1. TODENNÄKÖISYYSJAKAUMIEN ESTIMOINTI

Edellä esitelty Bayesiläinen luokittelusääntö ('Bayes Decision Theory') on optimaalinen tapa suorittaa luokittelu, kun luokkien tnjakaumat tunnetaan

Käytännössä tnjakaumia ei tunneta, vaan ne on estimoitava havainnoista

Voi olla, että tunnetaan tnjakauman tyyppi (esim. normaalijakauma), mutta ei parametrejä (esim. odotusarvo, varianssi); tilanne voi olla myös täysin päinvastainen

Seuraavaksi käydään läpi menetelmiä, joiden avulla voidaan estimoida tilastollisessa tunnistuksessa tarvittavia tnjakaumia

1.1 Suurimman uskottavuuden menetelmä

'Maximum Likelihood Estimation', ML-estimaatti

Tarkastellaan M:n luokan tunnistusongelmaa

Tehdään seuraavat oletukset:

- Ol., että tunnetaan tnjakaumien $p(\mathbf{x}|\omega_i)$ tyyppi ja että ne voidaan määrittää parametrivektoreiden $\mathbf{\Theta}_i$ avulla. Merkitään siis $p(\mathbf{x}|\omega_i; \mathbf{\Theta}_i)$
- Ol., että eri luokista tehdyt havainnot eivät riipu toisistaan ja tnjakaumat voidaan estimoida erikseen luokkakohtaisesti
- Ol., että tehdyt havainnot ovat toisistaan riippumattomia myös luokkien sisällä

Ongelma: kuinka valitaan tnjakauman $p(\mathbf{x}|\mathbf{\Theta})$ parametrit $\mathbf{\Theta}$, kun on tehty havainnot $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$?

Vastaus: Maksimoidaan parametrien uskottavuusfunktio ('likelihood function') $p(\mathbf{X}; \mathbf{\Theta})$:

$$p(\mathbf{X}; \mathbf{\Theta}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{\Theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k; \mathbf{\Theta}) \tag{1}$$

Eli valitaan estimaatti $\hat{\mathbf{\Theta}}_{ML}$ seuraavasti:

$$\hat{\mathbf{\Theta}}_{ML} = \arg\max_{\mathbf{\Theta}} \prod_{k=1}^{N} p(\mathbf{x}_k; \mathbf{\Theta}) \tag{2}$$

Välttämätön ehto optimiratkaisulle $\hat{\mathbf{\Theta}}_{ML}$:

$$\frac{\partial \prod_{k=1}^{N} p(\mathbf{x}_k; \mathbf{\Theta})}{\partial \mathbf{\Theta}} = \mathbf{0} \tag{3}$$

Voidaan yhtä hyvin tarkastella uskottavuusfunktion logaritmiä ('loglikelihood function') $L(\mathbf{\Theta})$:

$$L(\mathbf{\Theta}) = \ln(\prod_{k=1}^{N} p(\mathbf{x}_k; \mathbf{\Theta})) = \sum_{k=1}^{N} \ln(p(\mathbf{x}_k; \mathbf{\Theta})) \tag{4}$$

Silloin ehto (3) saa muodon

$$\begin{aligned} \hat{\mathbf{\Theta}}_{ML} : \frac{\partial L(\mathbf{\Theta})}{\partial \mathbf{\Theta}} &= \sum_{k=1}^{N} \frac{\partial \ln(p(\mathbf{x}_k; \mathbf{\Theta}))}{\partial \mathbf{\Theta}} \\ &= \sum_{k=1}^{N} \frac{1}{p(\mathbf{x}_k; \mathbf{\Theta})} \frac{\partial p(\mathbf{x}_k; \mathbf{\Theta})}{\partial \mathbf{\Theta}} \\ &= \mathbf{0} \end{aligned} \tag{5}$$

1.2 Suurimman a posteriori tn:n menetelmä

'Maximum *A Posteriori* Probability Estimation', MAP-estimaatti

Tarkastellaan M:n luokan tunnistusongelma ja tehdään samanlaiset lähtöoletuk
kuin ML-estimaatin tapauksessa

Maksimoidaan parametrien uskottavuusfunktion $p(\mathbf{X};\mathbf{\Theta})$ sijasta parametrien *a posteriori* tn:ttä $p(\mathbf{\Theta}|\mathbf{X})$

Bayes-säännön perusteella:

$$p(\mathbf{\Theta}|\mathbf{X}) = \frac{p(\mathbf{\Theta})p(\mathbf{X}|\mathbf{\Theta})}{p(\mathbf{X})} \tag{6}$$

MAP-estimaatti $\hat{\mathbf{\Theta}}_{MAP}$ löytyy $p(\mathbf{\Theta}|\mathbf{X})$:n (tai sen logaritmin!) maksimikohdasta eli

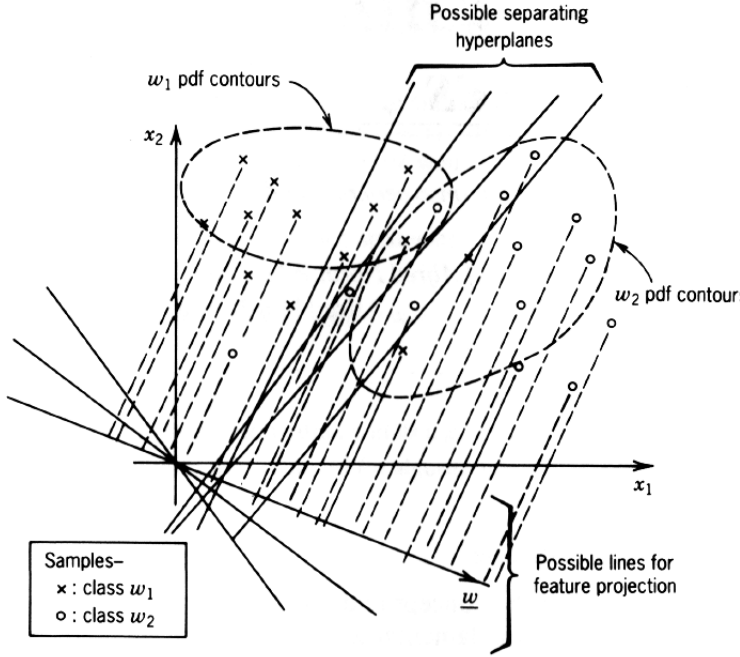
$$\hat{\mathbf{\Theta}}_{MAP} : \frac{\partial p(\mathbf{\Theta}|\mathbf{X})}{\partial \mathbf{\Theta}} = \frac{\partial p(\mathbf{\Theta})p(\mathbf{X}|\mathbf{\Theta})}{\partial \mathbf{\Theta}} = \mathbf{0} \tag{7}$$

Ero ML- ja MAP-estimaatin välillä on *a priori* tn:n $p(\mathbf{\Theta})$ käytössä. Jos $p(\mathbf{\Theta})$ on tasajakauma, ML- ja MAP-estimaatit ovat samat. Jos $p(\mathbf{\Theta})$:iin liittyvä

varianssi on suuri, estimaatit ovat lähestulkoon samat

Silloin kun $\frac{\sigma_\mu^2}{\sigma^2} \gg 1$ ML- ja MAP-estimaatit ovat lähestulkoon samat

Seuraavassa kuvassa kaksi tapausta, joissa erilaiset uskottavuusfunktiot ja *a priori* tnjakaumat. Tapauksessa a) ML- ja MAP-estimaatit ovat samankaltaiset, tapauksessa b) niiden ero on selvä


Possible separating
hyperplanes
w_1 pdf contours
x_2
w_2 pdf contours
x_1
Samples–
× : class w_1
o : class w_2
$\underline{w}$
Possible lines for
feature projection

1.3 Bayesiläinen päättely

ML- ja MAP-menetelmät löytävät tnjakaumien $p(\mathbf{x}|\omega_i; \mathbf{\Theta}_i)$ parametreille arvot käyttäen hyväksi tehtyjä havaintoja $\mathbf{X}_i$ ja *a priori* tnjakaumia $p(\mathbf{\Theta}_i)$

Miksi emme suoraan laskisi tnjakaumia $p(\mathbf{x}|\mathbf{X_i})$ ja suorittaisi luokittelua näiden avulla?

Tehdään taas samat lähtöoletukset kuin ML- ja MAP-estimaattien tapauksissa

Tiedetään, että

$$p(\mathbf{x}|\mathbf{X}) = \int p(\mathbf{x}|\mathbf{\Theta})p(\mathbf{\Theta}|\mathbf{X})d\mathbf{\Theta}, \tag{8}$$

missä

$$p(\mathbf{\Theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta})}{\int p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta})d\mathbf{\Theta}} \tag{9}$$

$$p(\mathbf{X}|\mathbf{\Theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k|\mathbf{\Theta}) \tag{10}$$

Tämän menetelmän haittapuoli on sen monimutkaisuus – analyyttinen ratkaisu on olemassa vain erikoistapauksille

Kun havaintojen määrä lähestyy ääretöntä, bayesiläisellä päättelyllä saatu normaalijakauman estimaatti lähestyy piikkiä, jonka huipun kohta on ML-estimaatti. Tämä tulos yleistyy useille muillekin paljon käytetyille tnjakaumille

Huom! Kun N lähestyy ääretöntä, ML- ja MAP-menetelmillä sekä bayesiläisellä päättelyllä saadut tulokset lähestyvät toisiaan. Menetelmien erot ovat merkittäviä, kun käytettävissä on vain rajallisesti havaintoja

1.4 Maksimientropiamenetelmä

Edellä esitetyissä menetelmissä oletettiin tnjakauman tyyppi tunnetuksi ja estimoitiin sille sopivat parametrit

Entä jos tunnetaan joitain tilastollisia tunnuslukuja, mutta ei tiedetä mikä tnjakaumatyyppi on kyseessä?

Valitaan sellainen tnjakauma, jolla on tunnetut ominaisuudet ja maksimaalinen entropia (ei aina mikään helppo ongelma!)

Tnjakauman $p(\mathbf{x})$ entropia H on määritelty seuraavasti:

$$H = -\int_{\mathbf{x}} p(\mathbf{x}) \ln(p(\mathbf{x})) d\mathbf{x} \tag{11}$$

Entropia kuvaa ilmiön satunnaisuutta tai yllätyksellisyyttä.

Tnjakauman entropian maksimointi vastaa minimimaalisinta *a priori* -tiedon käyttöä.

Huom! Voidaan osoittaa, että jos tunnetaan odotusarvo ja varianssi, maksimientropiaratkaisu on normaalijakauma

1.5 Mikstuurimallit

Mikstuurimalleissa lähdetään siitä, että $p(x)$:n muodostumiseen vaikuttaa joukko jakaumia. Tuntematon $p(x)$ mallinnetaan seuraavanlaisten tiheysjakaumien lineaariyhdistelmänä:

$$p(x) = \sum_{j=1}^{J} p(x|j)P_j \tag{12}$$

missä

$$\sum_{j=1}^{J} P_j = 1, \int_x p(x|j)dx = 1 \tag{13}$$

1.6 EM-algoritmi

EM = Expectation Maximization

EM-algoritmi sopii erityisen hyvin tilanteisiin, joissa on puuttuvaa dataa.

Algoritmi maksimoi loglikelihood-funktion odotusarvon ehdolla havainnot ja nykyinen estimaatti. Algoritmissa toistetaan kahta askelta: E (expectation) ja M (maximization).

1.7 Epäparametriset menetelmät

Edellä esitetyissä menetelmissä oletettiin, että tnjakauma voidaan määrittää parametrivektorin avulla. Seuraavaksi tarkastellaan menetelmiä, joissa tätä oletusta ei tehdä.

Sekä Parzen-ikkunat ja k:n lähimmän naapurin menetelmä ovat eräänlaisia variaatioita tnjakauman approksimoinnista histogrammin avulla.

Tnjakaumaa voidaan approksimoida histogrammin avulla seuraavasti:

- Tarkastellaan vain yhtä piirrettä x , eli 1-ulotteista tapausta, ja pyritään approksimoimaan tnjakauma $p(x)$
- Jaetaan x-akseli h:n pituisiksi intervalleiksi
- Approksimoidaan tn P, että x:n arvo osuu tietylle intervallille. Olkoot N havaintojen lkm ja k_N intervallille osuneiden havaintojen lkm. Silloin $P \approx k_N/N$
- Kun N lähestyy ääretöntä,

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{1}{h}\frac{k_N}{N},\ |x - \hat{x}| \leq \frac{h}{2}, \tag{14}$$

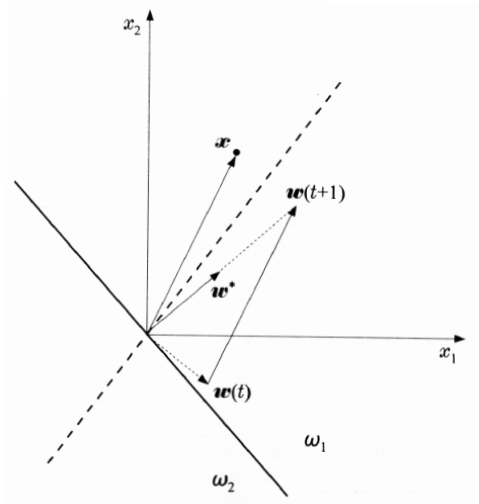
  missä $\hat{x}$ on intervallin keskikohta
- Approksimaatio on hyvä silloin, kun $p(x)$ on jatkuva ja h on riittävän pieni eli oletus $p(x) =$ vakio intervallilla on järkevä

- Voidaan osoittaa, että $\hat{p}(x) \to p(x)$, jos $N \to \infty$ ja seuraavat oletukset toteutuvat:

$$h_N \to 0$$
$$k_N \to \infty$$
$$\frac{k_N}{N} \to 0$$

- Käytännössä N on rajoitettu ja sopiva h pitää valita itse

Seuraavassa kuvassa tnjakaumaa approksimoidaan histogrammin avulla. Tapauksessa a) on valittu h:lle pieni ja tapauksessa b) suuri arvo


x_2
$\boldsymbol{x}$
$\boldsymbol{w}(t+1)$
$\boldsymbol{w}^*$
x_1
$\boldsymbol{w}(t)$
ω_1
ω_2

Parzen-ikkunat

Moniulotteisessa tapauksessa l-ulotteinen piirreavaruus jaetaan hyperkuutioksi, joiden tilavuus on h^l

Määritellään jokaiselle havainnolle $\mathbf{x}_i$ kantafunktio $\Phi(\mathbf{x}_i)$ seuraavasti:

$$\Phi(\mathbf{x}_i) = \begin{cases} 1, & \text{kun } |x_{ij}| \leq 1/2 \\ 0, & \text{muulloin} \end{cases}, \tag{15}$$

missä x_{ij} on $\mathbf{x}_i$:n j. komponentti

Kaava (14) voidaan esittää tällaisten funktioiden summana:

$$\hat{p}(\mathbf{x}) = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N}\Phi\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)\right) \tag{16}$$

Edellä jatkuvaa tnjakaumaa approksimoitiin epäjatkuvilla kuutiolla. Jatkuva approksimaatio saadaan, kun käytetään jatkuvia kantafunktioita $\Phi(\cdot)$

Kantafunktiot valitaan siten, että

$$\begin{aligned} \Phi(\mathbf{x}) &\geq 0 \\ \int_{\mathbf{x}} \Phi(\mathbf{x}) d\mathbf{x} &= 1 \end{aligned} \tag{17}$$

Kantafunktiot voivat olla esim. eksponentiaalisia, $N(\mathbf{0}, \mathbf{I})$

Mikä on $\hat{p}(\mathbf{x})$:n odotusarvo eli kuinka approksimaatio käyttäytyy, kun $N \to \infty$?

- $\hat{p}(\mathbf{x})$ on määritelty havaintojen $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ avulla, jotka noudattavat todellista tnjakaumaa $p(\mathbf{x})$
- Lasketaan $\hat{p}(\mathbf{x})$:n odotusarvo tnjakauman $p(\mathbf{x})$ suhteen:

$$\mathsf{E}[\hat{p}(\mathbf{x})] = \frac{1}{h^l}(\frac{1}{N}\sum_{i=1}^{N} \mathsf{E}[\Phi(\frac{\mathbf{x}_i - \mathbf{x}}{h})]) = \int_{\xi} \frac{1}{h^l}\Phi(\frac{\xi - \mathbf{x}}{h}) p(\xi) d\xi \tag{18}$$

- Yllä olevasta kaavasta nähdään, että $\hat{p}(\mathbf{x})$ on tasoitettu ('smoothed') versio todellisesta jakaumasta

- Kun $h \to 0$, $\frac{1}{h^l}\Phi(\frac{\xi-\mathbf{x}}{h})$ lähestyy deltafunktiota $\delta(\xi - \mathbf{x})$ ja nähdään, että $\hat{p}(\mathbf{x})$ on harhaton estimaatti $p(\mathbf{x})$:lle riippumatta N:stä

Jos N on kiinnitetty, estimaatin $\hat{p}(\mathbf{x})$ varianssi kasvaa, kun h pienenee.

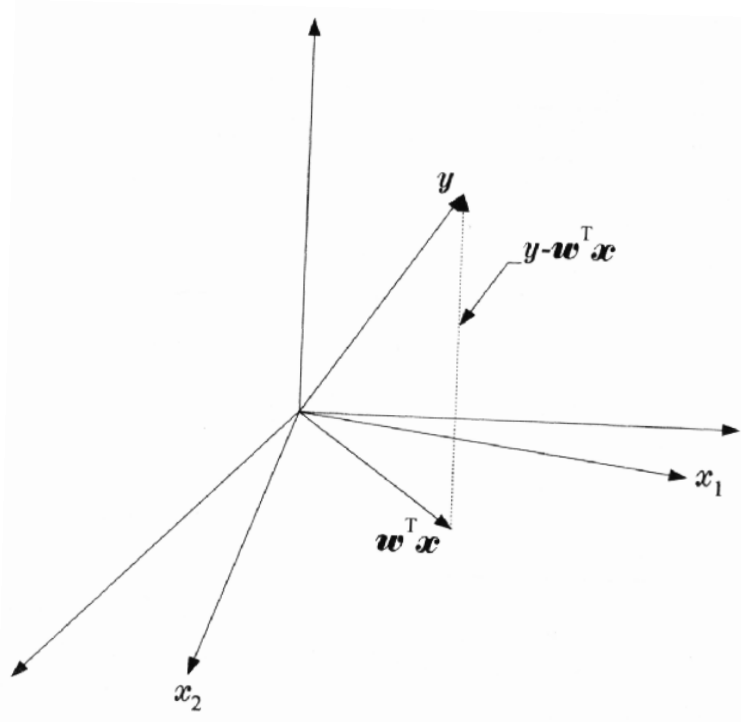
Jos h on kiinnitetty, estimaatin $\hat{p}(\mathbf{x})$ varianssi pienenee, kun $N \to \infty$

Useimmilla kantafunktiolla saatu estimaatti $\hat{p}(\mathbf{x})$ on harhaton ja asymptoottisesti konsistentti, jos $h \to 0$ siten, että $hN \to \infty$, kun $N \to \infty$

Käytännössä N on rajoitettu ja h joudutaan valitsemaan itse, esim. minimoimalla luokitteluvirhettä.

Mikäli ei haluta tinkiä estimaatin ominaisuuksista, tarvittavien havaintojen N lukumäärä kasvaa eksponentiaalisesti piirrevektoreiden dimension l suhteen ('curse of dimensionality!')

Seuraavassa kuvassa on esitetty Parzen-ikkunoilla saatuja tnjakauman (katkoviiva) approksimaatioita. Kantafunktiot ovat normaalijakaumia ja $h = 0.1$. Tapauksessa a) $N = 1000$ ja tapauksessa b) $N = 20\,000$


y
y-w^T x
x_1
w^T x
x_2

k:n lähimmän naapurin menetelmä

Edellisessä menetelmässä kantafunktiot olivat samanlaisia riippumatta siitä oliko tarkasteltavassa piirreavaruuden osassa havaintoja tiheässä vai harvassa.

Yleisempi versio kaavalle (14) :

$$\hat{p}(\mathbf{x}) = \frac{k}{NV(\mathbf{x})}, \tag{19}$$

missä $V(\mathbf{x})$ on $\mathbf{x}$:stä riippuva tilavuus. Eli kiinnitetään ensin k ja lasketaan vasta sitten havaintojen viemä tilavuus

Voidaan osoittaa, että estimaatti on harhaton ja asymptoottisesti konsistentti, jos $k \to \infty$ ja $k/N \to 0$, kun $N \to \infty$

Saatua estimaattia kutsutaan tnjakauman k:n lähimmän naapurin estimaatiksi.

k:n lähimmän naapurin päätössääntö on suboptimaalinen variaatio k:n lähimmän naapurin estimaatista:

- Etsi N:n opetusnäytteen joukosta luokiteltavan näytteen $\mathbf{x}$ k lähintä naapuria. Yleensä k on pariton eikä se ole luokkien lukumäärän M monikerta
- Laske kuinka moni (k_i) lähimmistä naapureista kuuluu luokkaan ω_i
- Valitse se luokka, jolle k_i on suurin

Lähimmän naapurin menetelmän luokitteluvirheelle P_{NN} voidaan laskea seuraavat teoreettiset rajat, kun $N \to \infty$:

$$P_{Bayes} \leq P_{NN} \leq P_{Bayes}(2 - \frac{M}{M-1}P_{Bayes}) \leq 2P_{Bayes}, \qquad (20)$$

missä P_{Bayes} on Bayes-säännön tuottama optimaalinen luokitteluvirhe ja M on luokkien lkm

Mikäli N on suuri, k:n lähimmän naapurin päätössääntö toimii paremmin isommilla k:n arvoilla kuin 1.

Käytännössä k:n lähimmän naapurin sääntö toimii usein erittäin hyvin yksinkertaisuudestaan huolimatta, mutta on laskennallisesti raskas suurilla N:n arvoilla.