1. LINEAARISET LUOKITTIMET

Edellisillä luennoilla tarkasteltiin luokitteluongelmaa tnjakaumien avulla ja esiteltiin menetelmiä, miten tarvittavat tnjakaumat voidaan estimoida. Tavoitteena oli löytää päätössääntö, joka minimoi luokitteluvirhetn:n tai riskin

Tietyin tnjakaumista tehdyin oletuksin näin saatu luokitin on lineaarinen

Nyt ei oteta kantaa luokkiin liittyviin tnjakaumiin, vaan oletetaan, että kaikki havainnot voidaan luokitella oikein/riittävän hyvin lineaarisilla luokittimilla

Lineaaristen luokittimien etuna on niiden yksinkertaisuus ja laskennallinen keveys

1.1 Lineaariset diskriminanttifunktiot

Lineaarinen luokitin jakaa piirreavaruuden eri luokkia vastaaviksi päätösalueiksi hypertasojen avulla

Lineaarinen diskriminanttifunktio $g(\mathbf{x})$:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0, \tag{1}$$

missä $\mathbf{w} = [w_1, \ldots, w_l]$ on painovektori ja w_0 on kynnysarvo ('threshold')

Tarkastellaan kahta pistettä, $\mathbf{x}_1$ ja $\mathbf{x}_2$, diskriminanttifunktion määrittelemällä hypertasolla:

$$\begin{aligned} 0 =& \mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0 \Rightarrow \\ & \mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0 \end{aligned} \tag{2}$$

Edellisestä nähdään, että painovektori $\mathbf{w}$ on ortogonaalinen hypertasoon nähden

Toisella puolella hypertasoa diskriminanttifunktio saa positiivisia arvoja, toisella puolella negatiivisia arvoja

Kahden luokan tapauksessa voidaan **luokittelusääntö** kirjoittaa seuraavasti:

- Merkitään $\tilde{\mathbf{x}} = [\mathbf{x}^T, 1]^T$ ja $\tilde{\mathbf{w}} = [\mathbf{w}^T, w_0]^T$
- Silloin $g(\mathbf{x}) \equiv g(\tilde{\mathbf{x}}) = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}$
- Valitaan luokka seuraavasti:

$$\begin{aligned} \omega_1 &: \ \tilde{\mathbf{w}}^T\tilde{\mathbf{x}} > 0 \\ \omega_2 &: \ \tilde{\mathbf{w}}^T\tilde{\mathbf{x}} < 0 \end{aligned} \tag{3}$$

Jos voidaan valita $\tilde{\mathbf{w}}$ siten, että edellinen päätössääntö ei tuota ainuttakaan luokitteluvirhettä, luokat ovat *lineaarisesti separoituvat*

Seuraavaksi käydään läpi menetelmiä, joiden avulla voidaan määrätä diskriminanttifunktion painokertoimet

Tästä eteenpäin oletetaan, että piirrevektoreihin on liitetty loppuun vakiotermi, kuten edellä esitetyssä päätössäännössä, ellei toisin mainita

1.2 Perseptroni-algoritmi

(Perseptroni-algoritmin nimi tulee sitä, että se kehitettiin alunperin aivojen neuronimallien, 'perceptrons', opetukseen. Niistä lisää myöhemmin!)

Ol., että luokat ovat lineaarisesti separoituvia

Tarkastellaan kahden luokan tapausta

Valitaan diskriminanttifunktion painokertoimet $\mathbf{w}$ ratkaisemalla seuraava optimointiongelma:

- Minimoi (perseptroni)kustannusfunktio $J(\mathbf{w})$, joka on määritelty seuraavasti:
$$J(\mathbf{w}) = \sum_{\mathbf{x} \in Y} (\delta_x \mathbf{w}^T \mathbf{x}), \tag{4}$$
missä Y on niiden opetusnäytteiden osajoukko, jotka luokittuvat väärin $\mathbf{w}$ määrittämän hypertason perusteella. Muuttuja $\delta_x = -1$, kun $\mathbf{x} \in \omega_1$, ja $\delta_x = 1$, kun $\mathbf{x} \in \omega_2$
- Huom! $J(\mathbf{w})$ saa vain positiivisia arvoja. Sen arvo on nolla, jos ei synny

lainkaan luokitteluvirheitä

- Huom! $J(\mathbf{w})$ on jatkuva ja paloittain lineaarinen, $J(\mathbf{w})$:n gradientti ei ole määritelty niissä kohdin, joissa osajoukko Y muuttuu
- Edellisestä huolimatta, suoritetaan minimointi 'gradient descent'-henkisellä, iteratiivisellä menetelmällä:

$$\begin{aligned} \mathbf{w}(t+1) &= \mathbf{w}(t) - \rho_t \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}|_{\mathbf{w}(t)} \\ &= \mathbf{w}(t) - \rho_t \sum_{\mathbf{x} \in Y} \delta_x \mathbf{x} \end{aligned} \tag{5}$$

  missä on $\mathbf{w}(0)$ alustettu satunnaisesti ja ρ_t on positiivinen oppimiskerroin

Iterointia toistetaan, kunnes painovektorin arvo konvergoi ratkaisuun eli kaikki havainnot luokitellaan oikein

Ratkaisun löytyminen ja tarvittavien iteraatioaskelien lkm riippuu kertoimen ρ_t valinnasta

Vaikka $J(\mathbf{w})$:n gradientti ei ole määritelty kaikkialla, menetelmä löytää ratkaisun äärellisellä määrällä iteraatioaskelia, kun ρ_t valitaan seuraavasti:

$$\lim_{t\to\infty} \sum_{k=0}^{t} \rho_k = \infty$$

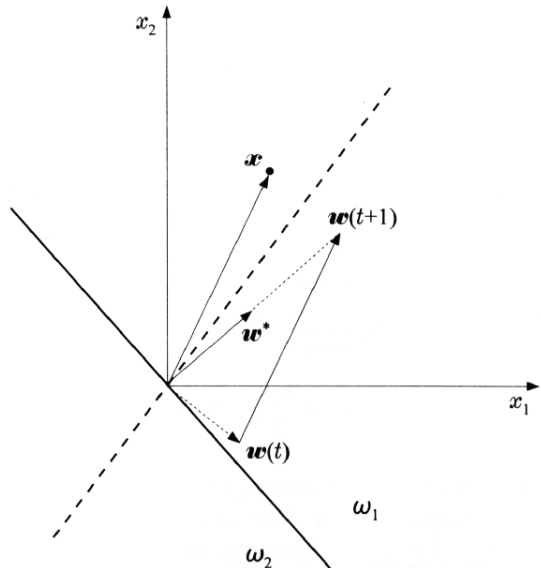
$$\lim_{t\to\infty} \sum_{k=0}^{t} \rho_k^2 < \infty$$

(todistus kirjassa)

Edelliset ehdot tarkoittavat käytännössä sitä, että ρ_t lähestyy nollaa, kun iteraatioaskelien lkm t kasvaa. Kerroin ρ_t ei saa kuitenkaan pienentyä liian nopeasti.

Sopivat arvot ρ_t:lle saadaan esim. seuraavasti: $\rho_t = c/t$, missä c on vakio. Kerroin ρ_t voi olla myös vakio, jos se on sopivasti rajoitettu

Perseptroni-säännön geometrinen tulkinta. Painovektorin päivitys, kun vain yksi piirrevektori luokittui väärin:



Perseptroni-säännön variaatioita

Edellä esitetyssä perseptroni-säännössä käytettiin jokaisella iteraatioaskeleella kaikkia N:ää opetusnäytettä. Painovektorin päivitys voidaan tehdä myös jokaisen havainnon perusteellä erikseen:

- Käydään opetusnäytteet läpi vuorotellen, päivitetään painovektoria seuraavasti:

$$\begin{aligned} &\mathbf{w}(t+1) = \mathbf{w}(t) + \rho\mathbf{x}_{(t)}, \quad \text{jos} \quad \mathbf{x}_{(t)} \in \omega_1 \quad \text{ja} \quad \mathbf{w}^T(t)\mathbf{x}_{(t)} \leq 0 \\ &\mathbf{w}(t+1) = \mathbf{w}(t) - \rho\mathbf{x}_{(t)}, \quad \text{jos} \quad \mathbf{x}_{(t)} \in \omega_2 \quad \text{ja} \quad \mathbf{w}^T(t)\mathbf{x}_{(t)} \geq 0 \\ &\mathbf{w}(t+1) = \mathbf{w}(t), \quad \text{muulloin} \end{aligned} \tag{6}$$

- Painovektoria päivitetään siis vain jos tarkasteltava opetusnäyte $\mathbf{x}_{(t)}$ luokittuu vääriin
- Opetusnäytteitä käydään läpi kunnes menetelmä konvergoi eli kaikki näytteet luokittuvat oikein
- Myös tämä menetelmä konvergoi äärellisellä määrällä iteraatioaskelia

Pocket-algoritmi

Ol., toisin kuin aikaisemmin, että luokat *eivät* ole lineaarisesti separoituvia

Seuraava menetelmä löytää lineaarisen diskriminanttifunktion, joka minimoi luokitteluvirheiden lkm:n:

- Alusta $\mathbf{w}(0)$ satunnaisesti. Lisäksi, alusta 'varasto'-vektori $\mathbf{w}_s$ ja laskurimuuttuja h_s nolliksi
- Päivitä painovektoria käyttäen perseptroni-sääntöä (5)
- Laske kuinka monta (h) opetusnäytettä luokittuu oikein käytettäessä päivitettyä painovektoria $\mathbf{w}(t+1)$
- Jos $h > h_s$, aseta $\mathbf{w}_s = \mathbf{w}(t+1)$ ja $h_s = h$
- Toista kunnes konvergoi

Keslerin konstruktio

Edellä esitetyissä menetelmissä tarkasteltiin vain kahden luokan tapausta

Keslerin konstruktion avulla näitä menetelmiä voidaan käyttää myös $M > 2$:n luokan tapauksessa

Piirrevektorin $\mathbf{x}$ luokittelupäätös tehdään lineaaristen diskriminanttifunktioiden avulla seuraavasti:

$$\omega_i : \ \mathbf{w}_i^T \mathbf{x} > \mathbf{w}_j^T \mathbf{x}, \ \forall j \neq i \tag{7}$$

Muodostetaan luokan ω_i jokaista opetusnäytettä kohti seuraavat $(l+1)M \times 1$-ulotteiset vektorit: $\mathbf{x}_{ij} = [\mathbf{0}^T, \ldots, \mathbf{x}^T, \ldots, -\mathbf{x}^T, \ldots, \mathbf{0}^T]^T$, $i \neq j$. Vektorit koostuvat siis M:stä blokista, joista $i.$ ja $j.$ ovat $\mathbf{x}$ ja $-\mathbf{x}$ ja muut ovat nollavektoreita

Vastaavasti luokkakohtaiset painovektorit kootaan yhdeksi vektoriksi: $\mathbf{w} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_M^T]^T$

Näiden avulla päätössääntö voidaan voidaan kirjoittaa uuteen muotoon:

$$\omega_i : \ \mathbf{w}^T \mathbf{x}_{ij} > 0 \ \forall j = 1, \ldots, M, \ j \neq i \tag{8}$$

Ongelmana on siis löytää painovektori $\mathbf{w}$, jonka positiivisella puolella ovat kaikki uudet vektorit $\mathbf{x}_{ij}$

Mikäli luokat ovat lineaarisesti separoituvia, voidaan painovektorin oppimiseen käyttää esim. perseptroni-sääntöä

Oppimisen jälkeen luokkakohtaiset painovektorit saadaan pilkkomalla $\mathbf{w}$ osiin

Huom! Vain painovektorin suunta on tärkeä, ei sen pituus. Edellä esitetyissä menetelmissä painovektorin pituus pyrkii kasvamaan: normalisoidaan painovektori yksikkövektoriksi jokaisen päivityksen jälkeen

1.3 Pienimmän neliön menetelmät

Usein tiedetään etukäteen, että luokat eivät ole lineaarisesti separoituvia, mutta silti halutaan käyttää luokitteluvirhetn:n kannalta suboptimaalista lineaarista luokitinta

Lineaarinen luokitin muodostetaan valitsemalla diskriminanttifunktion painovektori siten, että jokin ongelmaan sopiva kriteeri optimoituu

Yleensä määritellään diskriminanttifunktiolle tavoitearvot y erilaisille havainnoille $\mathbf{x}$ ja pyritään minimoimaan tavoitearvojen ja todellisten arvojen $g(\mathbf{x})$ neliöpoikkeamia ('Least Squares Methods')

Pienimmän neliösumman menetelmä

Pienimmän neliösumman menetelmässä minimoidaan kustannusfunktiota $J(\mathbf{w})$:

$$J(\mathbf{w}) = \sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2 \equiv \sum_{i=1}^{N} e_i^2, \tag{9}$$

missä N on opetusnäytteiden lkm ja y_i on $\mathbf{x}_i$:tä vastaava diskriminanttifunktion tavoitearvo, kahden luokan tapauksessa yleensä $y_i = \pm 1$

Kun derivoidaan $J(\mathbf{w})$ painovektorin $\mathbf{w}$ suhteen saadaan:

$$\begin{aligned} &\sum_{i=1}^{N}\mathbf{x}_i(y_i - \mathbf{x}_i^T\hat{\mathbf{w}}) = 0 \Rightarrow \\ &(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T)\hat{\mathbf{w}} = \sum_{i=1}^{N}(\mathbf{x}_i y_i) \end{aligned} \tag{10}$$

Käytetään seuraavia merkintöjä:

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \\ \mathbf{y} &= [y_1, \ldots, y_N]^T \end{aligned} \tag{11}$$

Kaava (10) voidaan kirjoittaa silloin matriisimuodossa:

$$\begin{aligned} &(\mathbf{X}^T\mathbf{X})\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y} \Rightarrow \\ &\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \end{aligned} \tag{12}$$

missä matriisi $\mathbf{X}^T\mathbf{X}$ on otoskorrelaatiomatriisi ja $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ on $\mathbf{X}$:n pseudoinverssi

Esimerkki: pienimmän neliösumman luokitin Tarkastellaan kahden luokan tapausta, jossa piirrevektorit ovat kaksiulotteisia. Luokista on tehty seuraavat havainnot:

$$\begin{aligned}\omega_1 : &[0.2, 0.7]^T\ [0.3, 0.3]^T\ [0.4, 0.5]^T \\ &[0.6, 0.5]^T\ [0.1, 0.4]^T \\ \omega_2 : &[0.4, 0.6]^T\ [0.6, 0.2]^T\ [0.7, 0.4]^T \\ &[0.8, 0.6]^T\ [0.7, 0.5]^T\end{aligned}$$

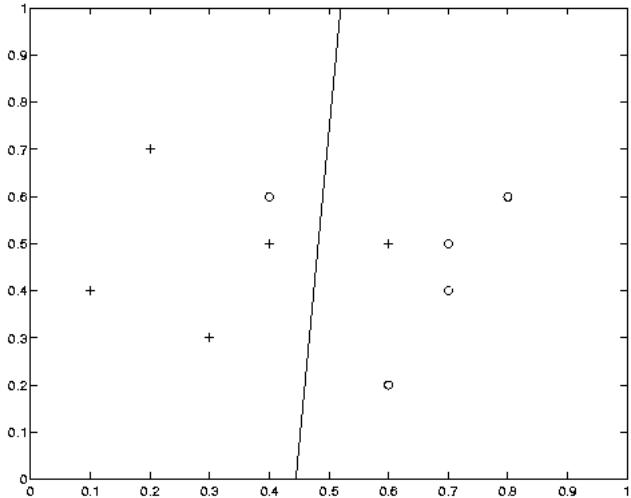
Luokat eivät ole lineaarisesti separoituvia. Yritetään löytää diskriminanttifunktio, joka on muotoa $g(\mathbf{x}; \mathbf{w}) = [\mathbf{x}, 1]^T\mathbf{w} = w_1x_1 + w_2x_2 + w_0$ ja joka minimoi virheiden neliösumman

Asetetaan diskriminanttifunktion tavoitearvoksi 1 tai -1, kun havainto on luokasta ω_1 tai ω_2

Silloin

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 2.8 & 2.24 & 4.8 \\ 2.24 & 2.41 & 4.7 \\ 4.8 & 4.7 & 10 \end{bmatrix}$$
$$\mathbf{X}^T\mathbf{y} = [-1.6, 0.1, 0.0]^T$$

ja ratkaisuksi saadaan kaavan (12) perusteella $\hat{\mathbf{w}} = [-3.218, 0.241, 1.431]^T$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

MSE-estimaatti

Seuraavaksi tarkastellaan neliöpoikkeamien summan sijasta neliöpoikkeaman odotusarvoa ('Mean Square Error Estimation')

Minimoitava kustannusfunktio $J(\mathbf{w})$:

$$J(\mathbf{w}) = \mathsf{E}[|y - \mathbf{x}^T\mathbf{w}|^2] \tag{13}$$

Kahden luokan tapauksessa, kun $y = \pm 1$, edellinen kaava voidaan kirjoittaa tnjakaumien avulla myös näin:

$$J(\mathbf{w}) = P(\omega_1)\int(1 - \mathbf{x}^T\mathbf{w})^2 p(\mathbf{x}|\omega_1)d\mathbf{x} + P(\omega_2)\int(1 + \mathbf{x}^T\mathbf{w})^2 p(\mathbf{x}|\omega_2)d\mathbf{x} \tag{14}$$

Välttämätön ehto $J(\mathbf{w})$:n minimille:

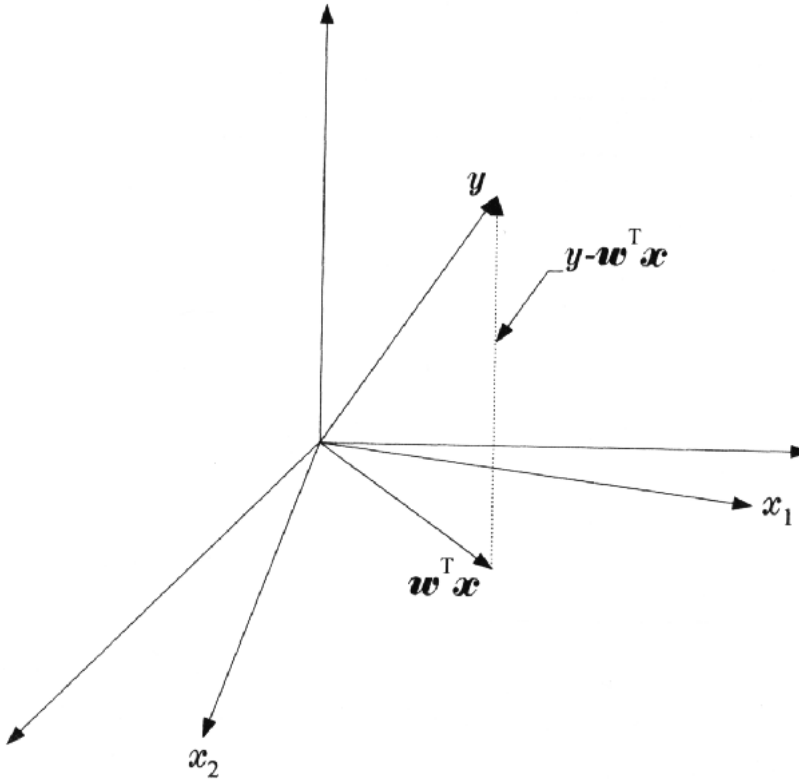
$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2\mathsf{E}[\mathbf{x}(y - \mathbf{x}^T\mathbf{w})] = \mathbf{0}, \tag{15}$$

Edellisestä kaavasta saadaan ratkaisu

$$\hat{\mathbf{w}} = \mathbf{R}_x^{-1}\mathrm{E}[\mathbf{x}y], \tag{16}$$

missä $\mathbf{R}_x$ on $\mathbf{x}$:n korrelaatiomatriisi ja $\mathrm{E}[\mathbf{x}y]$ on $\mathbf{x}$:n ja y:n ristikorrelaatiovektori

Ratkaisun geometrinen tulkinta? Diskriminanttifunktion tavoitearvoa approksimoidaan piirteiden lineaarikombinaatiolla, syntynyt virhe on ortogonaalinen piirreavaruuteen nähden:


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

Yleistys useammalle kuin kahdelle luokalle

Useamman ($M > 2$) kuin kahden luokan tapauksessa muodostetaan jokaiselle luokalle oma diskriminanttifunktio $g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x}$ ja asetetaan sen tavoitearvot y_i seuraavasti:

$$y_i = \begin{cases} 1, & \text{jos } \mathbf{x} \in \omega_i \\ 0, & \text{muulloin} \end{cases} \tag{17}$$

Huom! Kun $M = 2$, $y = \pm 1$ tuottaa saman ratkaisun kuin edellinen valinta, koska $\mathbf{w}^T\mathbf{x} = \frac{1}{2}(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x}$

Kootaan tavoitearvot vektoriksi $\mathbf{y} = [y_1, \ldots, y_M]^T$ ja painovektorit matriisiksi $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$

Ratkaisu löytyy minimoimalla seuraava kustannusfunktio $J(\mathbf{W})$

$$J(\mathbf{W}) = \mathsf{E}[\|\mathbf{y} - \mathbf{W}^T\mathbf{x}\|^2] = \mathsf{E}[\sum_{i=1}^{M}(y_i - \mathbf{w}_i^T\mathbf{x})^2] \tag{18}$$

Edellisestä kaavasta nähdään, että voidaan suunnitella jokainen diskriminanttifunktio erikseen.

LMS- eli Widrow-Hoff algoritmi

Yleensä ei tunneta MSE-estimaatin ratkaisussa eli kaavassa (16) esiintyviä korrelaatiomatriisia $\mathbf{R}_x$ ja ristikorrelaatiovektoria $\mathrm{E}[\mathbf{x}y]$

Voidaan osoittaa, että ratkaisu ongelmaan, joka on muotoa $\mathrm{E}[F(\mathbf{x}_k, \mathbf{w})] = 0$, löytyy seuraavalla iteratiivisellä algoritmillä:

$$\hat{\mathbf{w}}(k) = \hat{\mathbf{w}}(k-1) + \rho_k F(\mathbf{x}_k, \hat{\mathbf{w}}(k-1)) \tag{19}$$

kunhan

$$\begin{aligned} \sum_{k=1}^{\infty} \rho_k &\to \infty \\ \sum_{k=1}^{\infty} \rho_k^2 &< \infty \end{aligned} \tag{20}$$

$\mathbf{x}_k$ on sarja satunnaisia vektoreita, jotka ovat peräisin samanlaisista tnjakaumista, $F(\cdot,\cdot)$ on jokin funktio, ja $\mathbf{w}$ on sen tuntematon parametrivektori

Kun edellistä sovelletaan diskriminanttifunktion painokertoimien hakuun,

$$\hat{\mathbf{w}}(k) = \hat{\mathbf{w}}(k-1) + \rho_k \mathbf{x}_k (y_k - \mathbf{x}_k^T \hat{\mathbf{w}}(k-1)) \tag{21}$$

(kun $k \to \infty$, $\hat{\mathbf{w}}(k)$ lähestyy asymptoottisesti MSE-estimaattia)

Kerroin ρ_k voi olla myös sopivasti rajoitettu vakio $0 < \rho < 2/\text{trace}\{\mathbf{R}_x\}$. Voidaan osoittaa, että mitä pienempi ρ, sitä pienempi on estimaatin $\hat{\mathbf{w}}(k)$ varianssi. Toisaalta, konvergointi on hitaampaa pienellä ρ:lla

Kaavasta (21) nähdään, että estimaattia päivitetään jokaisen havainnon jälkeen. Kun ρ on vakio, pystyy estimaatti mukautumaan paremmin luokkien tnjakaumien muutoksiin

MSE-estimaatti ja luokkien a posteriori tn:t

Lähdetään tästä liikkeelle: voidaan helposti osoittaa, että

$$\begin{aligned}\hat{\mathbf{y}} &= \arg\min_{\tilde{\mathbf{y}}} \mathsf{E}[\|\mathbf{y}-\tilde{\mathbf{y}}\|^2] \\ &= \mathsf{E}[\mathbf{y}|\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{y} p(\mathbf{y}|\mathbf{x}) d\mathbf{y}\end{aligned} \tag{22}$$

(todistus kirjassa, odotusarvot lasketaan tnjakauman $p(\mathbf{y}|\mathbf{x})$ suhteen)

Edellinen tulos tarkoittaa sitä, että MSE-kriteerin mielessä paras mahdollinen estimaatti diskriminanttifunktion tavoitearvolle $\mathbf{y}$ on sen odotusarvo annettuna $\mathbf{x}$.

Diskriminanttifunktion voidaan ajatella olevan $\mathbf{y}$:n regressio annettuna $\mathbf{x}$

Muokataan kaavasta (18) yleisemmälle ongelmalle kustannusfunktio:

$$J = \mathsf{E}[\sum_{i=1}^{M} (g_i(\mathbf{x};\mathbf{w}_i) - y_i)^2], \tag{23}$$

missä diskriminanttifunktiot eivät ole välttämättä lineaarisia $\mathbf{x}$:n tai parametriensä $\mathbf{w}_i$ suhteen

Edellinen kaava voidaan kirjoittaa myös näin:

$$J = \mathsf{E}[\sum_{i=1}^{M}(g_i(\mathbf{x};\mathbf{w}_i) - \mathsf{E}[y_i|\mathbf{x}])^2] + \mathsf{E}[\sum_{i=1}^{M}(\mathsf{E}[y_i^2|\mathbf{x}] - (\mathsf{E}[y_i|\mathbf{x}])^2)] \qquad (24)$$

(välivaiheet kirjassa)

Jälkimmäinen termi ei riipu lainkaan diskriminanttifunktiosta, joten se voidaan jättää huomioimatta, kun minimoidaan J:tä

Nähdään kaavasta (22), että J minimoituu, kun $g_i(\mathbf{x};\hat{\mathbf{w}}_i)$ approksimoi MSE-mielessä mahdollisimman tarkasti $\mathsf{E}[y_i|\mathbf{x}]$:ää

Toisaalta:

$$\mathsf{E}[y_i|\mathbf{x}] = \sum_{j=1}^{M} y_i P(\omega_j|\mathbf{x}) \qquad (25)$$

Kun valitaan tavoitearvot siten, että $y_i = 1$ tai $y_i = 0$ riippuen kuuluuko $\mathbf{x}$ luokkaan ω_i vai ei, $g_i(\mathbf{x}; \hat{\mathbf{w}}_i)$ on $P(\omega_i|\mathbf{x})$:n MSE-estimaatti

$P(\omega_i|\mathbf{x})$ voidaan siis estimoida valitsemalla diskriminanttifunktioiden parametrit MSE-kriteerin ja LMS-menetelmän avulla (tuntematta tnjakaumia!) ja sitä voidaan käyttää Bayesiläisessä luokittelussa

Se kuinka hyvin luokkien *a posteriori* tnjakaumien estimointi sitten onnistuu riippuu diskriminanttifunktioiden tyypistä

1.4 Fisherin diskriminantti

Eräs tapa muodostaa lineaarinen luokitin: etsitään suora, jolle projisoidut piirrevektorit separoituvat mahdollisimman hyvin, ja jaetaan suora sitten päätösalueiksi

Tarkastellaan kahden luokan tapausta

Määritellään projektio seuraavasti: $y = \mathbf{w}^T\mathbf{x}$, missä $\|\mathbf{w}\| = 1$

Sopiva mitta $J(\mathbf{w})$ luokkien separoituvuudelle?

Eräs järkevä mitta saadaan luokkien projektioden odotusarvojen $\mu_{Y_i} = \mathsf{E}[\mathbf{w}^T\mathbf{x}|\mathbf{x} \in \omega_i]$ ja varianssien $\sigma^2_{Y_i} = \mathsf{E}[\|\mathbf{w}^T\mathbf{x} - \mu_{Y_i}\|^2|\mathbf{x} \in \omega_i]$ avulla:

$$J(\mathbf{w}) = \frac{(\mu_{Y_1} - \mu_{Y_2})^2}{\sigma^2_{Y_1} + \sigma^2_{Y_2}} \tag{26}$$

tai otoskeskiarvojen $m_{Y_i} = 1/N_i \sum_{j=1}^{N_i} \mathbf{w}^T \mathbf{x}_j$ ja sironnan $s_{Y_i}^2 = \sum_{j=1}^{N_i} (\mathbf{w}^T \mathbf{x}_j - m_{Y_i})^2$ avulla:

$$J(\mathbf{w}) = \frac{(m_{Y_1} - m_{Y_2})^2}{s_{Y_1}^2 + s_{Y_2}^2} \tag{27}$$

(N_i on luokkaan ω_i kuuluvien opetusnäytteiden lkm)

Suoraa $y = \hat{\mathbf{w}}^T \mathbf{x}$, joka löytyy maksimoimalla $J(\mathbf{w})$, kutsutaan *Fisherin diskriminantiksi*

Tapaus 1: tunnetaan luokkien odotusarvot ja kovarianssimatriisit

Kun tunnetaan luokan ω_i odotusarvo μ_i ja kovarianssimatriisi $\mathbf{\Sigma}_i$, voidaan vastaavat tunnusluvut laskea projektioille: $\mu_{Y_i} = \mathbf{w}^T \mu_i$ ja $\sigma_{Y_i} = \mathbf{w}^T \mathbf{\Sigma}_i \mathbf{w}$

Derivoidaan $J(\mathbf{w})$ (26) $\mathbf{w}$:n suhteen ja asetetaan nollavektoriksi. Ratkaisuksi saadaan:

$$\hat{\mathbf{w}} = \frac{1}{2} k (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1} (\mu_1 - \mu_2), \tag{28}$$

missä

$$k = \frac{\sigma_{Y_1}^2 + \sigma_{Y_2}^2}{\mu_{Y_1} - \mu_{Y_2}} \tag{29}$$

normalisoi $\hat{\mathbf{w}}$:n pituuden 1:ksi

Tapaus 2: ei tunneta luokkien odotusarvoja ja kovarianssimatriiseja

Määritellään aluksi seuraavat matriisit:

$$\begin{aligned} \mathbf{S}_i &= \sum_{j=1}^{N_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \\ \mathbf{S}_W &= \mathbf{S}_1 + \mathbf{S}_2 \\ \mathbf{S}_B &= (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \end{aligned} \tag{30}$$

missä $\mathbf{m}_i = 1/N_i \sum_{j=1}^{N_i} \mathbf{x}_j$

Näiden avulla $J(\mathbf{w})$ (27) voidaan kirjoittaa seuraavasti:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{31}$$

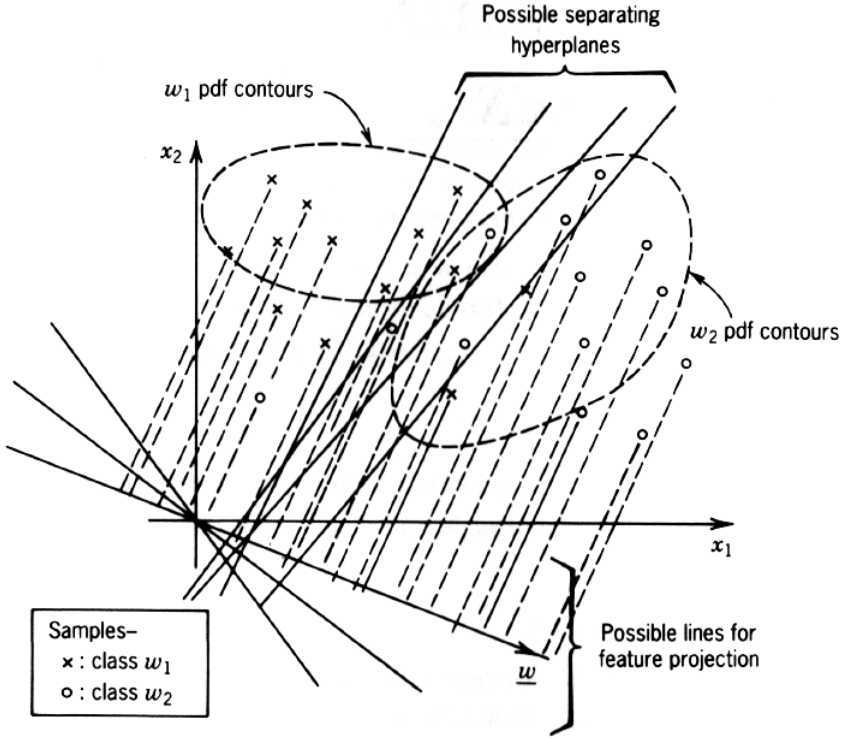
Derivoidaan $J(\mathbf{w})$ ja asetetaan nollavektoriksi. Tällöin saadaan seuraava yleinen ominaisvektoriongelma:

$$\lambda \mathbf{S}_W \hat{\mathbf{w}} = \mathbf{S}_B \hat{\mathbf{w}}, \tag{32}$$

ja ratkaisu:

$$\hat{\mathbf{w}} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{33}$$

(todistus kirjassa)


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