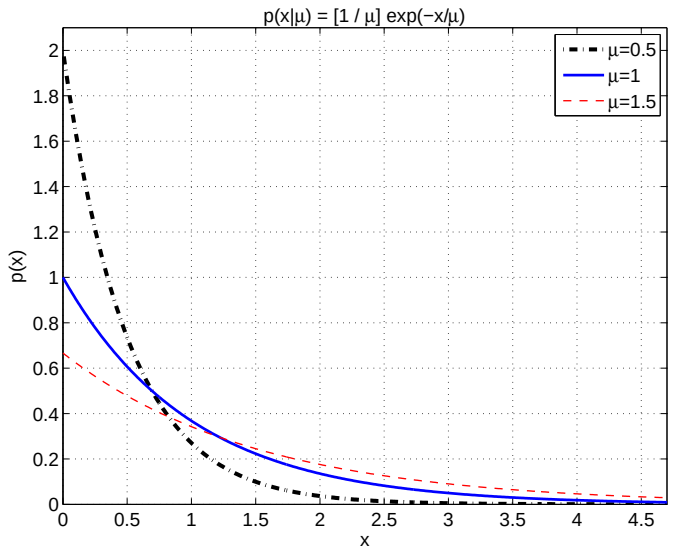
Kuva: Normaalijakauma kolmella eri σ:n arvolla

Perusjakaumat 1-ulotteisina

Eksponentiaalijakauman tiheysfunktio p(x)

Parametri λ ($1/\mu$, jossa μ keskiarvo)

$$p(x) = \lambda e^{-\lambda x}$$



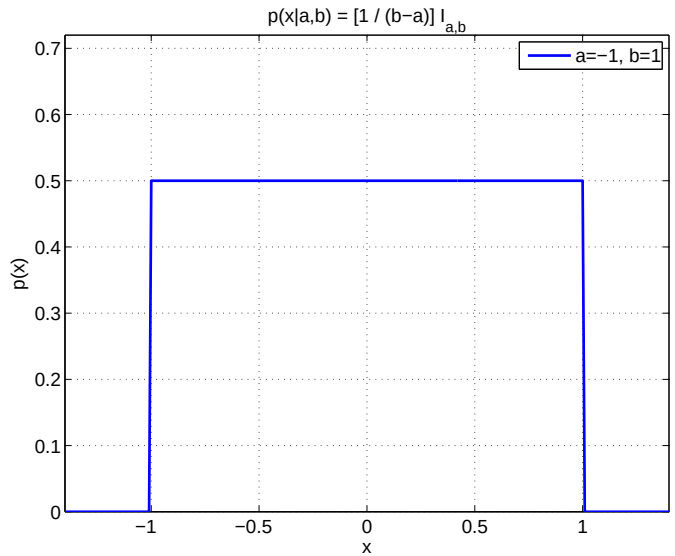
Kuva: Eksponenttijakauma kolmella eri λ:n arvolla

Perusjakaumat 1-ulotteisina

Tasainen jakauman tiheysfunktio p(x)

Tasainen välillä $[a, b]$

$$p(x) = \frac{1}{b-a}, \quad \text{kun } x \in [a, b],\ 0 \text{ muuten}$$



Kuva: Tasajakauma

Yleistys vektoridatalle

- Kun puhutaan vektoridatasta, on pakko yleistää yo. jakaumat moniulotteisiksi: olkoon taas datavektori $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$
- Sen todennäköisyysjakauma (kertymäfunktio) on

$$F(\mathbf{x}_0) = P(\mathbf{x} \leq \mathbf{x}_0) \tag{3}$$

missä relaatio "$\leq$" ymmärretään alkioittain

- Vastaava moniulotteinen tiheysfunktio $p(\mathbf{x})$ on tämän osittaisderivaatta kaikkien vektorialkioiden suhteen:

$$p(\mathbf{x}_0) = \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}\cdots\frac{\partial}{\partial x_d}F(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}_0} \tag{4}$$

- Käytännössä tiheysfunktio on tärkeämpi.

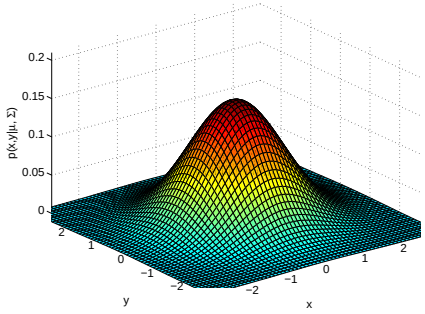
Yleistys vektoridatalle

Kahden muuttujan normaalijakauma, symmetrinen, $d = 2$

- "Symmetrinen" 2-ulotteinen ($d = 2$) normaalijakauma on

$$p(\mathbf{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{2\sigma^2}}$$

missä jakauman odotusarvo (keskipiste) on piste $\mathbf{m} = [\mu_1,\ \mu_2]$ ja keskihajonta joka suuntaan on sama σ.



Kuva: Symmetrinen 2D-normaalijakauma

Yleistys vektoridatalle

Kahden muuttujan normaalijakauma, $d = 2$

- Yllä oleva *2-ulotteinen normaalijakauma* voidaan yleisemmin kirjoittaa muotoon

$$p(\mathbf{x}) = K \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right)$$

jossa skaalaustermi K ($d = 2$)

$$K = \frac{1}{(2\pi)^{d/2}\det(\mathbf{C})^{1/2}} = \frac{1}{2\pi\det(\mathbf{C})^{1/2}}$$

ja $\mathbf{C}$ (toisinaan Σ) on (2×2)-kokoinen kovarianssimatriisi ja $\mathbf{m} = [\mu_1 \;\; \mu_2]^T$ keskiarvovektori (huippukohta)

Yleistys vektoridatalle (2)

Kahden muuttujan normaalijakauma, $d = 2$

- Keskihajonta / varianssi on 2-ulotteisessa tapauksessa monimutkaisempi kuin 1-ulotteisessa tapauksessa: varianssin $\sigma^2 = \mathrm{E}\{(x - \mu)^2\}$ korvaa kovarianssimatriisi

$$\mathbf{C} = \begin{bmatrix} \mathrm{E}\{(x_1 - \mu_1)^2\} & \mathrm{E}\{(x_1 - \mu_1)(x_2 - \mu_2)\} \\ \mathrm{E}\{(x_1 - \mu_1)(x_2 - \mu_2)\} & \mathrm{E}\{(x_2 - \mu_2)^2\} \end{bmatrix}$$

- Esimerkkejä 2-ulotteisen normaalijakauman tiheysfunktioista
  - $\mu_1 = 0$, $\mu_2 = 0$, symmetrinen diagonaalinen $\mathbf{C}$, jossa lisäksi $\sigma_1 = \sigma_2$
  - $\mu_1 = 0$, $\mu_2 = 0$, symmetrinen diagonaalinen $\mathbf{C}$, jossa $\sigma_1 \neq \sigma_2$
  - $\mu_1 = 0$, $\mu_2 = 0$, symmetrinen ei-diagonaalinen $\mathbf{C}$

Yleistys vektoridatalle

d:n muuttujan normaalijakauma

- Yleistys d:hen dimensioon suoraviivainen: matriisialkiot ovat $\mathbf{C}_{ij} = \mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = \mathrm{Cov}(x_i, x_j)$
- Kovarianssimatriisi $\mathbf{C}$ on siis symmetrinen neliömatriisi kokoa $(d \times d)$
- Silloin *d-ulotteisen normaalijakauman* tiheysfunktio voidaan kirjoittaa vektori-matriisimuotoon:

$$p(\mathbf{x}) = K \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{m})\right) \qquad (5)$$

- missä $\mathbf{m} = [\mu_1, \ldots, \mu_d]^T$ on keskiarvovektori (jakauman keskipiste, huippukohta) ja K on normalisoiva termi

$$K = \frac{1}{(2\pi)^{d/2} \det(\mathbf{C})^{1/2}}$$

Yleistys vektoridatalle (2)

d:n muuttujan normaalijakauma

- Termi K tarvitaan vain, jotta $p(\mathbf{x})$:n integraali d-ulotteisen avaruuden yli olisi 1
- Voidaan integroimalla johtaa että

$$\begin{aligned} \mathrm{E}\{\mathbf{x}\} &= \mathbf{m} = \int_{\mathcal{R}^{\lceil}} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ \mathrm{E}\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\} &= \mathbf{C} \end{aligned}$$

- Huomaathan, että vaikka $d > 1$, niin $p(\mathbf{x}) \in \mathbb{R}^1$ eli tiheysfunktion arvo on skalaari, sillä matriisikertolaskut eksponenttifunktiossa $(1 \times \underline{d})(\underline{d} \times \underline{\underline{d}})(\underline{\underline{d}} \times 1) = (1 \times 1)$

Yleistys vektoridatalle

Korreloimattomuus ja riippumattomuus

- Vektorin **x** alkiot ovat *korreloimattomat* jos niiden kovarianssit ovat nollia eli $\mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = 0$. Toisin sanoen, kovarianssimatriisi **C** on diagonaalinen

$$\mathbf{C} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & & \dots & \sigma_d^2 \end{bmatrix}$$

- Vektorin **x** alkiot ovat *riippumattomat* jos niiden *yhteistiheysjakauma* voidaan lausua *marginaalijakaumien* tulona:

$$p(\mathbf{x}) = p_1(x_1)p_2(x_2)....p_d(x_d)$$

Yleistys vektoridatalle (2)

Korreloimattomuus ja riippumattomuus

- Riippumattomuus tuottaa aina korreloimattomuuden, mutta ei välttämättä toisinpäin
- Osoitetaan että jos *normaalijakautuneen vektorin alkiot* ovat korreloimattomat, niin ne ovat myös riippumattomat: katsotaan termiä $(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$ tapauksessa missä $\mathrm{E}\{(x_i-\mu_i)(x_j-\mu_j)\}=0$ (korreloimattomuus).
- Mutta silloin $\mathbf{C}$ on lävistäjämatriisi ja

$$(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})=\sum_{i=1}^{d}(\mathbf{C}_{ii})^{-1}(x_i-\mu_i)^2$$

Yleistys vektoridatalle (3)

Korreloimattomuus ja riippumattomuus

- Kun otetaan tästä eksponenttifunktio kuten kaavassa (5) saadaan

$$
\begin{aligned}
p(\mathbf{x}) &= K \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right) \\
&= K \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{d}(\mathbf{C}_{ii})^{-1}(x_i-\mu_i)^2\right) \\
&= K \cdot \exp(-\frac{1}{2}\mathbf{C}_{11}^{-1}(x_1-\mu_1)^2)\ldots\exp(-\frac{1}{2}\mathbf{C}_{dd}^{-1}(x_d-\mu_d)^2)
\end{aligned}
$$

joten $p(\mathbf{x})$ hajaantuu marginaalijakaumien tuloksi, eli alkiot x_i ovat riippumattomia.

Suurimman uskottavuuden periaate

Parametrisen tiheysfunktion estimointi

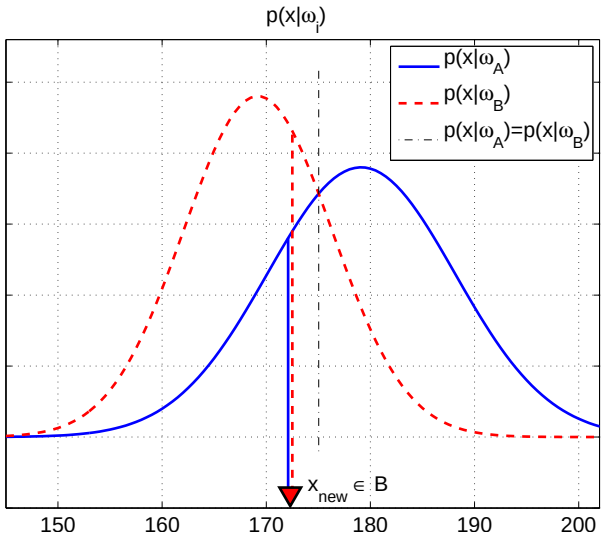
- Edellä esiteltiin teoriassa vektorimuuttujan jakauma, etenkin normaalijakauma.
- Jos on kuitenkin annettuna vain vektorijoukko datamatriisin **X** muodossa, mistä tiedämme sen jakauman?

  ▸ Data **X** ja eräs mahdollinen $p(x)$
- Jakauma olisi erittäin hyödyllinen, koska sen avulla voidaan esim. vastata seuraavaan kysymykseen: kun on annettuna uusi vektori **x**, millä todennäköisyydellä se kuuluu samaan joukkoon kuin datajoukko **X** (tai jokin sen osajoukko).

Suurimman uskottavuuden periaate (2)

Parametrisen tiheysfunktion estimointi

- Etenkin luokittelu perustuu usein jakaumien estimointiin



Kuva: Esimerkki luokittelusta. Olkoon data $\mathbf{X}_A$ naisten ja $\mathbf{X}_B$ miesten pituuksia (cm), ja niistä estimoidut normaalijakaumat $p_A(x) \equiv p(x|\omega_A)$ ja $p_B(x) \equiv p(x|\omega_B)$. Jakaumia voidaan käyttää luokittelemaan uusi datapiste $p(x_{new}|\omega_B) > p(x_{new}|\omega_A)$. Havainto x_{new} luokiteltaisiin naiseksi.

Suurimman uskottavuuden periaate (3)

Parametrisen tiheysfunktion estimointi

- Eräs mahdollisuus on d-ulotteinen histogrammi: mutta dimensionaalisuuden kirous tekee tästä hyvin hankalan jos d vähänkin suuri
- Parempi menetelmä on yleensä *parametrinen tiheysestimointi*, joka tarkoittaa, että oletetaan tiheysfunktiolle $p(\mathbf{x})$ jokin muoto (esim. normaalijakauma) ja pyritään estimoimaan datajoukosta vain sen *parametrit* – normaalijakaumalle siis keskiarvovektori $\mathbf{m} = [\mu_1 \ \ldots \ \mu_d]^T$ ja kovarianssimatriisi $\mathbf{C} = (\mathbf{C}_{ij})$
- Tämä tekee yhteensä vain $d + d(d+1)/2$ parametria, koska $\mathbf{C}$ on symmetrinen matriisi kokoa $(d \times d)$
- Vertaa histogrammiin, jossa lokeroiden määrä kasvaa eksponentiaalisesti l^d.

Suurimman uskottavuuden periaate (4)

Parametrisen tiheysfunktion estimointi

- Merkitään tuntemattomien parametrien muodostamaa järjestettyä joukkoa vektorilla $\boldsymbol{\theta}$ ja tiheysjakaumaa $p(\mathbf{x}|\boldsymbol{\theta})$
- Tarkoittaa: funktion varsinainen argumentti ovat vektorialkiot $x_1, \ldots, x_d$ mutta funktio riippuu myös parametreista ($\boldsymbol{\theta}$:n alkioista) aivan niinkuin vaikkapa normaalijakauman muoto muuttuu kun kovarianssimatriisi muuttuu
- Funktion yleinen muoto oletetaan siis tunnetuksi (parametreja vaille) ja parametrivektorin $\boldsymbol{\theta}$ alkiot ovat vakioita mutta tuntemattomia.
- Miten estimaatit $\hat{\boldsymbol{\theta}}$ sitten ratkaistaan?

Suurimman uskottavuuden periaate

- *Suurimman uskottavuuden* menetelmässä (maximum likelihood) ▸ Milton: ML method valitaan se parametrivektori θ joka maksimoi datavektorijoukon yhteisen tiheysfunktion, ns. *uskottavuusfunktion* $L(\theta)$

$$L(\theta) = p(\mathbf{X} \mid \theta) = p(\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n) \mid \theta) \qquad (6)$$

- Tässä $\mathbf{X}$ on siis koko datamatriisi ja $\mathbf{x}(1), ..., \mathbf{x}(n)$ ovat sen sarakkeet
- *Ajatus: valitaan sellaiset arvot parametreille, että havaittu datajoukko on todennäköisin mahdollinen* ▸ Esimerkki
- Huomaa että kun numeerinen datamatriisi $\mathbf{X}$ sijoitetaan tähän funktioon, se ei olekaan enää $\mathbf{x}$:n funktio vaan ainoastaan θ:n.

Suurimman uskottavuuden periaate (2)

- Siten parametrit saadaan maksimoimalla (“derivaatan nollakohta”)

$$\frac{\partial}{\partial \boldsymbol{\theta}} p(\mathbf{X} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{ML}} = \mathbf{0} \tag{7}$$

- Yleensä aina ajatellaan että datavektorit ($\mathbf{X}$:n sarakkeet) ovat riippumattomia, jolloin uskottavuusfunktio hajoaa tuloksi:

$$L(\boldsymbol{\theta}) = p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} p(\mathbf{x}(j) \mid \boldsymbol{\theta}) \tag{8}$$

- Useimmiten uskottavuusfunktiosta $L(\boldsymbol{\theta})$ otetaan vielä logaritmi $\ln L(\boldsymbol{\theta})$, koska silloin (a) laskenta muuttuu yleensä helpommaksi (useimpien tiheysfunktioiden tulo muuttuu summaksi) ja (b) sekä $L(\boldsymbol{\theta})$:n että $\ln L(\boldsymbol{\theta})$:n ääriarvo (maksimi) on samassa pisteessä.

Suurimman uskottavuuden periaate

Esimerkki 1-ulotteiselle normaalijakaumalle

- Esimerkki: otetaan 1-ulotteinen (skalaari)data, eli datamatriisissa on vain skalaarilukuja $x(1), \ldots, x(n)$. Oletetaan että näytteet ovat riippumattomia toisistaan.
- Oletetaan että ne ovat normaalijakautuneita, mutta emme tiedä niiden odotusarvoa μ emmekä varianssia σ^2. Lasketaan nämä suurimman uskottavuuden menetelmällä.
- Uskottavuusfunktio “ensimmäiselle datapisteelle”:

$$p(x(1) \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}[x(1) - \mu]^2\right]$$

Suurimman uskottavuuden periaate (2)

Esimerkki 1-ulotteiselle normaalijakaumalle

- Uskottavuusfunktio $L(\theta)$ koko datasetille:

$$p(\mathbf{X} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2\right] \quad (9)$$

- Otetaan logaritmi $\ln L(\theta)$:

$$\ln p(\mathbf{X} \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2 \quad (10)$$

Suurimman uskottavuuden periaate (3)

Esimerkki 1-ulotteiselle normaalijakaumalle

- Etsitään ääriarvo asettamalla derivaatta nollaksi ja saadaan yhtälö odotusarvolle μ

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{X} \mid \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{j=1}^{n} [x(j) - \mu] = 0 \qquad (11)$$

- Ratkaistaan tämä μ:n suhteen, saadaan otoksen keskiarvo

$$\mu = \hat{\mu}_{ML} = \frac{1}{n} \sum_{j=1}^{n} x(j) \qquad (12)$$

Suurimman uskottavuuden periaate (4)

Esimerkki 1-ulotteiselle normaalijakaumalle

- Vastaava yhtälö varianssille σ^2 on

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{X} \mid \mu, \sigma^2) = 0 \tag{13}$$

josta tulee ML-estimaattina otoksen varianssi

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{j=1}^{n} [x(j) - \hat{\mu}_{ML}]^2. \tag{14}$$

Bayes-estimointi

- Bayes-estimointi on periaatteeltaan erilainen kuin ML-estimointi
- Oletamme että tuntematon parametrijoukko (vektori) θ ei olekaan kiinteä vaan silläkin on jokin oma jakaumansa $p(\theta)$
- Kun saadaan mittauksia / havaintoja, niiden avulla θ:n jakauma *tarkentuu* (esim. sen varianssi pienenee).
- Käytännössä Bayes-estimointi ei välttämättä ole paljonkaan ML-menetelmää raskaampi mutta antaa parempia tuloksia
- Muistetaan todennäköisyyslaskennasta *ehdollisen todennäköisyyden* kaava: $P(A|B) = P(A,B)/P(B)$ missä A ja B ovat joitakin tapahtumia (ajattele vaikka nopanheittoa)

Bayes-estimointi (2)

- Tätä voi hyödyntää ajattelemalla datajoukkoa ja sitä mallia (todennäköisyysjakauman parametreja) josta datajoukko on tullut:

$$P(malli, data) = P(malli|data)P(data)$$

$$= P(data, malli) = P(data|malli)P(malli)$$

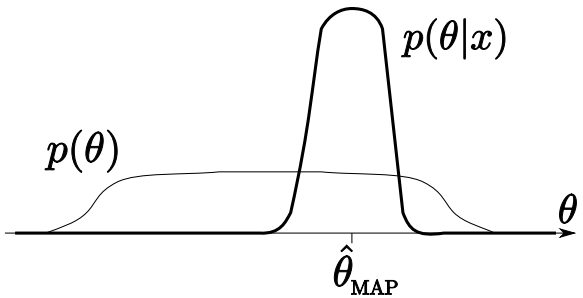
jossa vain on kirjoitettu datan ja mallin yhteisjakauma kahdella eri tavalla

- Mutta tästä seuraa ns. *Bayesin kaava*

$$P(malli|data) = \frac{P(data|malli)P(malli)}{P(data)} \qquad (15)$$

Bayes-estimointi (3)

- Ajatus on että meillä on jokin käsitys mallin todennäköisyydestä ennen kuin mitään dataa on olemassa: $P(\textit{malli})$, ns. *prioritodennäköisyys*
- Kun dataa sitten saadaan, tämä tarkentuu todennäköisyydeksi $P(\textit{malli}|\textit{data})$, ns. *posterioritodennäköisyys*



Kuva: Bayesiläinen laskenta: Priori $p(\theta)$ ja posteriori $p(\theta|\mathbf{X})$, kun havaittu dataa $\mathbf{X}$. Posteriorista laskettu piste-estimaatti θ_{MAP}.

Bayes-estimointi (4)

- Bayesin kaava kertoo kuinka tämä tarkentuminen lasketaan.
- Jos meillä on taas datamatriisi $\mathbf{X}$ ja tuntematon parametrivektori θ, niin Bayes antaa

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

- (HUOM matemaattinen merkintätapa: kaikkia todennäköisyysjakaumia (tiheysfunktioita) merkitään vain p:llä, ja argumentti kertoo mistä p:stä on kysymys. Tämä on tietysti matemaattisesti väärin mutta yleisesti käytetty tapa.)

Bayes-estimointi (5)

- Yo. kaavassa nähdään että priorijakauma kerrotaan *uskottavuusfunktiolla* ja jaetaan *datan jakaumalla*, jotta saataisiin posteriorijakauma.
- Bayes-analyysissä “keksitään” priorijakauma ja pyritään muodostamaan posteriorijakauma
- Usein kuitenkin tyydytään etsimään posteriorin maksimikohta θ:n suhteen: Maksimi-posteriori eli MAP-estimointi
- Tällä on selkeä yhteys ML-estimointiin: nimittäin Bayesin kaava antaa

$$\ln p(\theta|\mathbf{X}) = \ln p(\mathbf{X}|\theta) + \ln p(\theta) - \ln p(\mathbf{X})$$

Bayes-estimointi (6)

ja etsittäessä maksimikohtaa $\boldsymbol{\theta}$:n suhteen tulee gradienttiyhtälö

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{X}|\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{\theta}) = \mathbf{0} \tag{16}$$

- Huomataan että ML-menetelmään verrattuna tulee ylimääräinen termi $\partial(\ln p(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}$.
- Siitä on joskus suurta hyötyä, kuten seuraavassa kohdassa nähdään.

Regressiosovitus

- Regressio eroaa tähänastisista ohjaamattoman oppimisen menetelmistä (PCA, DSS, todennäköisyysjakaumien estimointi) siinä, että kuhunkin datavektoriin $\mathbf{x}(i)$ liittyy jokin lähtösuure $y(i)$, jonka arvellaan noudattavan mallia

$$y(i) = f(\mathbf{x}(i), \boldsymbol{\theta}) + \epsilon(i)$$

- Tässä f on jokin funktio, ns. regressiofunktio, ja $\epsilon(i)$ on virhe (y-akselin suunnassa)
- Funktiolla f on tunnettu (tai oletettu) muoto mutta tuntematon parametrivektori $\boldsymbol{\theta}$

▸ Yleinen esimerkki polynomisesta sovituksesta

Regressiosovitus

ML-estimointi lineaarisessa regressiossa

- Esimerkkinä lineaarinen regressio:

$$y(i) = \theta_0 + \sum_{j=1}^{d} \theta_j x_j(i) + \epsilon(i) = \theta_0 + \boldsymbol{\theta}^T \mathbf{x}(i) + \epsilon(i)$$

- Tunnettu tapa ratkaista parametrit $\hat{\theta}$ on *pienimmän neliösumman* keino: minimoi θ:n suhteen

$$\frac{1}{n} \sum_{i=1}^{n} [y(i) - f(\mathbf{x}(i), \boldsymbol{\theta})]^2$$

Regressiosovitus (2)
ML-estimointi lineaarisessa regressiossa

- Itse asiassa tämä on *ML-estimaatti tietyllä ehdolla*: jos regressiovirhe $\epsilon(i)$ (kaikille arvoille i) on riippumaton $\mathbf{x}(i)$:n arvosta ja normaalijakautunut odotusarvolla 0 ja hajonnalla σ. (Hyvin luonnollinen oletus!)
- Silloin ML-estimaatti parametrille θ saadaan uskottavuusfunktiosta ($Y = (y(i))_i$)

$$p(\mathbf{X}, Y|\theta) = p(\mathbf{X})p(Y|\mathbf{X}, \theta) = p(\mathbf{X}) \prod_{i=1}^{n} p(y(i)|\mathbf{X}, \theta)$$

jossa on vain käytetty ehdollisen todennäköisyyden kaavaa, huomattu että $\mathbf{X}$ ei riipu regressiomallin parametreistä ja oletettu eri mittapisteissä olevat $y(i)$:t riippumattomiksi (kuten ML-estimoinnissa yleensä oletetaan)

Regressiosovitus (3)
ML-estimointi lineaarisessa regressiossa

- Otetaan logaritmi

$$\ln p(\mathbf{X}, Y|\theta) = \ln p(\mathbf{X}) + \sum_{i=1}^{n} \ln p(y(i)|\mathbf{X}, \theta)$$

- Mutta jos $\mathbf{X}$ eli sen sarakkeet $\mathbf{x}(i)$ ovat kiinteitä niinkuin ehdollisessa todennäköisyydessä ajatellaan, on $y(i)$:n jakauma sama kuin $\epsilon(i)$:n mutta keskiarvo on siirtynyt kohtaan $f(\mathbf{x}(i), \theta)$ - siis normaalijakauma:

$$p(y(i)|\mathbf{X}, \theta) = \mathit{vakio} \times \exp(-\frac{1}{2\sigma^2}[y(i) - f(\mathbf{x}(i), \theta)]^2)$$

Regressiosovitus (4)
ML-estimointi lineaarisessa regressiossa

- Vihdoin saadaan uskottavuusfunktion logaritmiksi:

$$\ln p(\mathbf{X}, Y|\theta) = \ln p(\mathbf{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y(i) - f(\mathbf{x}(i), \theta)]^2 + n\ln(\mathit{vakio})$$

- Maksimoinnissa (derivointi θ:n suhteen) θ:sta riippumattomat termit katoavat
- Täten maksimointi on täsmälleen sama kuin pienimmän neliösumman minimointi! (Koska $p(\mathbf{X})$ on datamatriisin jakauma eikä riipu mistään mallista)

Regressiosovitus

Bayesin priorijakauman hyväksikäyttö lineaarisessa regressiossa

- Mitä Bayes voi antaa tähän lisää?
- ML:ssä maksimoitiin $p(\mathbf{X}, Y|\theta)$. Bayes-estimoinnissa maksimoidaan $p(\theta|\mathbf{X}, Y) \propto p(\mathbf{X}, Y|\theta) \cdot p(\theta)$, joka logaritmin jälkeen on $\ln p(\mathbf{X}, Y|\theta) + \ln p(\theta)$
- Eli maksimoitavassa funktiossa on priorijakaumasta tuleva termi $\ln p(\theta)$ ja maksimoitava funktio on

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y(i) - f(\mathbf{x}(i), \theta)]^2 + \ln p(\theta)$$

- Tästä näkee, että jos kohinan $\epsilon(i)$ varianssi σ^2 on hyvin suuri, niin 1. termi on pieni ja θ:n priorijakauma määrää pitkälle sen arvon.

Regressiosovitus (2)

Bayesin priorijakauman hyväksikäyttö lineaarisessa regressiossa

- Päinvastoin jos σ^2 on pieni, 1. termi dominoi ja priorijakaumalla on hyvin vähän vaikutusta lopputulokseen
- Tämä tuntuu hyvin luonnolliselta ja hyödylliseltä!
- Esimerkiksi jos on syytä olettaa että kaikki parametrit ovat (0,1)-normaalijakautuneita, on

$$p(\boldsymbol{\theta}) = \mathit{vakio} \times \exp(-1/2 \sum_{k=1}^{K} \theta_k^2),$$

$$\ln p(\boldsymbol{\theta}) = \ln(\mathit{vakio}) - 1/2 \sum_{k=1}^{K} \theta_k^2,$$

ja prioritermin maksimointi johtaa pieniin arvoihin parametreille.

Regressiosovitus (3)

Bayesin priorijakauman hyväksikäyttö lineaarisessa regressiossa

- Tällöin regressiomalli yleensä käyttäytyy paremmin kuin jos parametreilla on hyvin suuria arvoja.

Esimerkki lineaarisesta regressiosta

- Katsotaan esimerkkinä lineaarista regressiota
- Nyt oletetaan että datasta tiedetään seuraavaa: yhteys on joko 1. asteen (lineaarinen) tai ehkä toisen asteen polynomi, joka kulkee hyvin luultavasti origon kautta. Mallin virhe (mitattujen $y(i)$- arvojen poikkeama mallin antamista arvoista) on normaalijakautunut hajonnalla σ.
- Malli on silloin

$$y(i) = a + bx(i) + cx(i)^2 + \epsilon(i)$$

- Käytetään Bayes-estimointia. Priorit valitaan normaalijakautuneiksi siten että a:n keskiarvo on 0 hajonta on 0.1, b:n keskiarvo on 1 ja hajonta 0.5 ja c:n keskiarvo on 0 ja hajonta 1.

Esimerkki lineaarisesta regressiosta (2)

- Näiden tiheydet ovat siis
  $vakio \times e^{-50a^2}$, $vakio \times e^{-2(b-1)^2}$, $vakio \times e^{-0.5c^2}$ ja maksimoitava funktio on (kun vakiot tiputetaan pois)

$$-\frac{1}{2\sigma^2}\sum[y(i)-a-bx(i)-cx(i)^2]^2-50a^2-2(b-1)^2-0.5c^2$$

- Derivoimalla a:n suhteen ja ratkaisemalla saadaan

$$a=\frac{1}{n+100\sigma^2}\sum[y(i)-bx(i)-cx(i)^2]$$

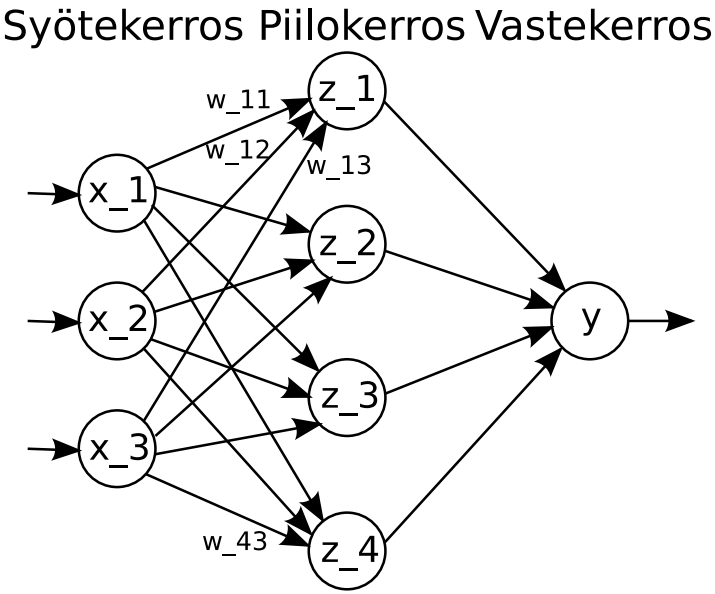
ja vastaavat yhtälöt b:lle ja c:lle, joista ne kaikki voidaan ratkaista (lineaarinen yhtälöryhmä kolmelle muuuttujalle).

Esimerkki lineaarisesta regressiosta (3)

- Ilman Bayesia yo. kaavasta putoaa kokonaan pois termi $100\sigma^2$, joten Bayes pakottaa *a*:n aina pienempään arvoon. Ainoastaan kun mallin virhevarianssi σ^2 on hyvin pieni, Bayesin vaikutus menee vähäiseksi (silloin data on hyvin luotettavaa ja etukäteistiedon merkitys vähenee).

Esimerkki regressiosta: neuroverkko

- *Neuroverkko* on oheisen kuvan esittämä laskentamalli, joka laskee kohtalaisen monimutkaisen funktion tulosuureistaan (inputeistaan)



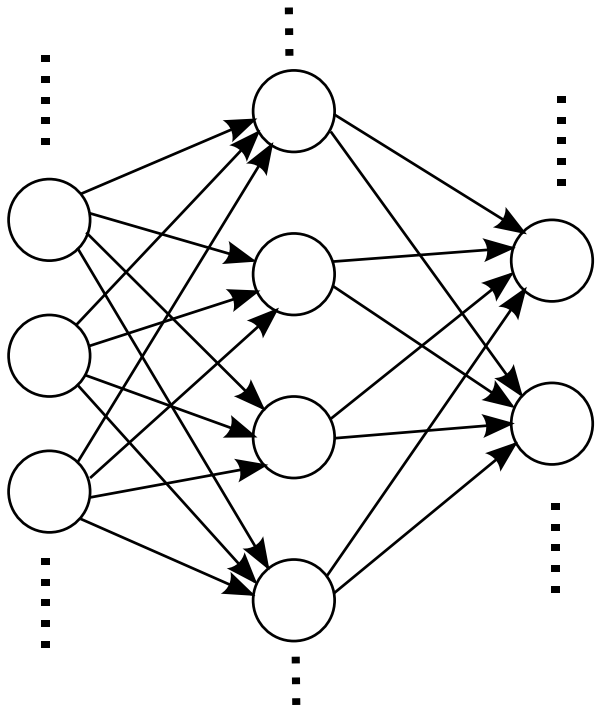
Kuva: Monikerrosperseptroniverkko (MLP) yhdellä ulostulolla.

Esimerkki regressiosta: neuroverkko (2)

- Malli koostuu “neuroneista” jotka laskevat funktioita tulosuureistaan
- Ensimmäisen kerroksen, ns. piilokerroksen neuronit laskevat kukin funktion $z_i = f(\mathbf{x}, \mathbf{w}_i)$ tulosuureista jotka ovat syötteenä olevan vektorin $\mathbf{x}$ alkiot
- Funktio on epälineaarinen ja $\mathbf{w}_i$ on i:nnen neuronin parametrivektori (jossa alkioita yhtä monta kuin $\mathbf{x}$:ssä)
- Toisen kerroksen neuroni laskee vain painotetun summan piilokerroksen lähtösuureista z_i: $y = \sum_i v_i z_i$

Esimerkki regressiosta: neuroverkko (3)

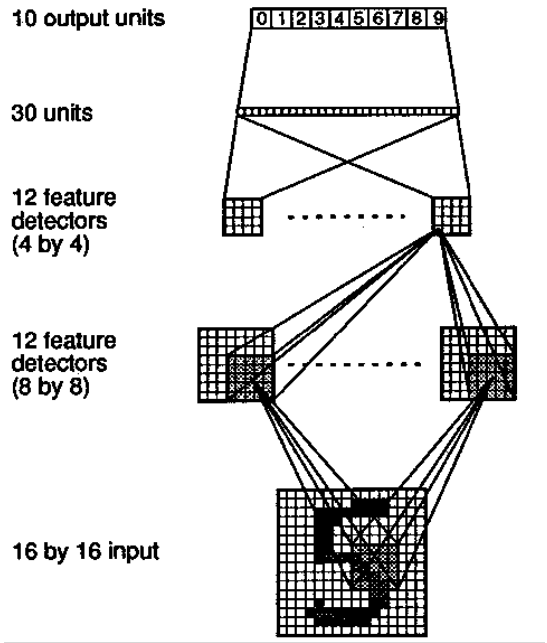
- Toisessa kerroksessa voi olla myös useampia rinnakkaisia neuroneita, jolloin verkko laskee regressiofunktion vektoreista **x** vektoreihin **y**;



Kuva: Yleinen monikerrosperseptroniverkko (MLP).

Esimerkki regressiosta: neuroverkko (4)

- MLP-verkon sovelluksena vaikkapa numerontunnistus



Kuva: Esimerkki neuroverkon käytöstä käsinkirjoitettujen merkkien tunnistamisessa. Neuroverkko oppii aineistosta kunkin numeron ja antaa suurimman vasteen kyseiselle ulostuloneuronille.

Esimerkki regressiosta: neuroverkko (5)

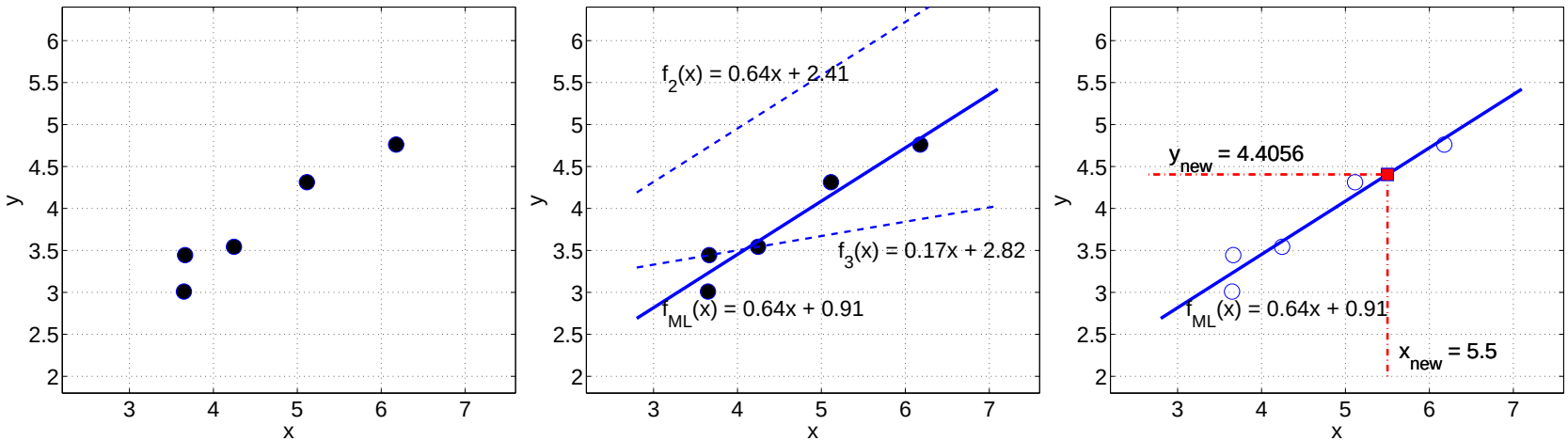
- Neuroverkoista on tullut tavattoman suosittuja hyvin monilla aloilla, koska ne laskevat hyvin isoja regressiomalleja, ovat osoittautuneet suhteellisen tarkoiksi ja tehokkaiksi, ja parametrien laskemiseksi isoissakin malleissa (satoja tai tuhansia neuroneita) on hyvin tehokkaita koneoppimisalgoritmeja ja helppokäyttöisiä ohjelmistopaketteja
- Raportoituja sovelluksia on satoja ellei tuhansia
- Enemmän kurssissa T-61.5130 Machine Learning and Neural Networks.

Yhteenveto

Tässä luvussa

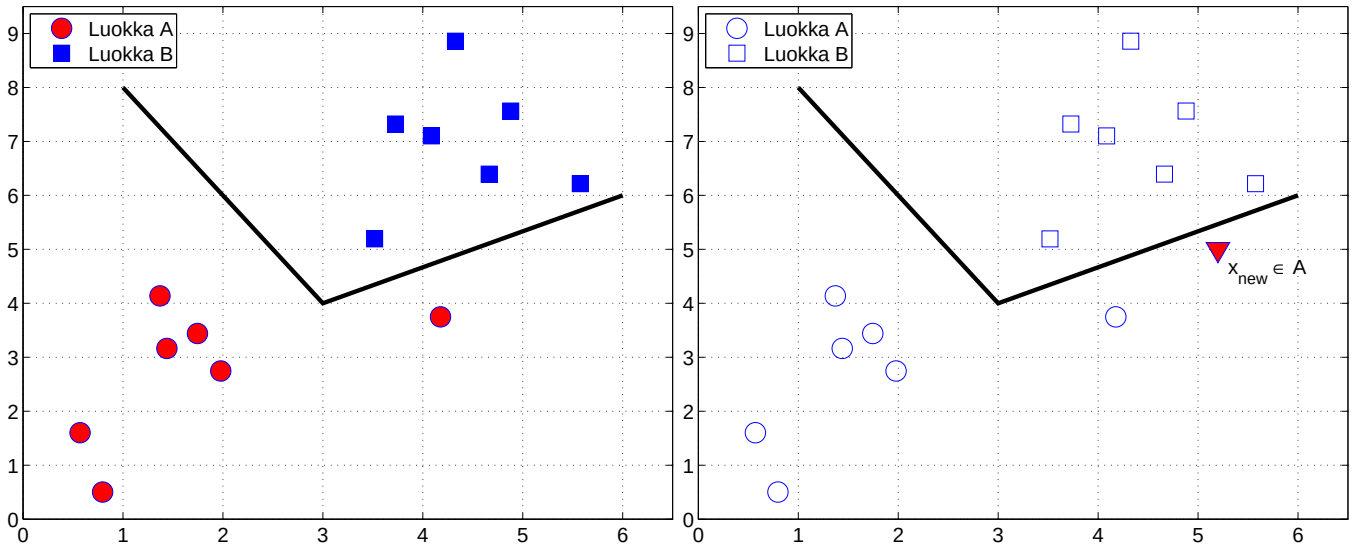
- tutustuttiin tiheysfunktioihin
- esiteltiin suurimman uskottavuuden menetelmä parametrien estimointiin
- liitettiin etukäteistieto (a priori) parametrien estimointiin Bayesin teoreemaa käyttäen
- laskettiin regressiota eri menetelmillä (ML, Bayes, neuroverkko)

Esimerkki yksinkertaisesta lineaarisesta regressiosta



Kuva: Vasen kuva: datapisteitä (x_i, y_i). Tavoite selittää y x:n avulla käyttämällä funktiota $f(x, \theta) = kx + b$. Keskikuva: Kolme eri parametrisovitusta θ_{ML}, θ_2, θ_3. Parametrit estimoidaan pienimmän neliösumman kriteerin mukaisesti ja $f_{ML}(x) = 0.64x + 0.91$ antaa parhaimman sovituksen. Oikea kuva: opittua funktiota f voidaan käyttää arvioimaan uudelle x_{new}:lle “sopiva” y_{new}. ◂ Takaisin kalvoihin

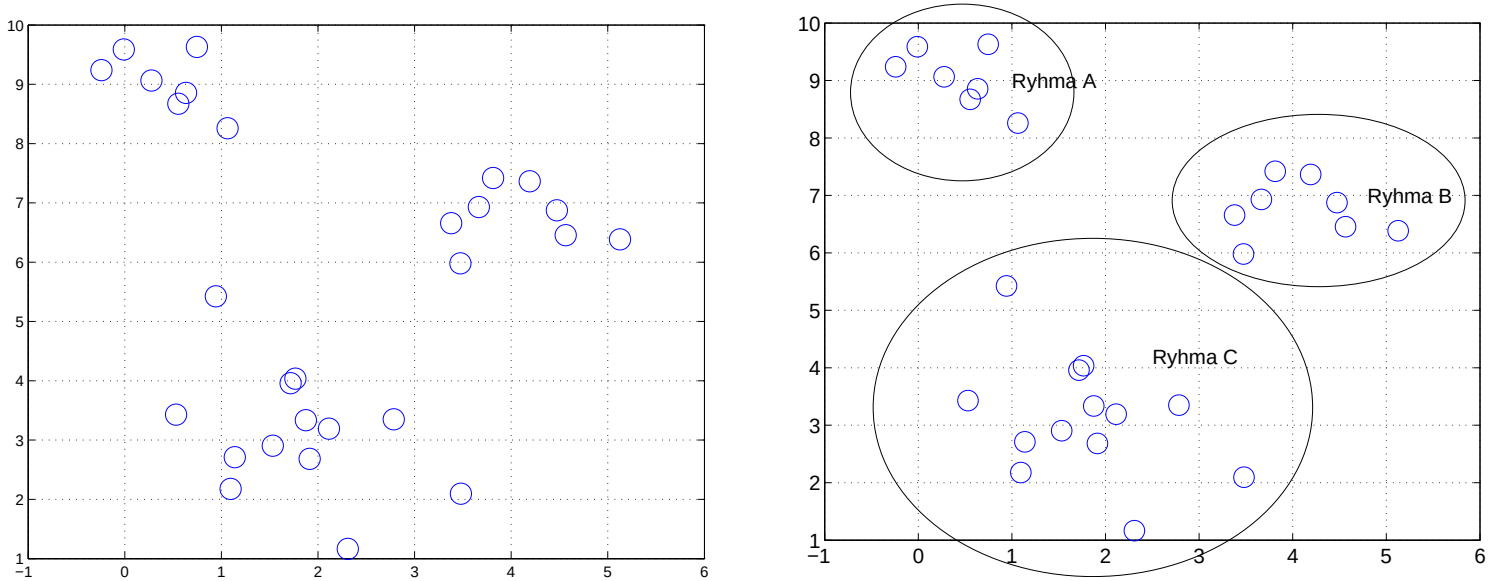
Esimerkki yksinkertaisesta luokittelusta



Kuva: Vasemmalla 2-ulotteinen data-aineisto, jossa lisänä luokkainformaatio jokaisesta datapisteestä joko A (pallo) tai B (neliö). Datasta on opittu luokittelija (luokkaraja). Oikealla luokittelijaa käytetään antamaan "sopiva" luokka-arvo uudelle datapisteelle x_{new}. Tässä esimerkissä luokittelija voi tuntua antavan "väärän" tuloksen.

Takaisin kalvoihin

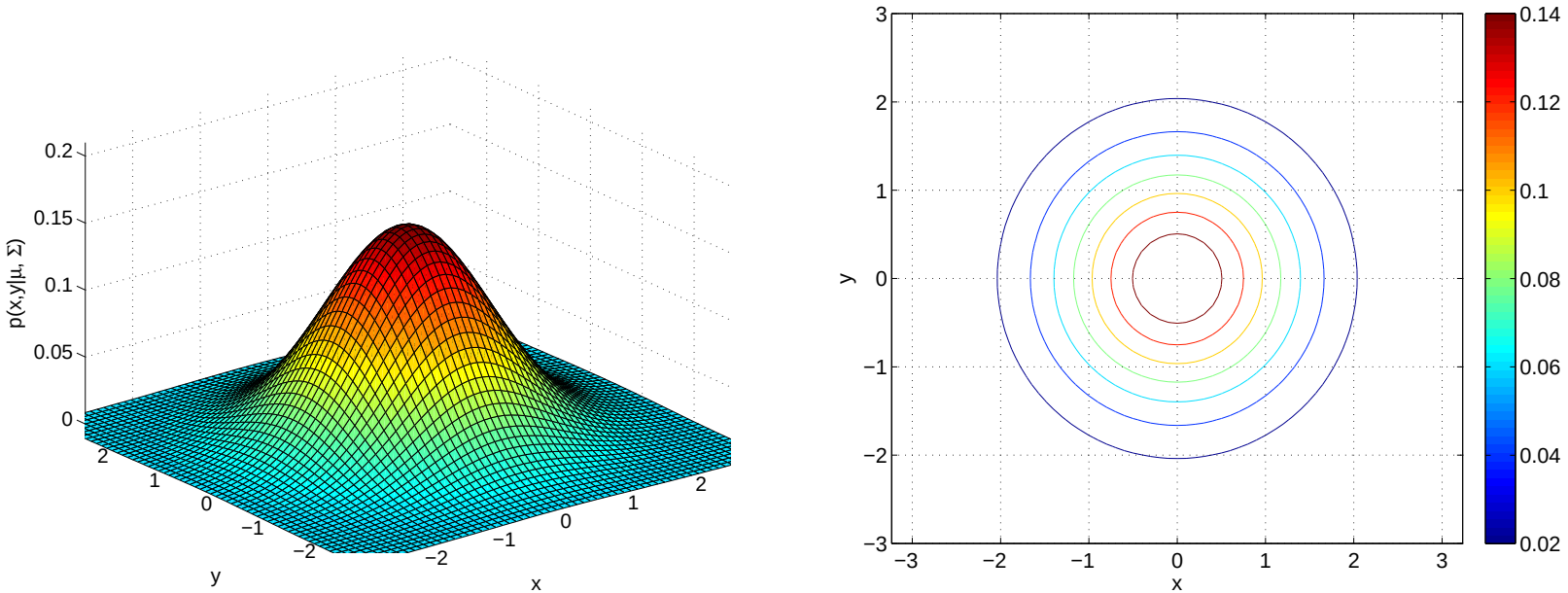
Esimerkki ryhmittelystä



Kuva: Vasemmalla 2-ulotteista dataa ilman luokkainformaatiota. Oikealla näytteet ryhmitelty kolmeen ryppääseen, joita aletaan kutsua ryhmiksi A, B, C. ◂ Takaisin kalvoihin

Esimerkki 2-ulotteisesta normaalijakaumasta

$\mu_1 = 0$, $\mu_2 = 0$, symmetrinen diagonaalinen **C**, jossa lisäksi $\sigma_1 = \sigma_2$

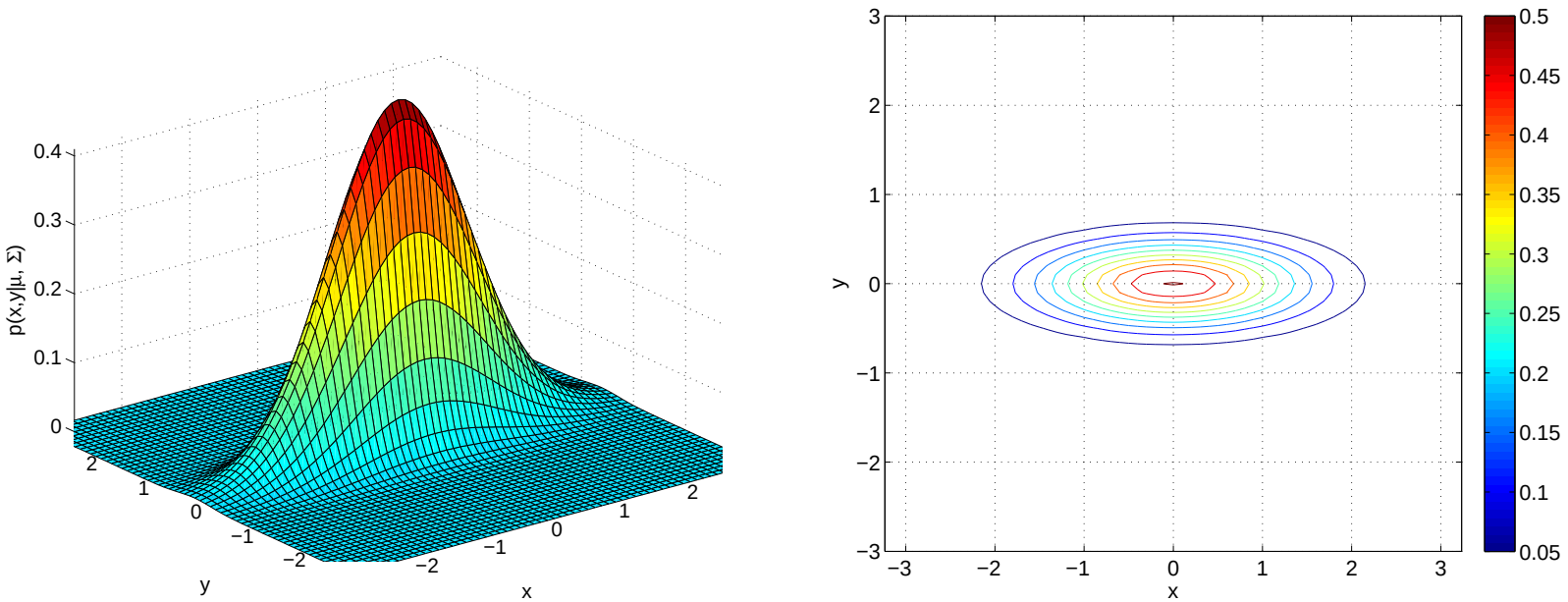


Kuva: Esimerkki 2-ulotteisesta normaalijakaumasta: $\mu_1 = 0$, $\mu_2 = 0$, symmetrinen diagonaalinen **C**, jossa lisäksi $\sigma_1 = \sigma_2$: $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Tässä esim. $p(\begin{bmatrix} 0 & 0 \end{bmatrix}^T) \approx 0.159$ ja $p(\begin{bmatrix} 1 & -2 \end{bmatrix}^T) \approx 0.0131$.

toinen esimerkki | kolmas esimerkki | Takaisin kalvoihin

Esimerkki 2-ulotteisesta normaalijakaumasta

$\mu_1 = 0$, $\mu_2 = 0$, symmetrinen diagonaalinen **C**, jossa $\sigma_1 \neq \sigma_2$



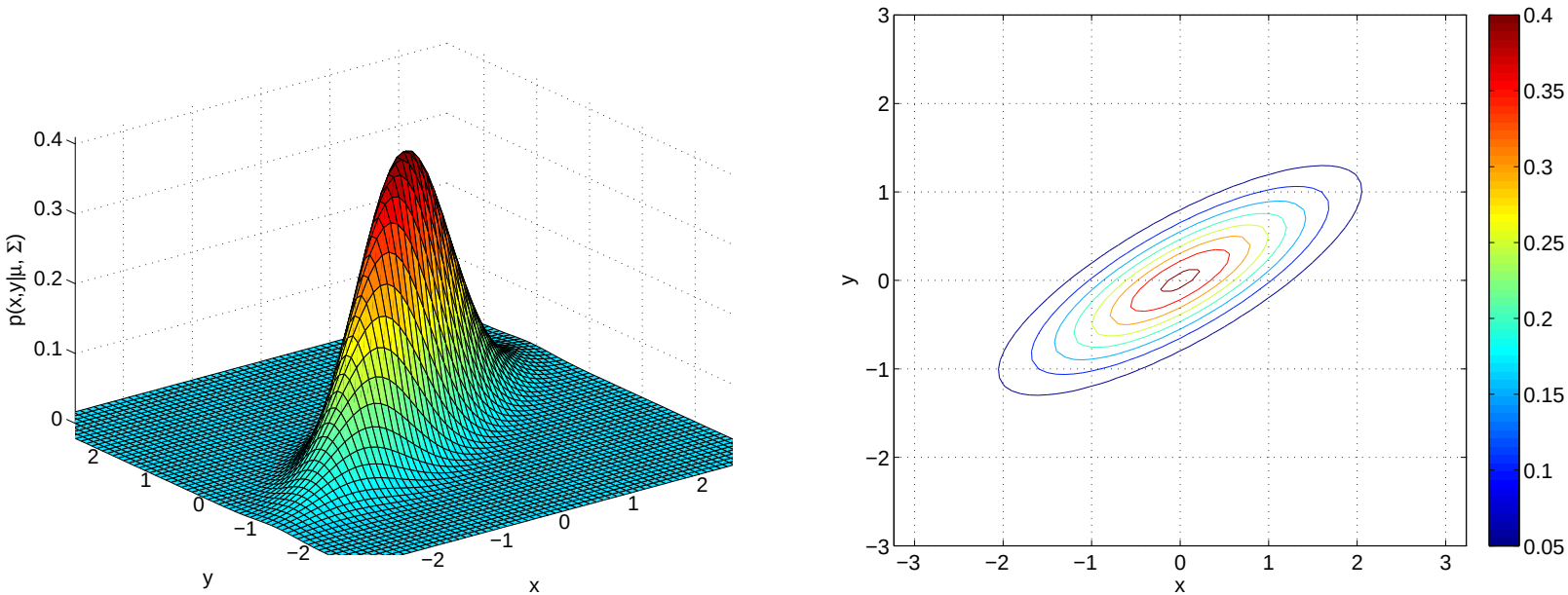
Kuva: Esimerkki 2-ulotteisesta normaalijakaumasta: $\mu_1 = 0$, $\mu_2 = 0$, symmetrinen diagonaalinen **C**, jossa $\sigma_1 \neq \sigma_2$: $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}$.

Tässä esim. $p(\begin{bmatrix} 0 & 0 \end{bmatrix}^T) \approx 0.503$ ja $p(\begin{bmatrix} 1 & -0.3 \end{bmatrix}^T) \approx 0.195$.

▸ ensimmäinen esimerkki ▸ kolmas esimerkki ◂ Takaisin kalvoihin

Esimerkki 2-ulotteisesta normaalijakaumasta

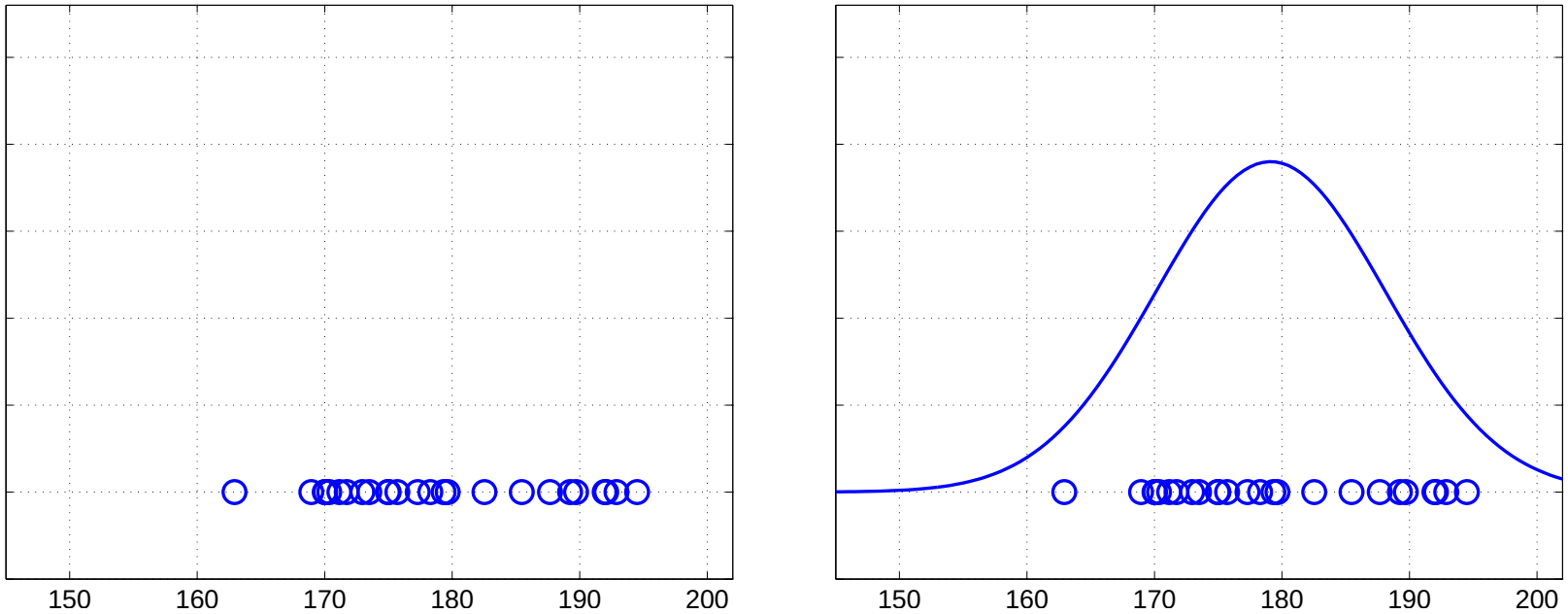
$\mu_1 = 0$, $\mu_2 = 0$, symmetrinen ei-diagonaalinen **C**



Kuva: Esimerkki 2-ulotteisesta normaalijakaumasta: $\mu_1 = 0$, $\mu_2 = 0$, symmetrinen ei-diagonaalinen **C**: $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.4 \end{bmatrix}$. Tässä esim. $p(\begin{bmatrix} 0 & 0 \end{bmatrix}^T) \approx 0.411$ ja $p(\begin{bmatrix} 1 & 1 \end{bmatrix}^T) \approx 0.108$.

▸ ensimmäinen esimerkki ▸ toinen esimerkki ◂ Takaisin kalvoihin

Data **X** ja siitä eräs tiheysfunktio $p(x)$



Kuva: Vasemmalla 1-ulotteinen datamatriisi **X**, jossa 25 näytettä, esimerkiksi ihmisten pituuksia (cm). Oikealla siihen sovitettu tiheysfunktio $p(x)$, joka tällä kertaa normaalijakauma. Normaalijakauman parametrit $\hat{\mu}$ ja $\hat{\sigma}$ on *estimoitu* datasta. Silmämääräisesti $\hat{\mu} \approx 179$ ja $\hat{\sigma} \approx 9$. ◂ Takaisin kalvoihin

Suurimman uskottavuuden menetelmä
Maximum likelihood method by Milton and Arnold

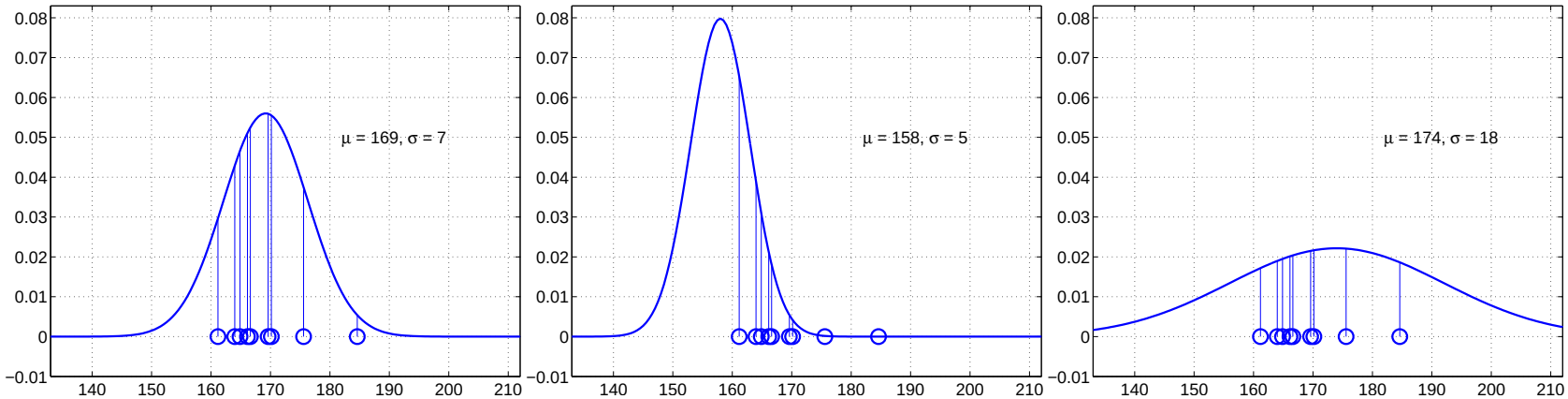
1. Obtain a random sample $x(1), x(2), \ldots, x(n)$ from the distribution of a random variable X with density p and associated parameter θ
2. Define a likelihood function $L(\theta)$ by
$$L(\theta) = \prod_{i=1}^{n} p(x(i))$$
3. Find the expression for θ that maximizes the likelihood function. This can be done directly or by maximizing $\ln L(\theta)$
4. Replace θ by $\hat{\theta}$ to obtain an expression for ML estimator for θ
5. Find the observed value of this estimator for a given sample

p. 233. Milton, Arnold: Introduction to probability and statistics. Third edition. McGraw-Hill, 1995.

◂ Takaisin kalvoihin

ML:n perusajatus

“Valitaan todennäköisimmin datan generoinut jakauma”

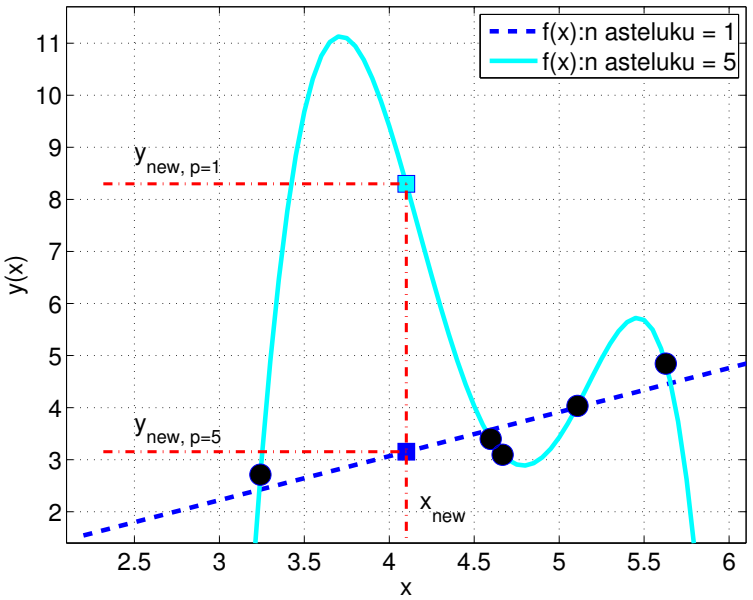


Kuva: Annettuna datajoukko $x(1), \ldots, x(9) \in \mathbb{R}^1$ palloina ja kolme gaussista tiheysfunktiota eri parametriyhdistelmillä $\boldsymbol{\theta}_1 = \{\mu = 169,\ \sigma = 7\}$, $\boldsymbol{\theta}_2 = \{\mu = 158,\ \sigma = 5\}$ ja $\boldsymbol{\theta}_3 = \{\mu = 174,\ \sigma = 18\}$. Lasketaan uskottavuusfunktio $L(\boldsymbol{\theta}) = p(x(1)|\boldsymbol{\theta}) \cdot p(x(2)|\boldsymbol{\theta}) \cdot \ldots \cdot p(x(9)|\boldsymbol{\theta})$. Ensimmäinen tuottaa selvästi suurimman $L(\boldsymbol{\theta})$:n arvon annetulla datasetillä: $L(\boldsymbol{\theta}_1) = 0.029 \cdot 0.042 \cdot \ldots \cdot 0.005 > L(\boldsymbol{\theta}_3) > L(\boldsymbol{\theta}_2)$. Valitaan siten $\boldsymbol{\theta}_1 = \{\mu = 169,\ \sigma = 7\}$ (*näistä kolmesta vaihtoehdosta!*), koska $\boldsymbol{\theta}_1$ on todennäköisimmin generoinut datan $\mathbf{X}$.

◂ Takaisin kalvoihin

Esimerkki lineaarisesta regressiosta ja ylioppimisesta

Polynomisovitus asteilla p=1 ja p=5 pieneen datamäärään N=5



Kuva: Regressiosovitus $y(i) = f(\mathbf{x}(i), \theta) + \epsilon(i)$. Tunnetaan viisi havaintoa $\{x(i), y(i)\}_i$ (mustat pallot). Asteluvun $p = 1$ polynomifunktio $f_{p=1}(.)$ (sininen katkoviiva) sovittuu melko hyvin (pienehköt virheet $(y(i) - f(x(i), \theta_1))^2$). Sen sijaan polynomisovitus $f_{p=5}(.)$ asteluvulla $p = 5$ vie neliövirheen nollaan, mutta on ylioppinut tunnettuun dataan ja siten huono uusille havainnoille (esim. x_{new}).

Takaisin kalvoihin