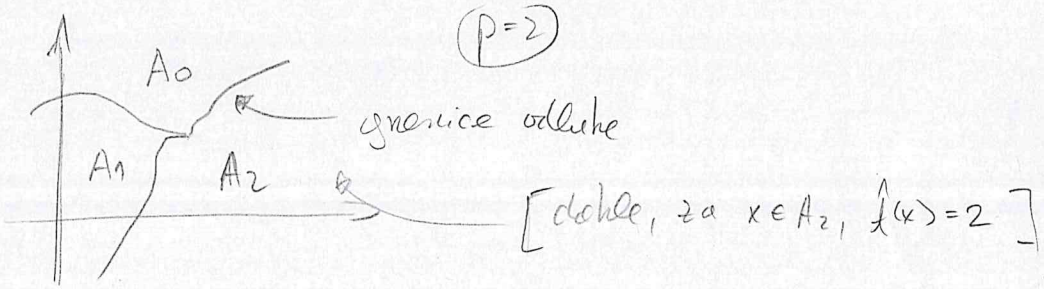
(7) Osnovne metode za klasifikaciju

(7.1) Bayesov klasifikator

↳ $Y \in S = \{0, 1, \dots, K-1\}$ za $K \geq 2$, $X \in \mathbb{R}^p$

→ trebamo $\hat{f} = \hat{f}(\cdot) : \mathbb{R}^p \to S$

Svaka $\hat{f}$ dijeli $\mathbb{R}^p$ na $A_k := \hat{f}^{-1}(\{k\})$, $k \in S$



Svaka f-ja gubitka $L : S \times S \to \mathbb{R}$ je zapravo $K \times K$ matrica

[$L(a,b)$ = trošak ako stvarnu klasu $\underline{\underline{a}}$ predviđamo kao $\underline{\underline{b}}$]

↳ $L(f) := \mathbb{E}[L(Y, f(X))]$, te $\forall x \in \mathbb{R}^p$

$$L_x(f) := \mathbb{E}\left[L(Y, f(x)) \mid X = x\right]$$

$$= \sum_{k \in S} L(k, f(x)) \underbrace{\mathbb{P}(Y = k \mid X = x)}_{=: p_k(x)} \qquad (7.1)$$

↳ Bayesov klasifikator je $f^* : \mathbb{R}^p \to S$ dan s

$$f^*(x) = \underset{c \in S}{\operatorname{argmin}} \sum_{k \in S} L(k, c)\, p_k(x), \qquad (7.2)$$

te očito vrijedi $\forall f : \mathbb{R}^p \to S$, $\forall x \in \mathbb{R}^p$,

$$L_x(f^*) \leq L_x(f), \quad L(f^*) \leq L(f).$$

"Bayesov nzih"

[Naravno, f^* ne znamo jer ne znamo $p_k(x)$]

(P2) Ako je $L(a,b) = \begin{cases} 1 & , a \neq b \\ 0 & , a = b \end{cases}$ ("0-1 gubitak"),

onda

$$f^*(x) = \underset{h \in \mathcal{H}}{\operatorname{argmax}}\, p_h(x), \quad x \in \mathbb{R}^p \qquad (7.3)$$

[Naravno, f^* iz (7.2) nezavisno, jer nezavisno $p_h(x)$!]

U većini metoda prvo procjenjuje $p_h(x) = P(Y=h \mid X=x)$, $x \in \mathbb{R}^p$ te se $\hat{p}_h(x)$ te umjesto f^*:

$$\hat{f}(x) := \underset{c \in \mathcal{C}}{\operatorname{argmin}} \sum_{h \in \mathcal{H}} L(h,c)\, \hat{p}_h(x), \quad x \in \mathbb{R}^p \qquad (7.4)$$

$$\left(\text{0-1 gub.} \Rightarrow \hat{f}(x) = \underset{c}{\operatorname{argmax}}\, \hat{p}_c(x)\right) \qquad (7.5)$$

(7.2) Logistička regresija

↳ za $K=2$, tj. $S=\{0,1\}$, to je GLM za $\{B(1,p)\}$ familiju uz kanonsku f-ju veze g (tzv. logit f-ju): za $\beta \in \mathbb{R}^p$ pretpost.

$$\log\left(\frac{p_1(x)}{\underbrace{1-p_1(x)}_{=p_0(x)}}\right) = g(p_1(x)) \equiv x^T\beta, \quad x \in \mathbb{R}^p \qquad (7.6)$$

tj.

$$p_1(x) = g^{-1}(x^T\beta) \equiv \frac{e^{x^T\beta}}{1+e^{x^T\beta}}, \quad x \in \mathbb{R}^p \qquad (7.7)$$

↳ $\frac{p_1(x)}{p_0(x)}$ su "izgledi" (engl. odds), npr.

- $p_1 = \frac{2}{3} \Rightarrow p_0 = \frac{1}{3}$; odds $= \frac{2}{1}$
- $p_1 = \frac{2}{5} \Rightarrow p_0 = \frac{3}{5}$; odds $= \frac{2}{3}$. ⊥

Interpretacija parametara β?

↳ uz (7.6) (tj. (7.7)), za $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ te

$x' = (x_1, \dots, x_{j-1}, x_j+1, x_{j+1}, \dots, x_p) \in \mathbb{R}^p$, imamo

$$\log\left(\frac{p_1(x')}{p_0(x')}\right) = \log\left(\frac{p_1(x)}{p_0(x)}\right) + \beta_j, \text{ tj.}$$

$$\frac{p_1(x')}{p_0(x')} = \frac{p_1(x)}{p_0(x)} \cdot e^{\beta_j} \quad \begin{cases} > 1, & \text{ako } \beta_j > 0 \\ < 1, & \text{ako } \beta_j < 0 \end{cases} . \perp$$

→ ~~pweraje. R~~

(7.2.1) Multinomijalna logistička regresija ($K \in \mathbb{N}$)

↳ pretp. da za neke $\beta^{(1)}, \dots, \beta^{(K-1)} \in \mathbb{R}^p$,

$$\log\left(\frac{p_h(x)}{p_0(x)}\right) = x^T\beta^{(h)}, \quad x \in \mathbb{R}^p, \quad h = 1, \dots, K-1 \qquad (7.8)$$

↳ [$K-1$ nezavisnih "log. regr." za (R) u odnosu na (0).]

Napr. Kategorija 0 je tzv. bazna kategorija (engl. baseline).
↳ proizvoljna!

(7.8) $\Rightarrow$ za x, x' koji gase $(x' = x + e_j)$,
$\forall j = 1, \dots, p$

$$\frac{p_k(x')}{p_0(x')} = \frac{p_k(x)}{p_0(x)} \cdot e^{\beta_j^{(k)}}.$$

(Pr) Pokažite da je (7.8) ekvivalentno:

$$p_k(x) = \frac{e^{x^T \beta^{(k)}}}{1 + \sum_{i=1}^{K-1} e^{x^T \beta^{(i)}}}, \quad k = 1, \dots, K-1, \qquad (7.9)$$

$$p_0(x) = \frac{1}{1 + \sum_{i=1}^{K-1} e^{x^T \beta^{(i)}}}$$

↳ ako su procjene $\hat\beta^{(1)}, \dots, \hat\beta^{(K-1)}$ te $\hat\beta^{(0)} := (0, \dots, 0) \in \mathbb{R}^p$,

$$\hat p_k(x) := \frac{e^{x^T \hat\beta^{(k)}}}{\sum_{i=0}^{K-1} e^{x^T \hat\beta^{(i)}}}, \quad k = 0, 1, \dots, K-1. \quad \left(= \text{softmax}_k\left(x^T\hat\beta^{(0)}, \dots, x^T\hat\beta^{(K-1)}\right)\right)$$

[tj. 0-1 gukitak]

↳ ako je $\hat f(x) = \underset{k \in S}{\arg\max}\ \hat p_k(x)$, imamo

$$\hat f(x) = \underset{k \in S}{\arg\max}\ \underbrace{x^T \hat\beta^{(k)}}_{=: \hat\delta_k(x)}, \quad x \in \mathbb{R}^p$$

"diskriminacijske f-je"

↳ Granica odluke između klasa $k, \ell \in S$ je

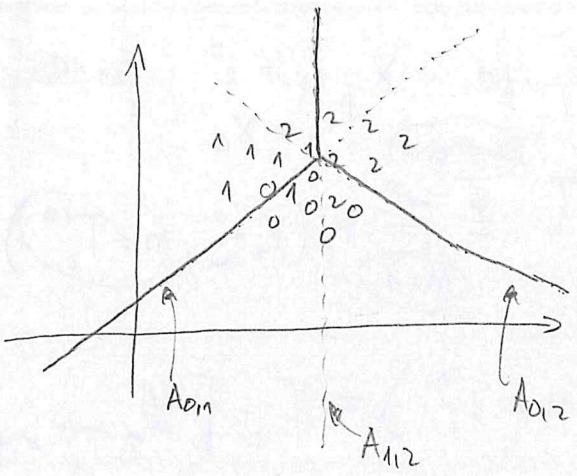
$$A_{k,\ell} = \{x \in \mathbb{R}^p : \hat\delta_k(x) = \hat\delta_\ell(x)\} = \{x : x^T(\hat\beta^{(k)} - \hat\beta^{(\ell)}) = 0\}$$

↳ hiperravnina u $\mathbb{R}^p$, $\forall k, \ell$

↳ linearne granice odluke!

[tipično imamo i $x_0 = 1$ pa $A_{k,\ell}$ ne prolazi nužno kroz 0.]

npr.:



$K=3,\ p=2$

(1°) A_{01} i $A_{01,2}$ → nacrtaj $\{x : f(x)=0\}$

(2°) $A_{11,2}$ → $\{f(x)=1\}$ i $\{f(x)=2\}$

[Mogući li se A_{ij} sjeći u jednoj točci? DA]

[Nap.! Multinomijalna log-regresija je zapravo specijalan slučaj tzv. "feedforward" neuronske mreže koja nema "skrivenih" slojeva. ∎]

7.3 Generativni modeli

- u log. regresiji direktno modeliramo $p_k(x)$, $k \in S$
  → [primjer] "diskriminacijska" metoda
- alternativno, ako modeliramo $f_k(x) := \mathbb{P}(X = x \mid Y = k)$, $x \in \mathbb{R}^p$, $k \in S$, te $\pi_k := \mathbb{P}(Y = k)$,

$$p_k(x) \overset{\text{Bayes}}{=} \frac{f_k(x)\,\pi_k}{\sum_{i \in S} f_i(x)\,\pi_i} \qquad (7.10)$$

↳ zapravo modeliramo $\mathbb{P}(X = x, Y = k) = f_k(x)\,\pi_k$, tj.
$\mathbb{P}((X,Y) \in \cdot)$ → "generativne" metode

↳ (1°) Linearna diskriminacijska analiza (LDA)
(2°) Kvadratna —— " —— (QDA)
(3°) Naivni Bayes

Uočimo, (7.10) povlači (uz 0-1 gubitak)

$$f^*(x) = \underset{k \in S}{\operatorname{argmax}}\, f_k(x)\,\pi_k. \qquad (7.11)$$

7.3.1 LDA

↳ Pretpostavka je da $X \mid Y = k \sim N(\mu_k, \Sigma_k)$, $\mu_k \in \mathbb{R}^p$, $\Sigma_k \in \mathbb{R}^{p \times p}$, tj.

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma_k)}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\}, \quad x \in \mathbb{R}^p, \qquad (7.12)$$

te dodatno

$$\Sigma_k = \Sigma, \ \forall k \in S\,! \qquad (7.13)$$

$$\hat{f}^*(x) = \underset{k\in S}{\operatorname{argmax}} \underbrace{\log\left(f_k(x)\cdot\pi_k\right)}_{=\log(f_k(x))+\log(\pi_k)}$$

(Mahalanobisova udaljenost od x i μ_k)

(7.14)

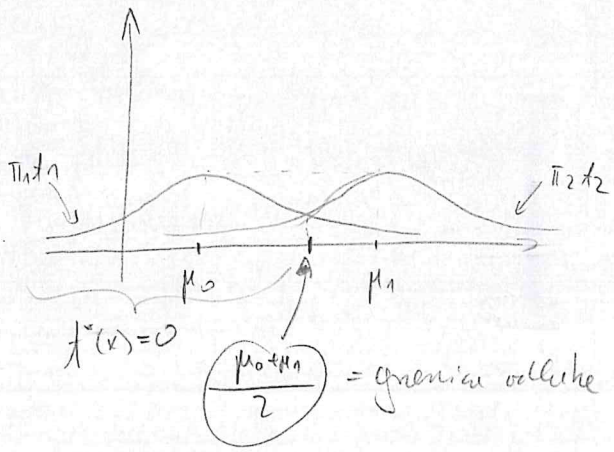
$$\overset{(7.12)}{\underset{(7.13)}{=}} \underset{k\in S}{\operatorname{argmin}} \Big\{ \underbrace{\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}_{=:d_k(x)} - \log(\pi_k) \Big\}, \quad x\in\mathbb{R}^p$$

Ako je $\boxed{\Sigma = I_p}$ te $\boxed{\pi_0 = \dots = \pi_{K-1}}$,

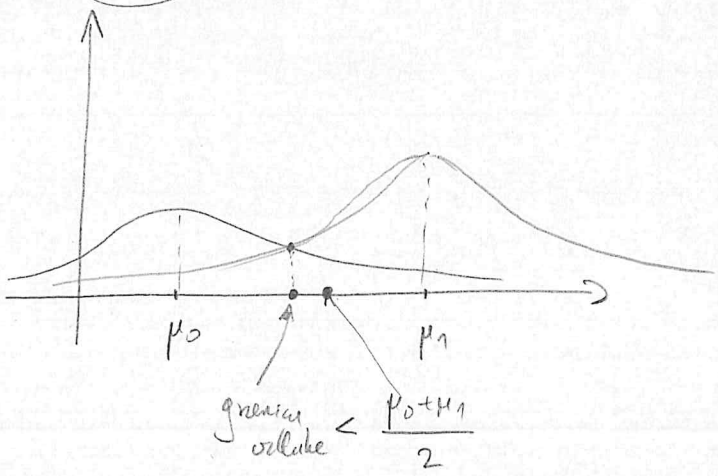
$$\hat{f}^*(x) = \underset{k\in S}{\operatorname{argmin}} \|x-\mu_k\|_2^2 .$$

Pr. $p=1$, $K=2$ $(\Sigma = \sigma^2 > 0)$ $\longrightarrow$ $\boxed{f_k = f_N(\mu_k, \sigma^2)}$

1° $\pi_0 = \pi_1$



2° $\pi_1 > \pi_2$



[uzimamo u obzir da je unaprijed (x_i, y_i) s $\boxed{y_i = 1}$] ∎

Iz $d_k(x) = \underbrace{\frac{1}{2}x^T\Sigma^{-1}x}_{\text{ne ovisi o } k} - x^T\underbrace{\Sigma^{-1}\mu_k}_{=:w_k} + \underbrace{\mu_k^T\Sigma^{-1}\mu_k - \log(\pi_k)}_{=:-b_k}$

(7.14)

$$\hat{f}^*(x) = \underset{k\in S}{\operatorname{argmax}} \underbrace{x^T w_k + b_k}_{k}$$

$\rightarrow$ linearne diskriminacijske funkcije $\longrightarrow$ "LDA"

Specijalno, LDA $\longrightarrow$ linearne granice odluke.

Parametre projekcijskog:

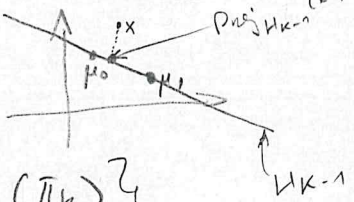
- $\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i = k\}}$,
- $\hat{\mu}_k = \frac{\sum_{i=1}^{n} x^{(i)} \mathbb{1}_{\{y_i = k\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{y_i = k\}}}$,
- $\hat{\Sigma} = \frac{1}{n-K} \sum_{i=1}^{n} (x^{(i)} - \hat{\mu}_{y_i})(x^{(i)} - \hat{\mu}_{y_i})^T$. ("pooled" uzoračka kov. matrica)

$\hookrightarrow \hat{f}(x) = \underset{k \in S}{\operatorname{argmin}} \hat{d}_k(x)$.

dim. $K-1$!

Nap. 1 (!) (i) Pretp. da je $\hat{\Sigma} = I_p$, te $H_{K-1} := \text{aff}(\mu_0, \dots, \mu_{K-1})$

$\hookrightarrow \hat{f}(x) = \underset{k \in S}{\operatorname{argmin}} \{\frac{1}{2} \| x - \mu_k \|_2^2 - \log(\pi_k)\}$



$= \underset{k \in S}{\operatorname{argmin}} \{\frac{1}{2} \| \text{Proj}_{H_{K-1}}(x) - \mu_k \|_2^2 - \log(\pi_k)\}$

$\Rightarrow$ dim je $K-1 < p$, imamo smanjenje dimenzije

(ii)* općenito, prvo napravimo transformaciju

$x \longmapsto x^* := \hat{\Sigma}^{-\frac{1}{2}} x$ $\left(\Rightarrow (x-\hat{\mu}_k)^T \hat{\Sigma}^{-1} (x-\hat{\mu}_k) = \| x^* - \hat{\mu}_k^* \|_2^2, \ \hat{\Sigma}^* = I_p\right)$

te gledamo projekcije na $H_{K-1}^* = \text{aff}(\mu_0^*, \dots, \mu_{K-1}^*)$.

$\hookrightarrow$ Može se pokazati da postoje vektori $v_1, \dots, v_{K-1} \in \mathbb{R}^p$

t.d.

$$\hat{f}(x) = \underset{k \in S}{\operatorname{argmin}} \{\frac{1}{2} \| V \cdot x - V \cdot \hat{\mu}_k \|_2^2 - \log(\pi_k)\}$$

za $V = \begin{bmatrix} - v_1^T - \\ \vdots \\ - v_{K-1}^T - \end{bmatrix} \in \mathbb{R}^{(K-1) \times p}$

$\hookrightarrow$ $v_i^T \cdot x$ je tzv. i-ta diskriminacijska koordinata

$\hookrightarrow$ Primjerena statistika i ESL 4.3.3

$\rightarrow$ wine-LDA.R

7.3.2 QDA

↳ sve kao kod LDA, ali bez pretp. $\Sigma_k = \Sigma$, $\forall k \in S$.

$$\Rightarrow f^*(x) \overset{pt}{=} \underset{k \in S}{\operatorname{argmax}} \left\{ -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log(\det(\Sigma_k)) + \log \pi_k \right\}$$

$$= \underset{k \in S}{\operatorname{argmax}} \left\{ -\frac{1}{2} x^T \Sigma_k^{-1} x + \dots \right\}$$

↳ kvadratne granice odluke.

Ipak, u odnosu na LDA, QDA procjenjuje $\hat{\Sigma}_k$, $k \in S$

→ dodatnih $\frac{p(p-1)}{2} \cdot (K-1)$ parametara!

[Kada je p velik, čak i LDA procjenjuje previše parametara

→ Σ, μ_k, π_k

$\frac{p(p-1)}{2}$ $K \cdot p$ $K-1$]

7.3.3 Naivni Bayes (NB)

↳ pretp. da su $\forall k \in S$, $X_1, \dots, X_p$ nezavisne uz dano $Y=k$, tj.

$$f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j), \quad \forall x = (x_1, \dots, x_p) \in \mathbb{R}^p \qquad (7.15)$$

[nezavisnost je oblik regularizacije]

Marginalne gustoće možemo procijeniti npr.

- ako $X_j \in \mathbb{R}$ => pretpostavimo

$$X_j \mid Y=k \sim N(\mu_{kj}, \sigma^2_{kj})$$

(kao QDA uz Σ_k = dijagonalna matrica)

- ako $X_j \in S_j = \{0, \dots, K_j - 1\}$ $\Rightarrow$ npr. neparametarski

$$\hat{f}_{k,j}(x) = \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = k, x_{ij} \in x\}_j}{\sum_{i=1}^{n} \mathbb{1}\{y_i = k\}}$$

7.3.4 Usporedba metoda [sami]

- log. regresija — $\log\left(\frac{p_k(x)}{p_0(x)}\right) = \boxed{x^T \beta^{(k)}}$ (tipično $x_0 = 1$)

- LDA — $\log\left(\frac{p_k(x)}{p_0(x)}\right) \overset{(BZ)}{=} \log\frac{\pi_k}{\pi_0} - \frac{1}{2}(\mu_k + \mu_0)^T \Sigma^{-1}(\mu_k - \mu_0) + x^T \Sigma^{-1}(\mu_k - \mu_0)$

$$=: \alpha_{k,0} + \sum_{i=1}^{K-1} \alpha_{k,i} \cdot x_i$$

$$=: \boxed{x^T \alpha^{(k)}}$$

($x_0 := 1$)

↳ ipak, $\hat{\alpha}^{(k)} \neq \hat{\beta}^{(k)}$ [, te je log. regresija općenitija jer ne pretpostavlja ništa na rezd. $\mathbb{P}(X \in \cdot)$].

- NB — $\log\left(\frac{p_k(x)}{p_0(x)}\right) \overset{(17.15)}{=} \log\left(\frac{\pi_k}{\pi_0}\right) + \sum_{j=1}^{p} \log\left(\frac{f_{k,j}(x_j)}{f_{0,j}(x_j)}\right)$ ($= \pi_k f_k(x)$)

$$=: \alpha_k + \sum_{j=1}^{p} g_{k,j}(x_j)$$

↳ fleksibilne granice odluke

↳ ipak, nema interakcija

- QDA — $\log\left(\frac{p_k(x)}{p_0(x)}\right) = a_k' + \sum_{j=1}^{p} b_{k,j} x_j + \boxed{\sum_{j,\ell=1}^{p} c_{k,j,\ell} x_j x_\ell}$

interakcije.

7.4 kNN metoda

↳ za $h_0 \in \{1, \dots, n\}$,

$$\hat{p}_h(x) = \frac{1}{h_0} \sum_{x^{(i)} \in \mathcal{U}_{h_0}(x)} \mathbb{1}\{y_i = h\}, \quad h \in S, \ x \in \mathbb{R}^p$$

↳ $$\hat{f}(x) = \underset{h \in S}{\operatorname{argmax}}\ \hat{p}_h(x) = \underset{h \in S}{\operatorname{argmax}} \sum_{x^{(i)} \in \mathcal{U}_{h_0}(x)} \mathbb{1}\{y_i = h\}$$

"majority vote"

[ez 0-1 veszteség]

Ucsp.] za velike p

→ problem dimenzionalnosti! ☹

→ S-and-P.R