Klasická diskriminační analýza

data: $\begin{pmatrix} x_{11} & \dots & x_{1p} & y_1 \\ \vdots & & \vdots & \vdots \\ x_{n1} & \dots & x_{np} & y_n \end{pmatrix}$

$x_i = (x_{i1}, \dots, x_{ip})^T$ .... vektor regresorů (prediktorů) pro i-té pozorování

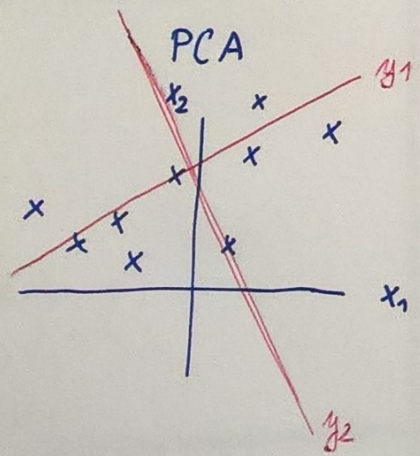
y_i - udává příslušnost i-tého pozorování k dané skupině
- kategoriální proměnná o kategoriemi $1, 2, \dots, J$

Cíl: Vybrat rozhodovací pravidlo, které bude co nejlépe klasifikovat nové pozorování $x^* = (x_1^*, \dots, x_p^*)^T$ do správné skupiny.

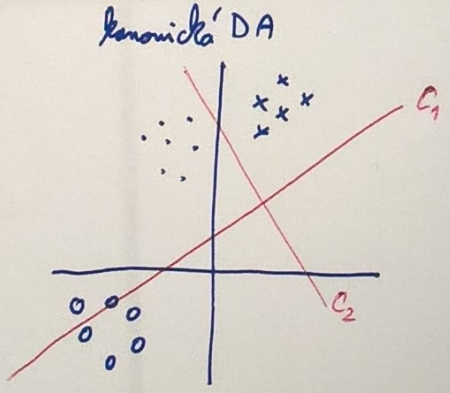
Budeme předpokládat, že naše data jsou generována sdruženým náhodným vektorem $X = (X_1, \dots, X_p)^T$.

1.) Kanonická diskriminační analýza

- kombinace PCA a MANOVA



$Y_1 = a_1^T X$
$Y_2 = a_2^T X$
⋮



$C_1 = a_1^T X$
$C_2 = a_2^T X$
⋮

maximalizuj se podíl meziskupinové a vnitroskupinové variability

$$\lambda = \frac{a_1^T B a_1}{a_1^T \Sigma a_1}$$

$$B = \frac{1}{J} \sum_{i=1}^{J} (\mu_i - \mu)(\mu_i - \mu)^T$$

Σ je společný rozptyl skupin

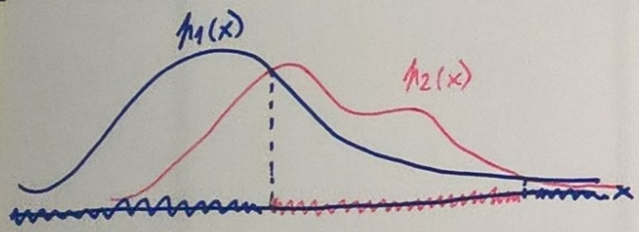
2) Fisherova diskriminační analýza

- nechť náh. vektor X má v j-té skupině sdruženou hustotu $f_j(x)$, j.

$X | Y=j$ má hustotu $f_j(x)$ pro $j=1,\dots,J$

↑ známá p-rozměrná hustota

Př.: $p=1$, 2 skupiny ($J=2$)



x^* klasifikuj do „pravděpodobnější" skupiny

a co když skupiny nejsou rozloženy rovnoměrně?

označme $\pi_j = P(Y_i = j) = P(\text{dané pozorování patří do třídy } j)$ pro $j=1,\dots,J$ ... apriorní pravděpodobnost

zřejmě $\pi_1 + \pi_2 + \dots + \pi_J = 1$.

Bayesova věta: $$P(Y=j \mid X=x) = \frac{P(X=x \mid Y=j)\cdot P(Y=j)}{P(X=x \mid Y=1)\cdot P(Y=1) + \dots + P(X=x \mid Y=J)\cdot P(Y=J)} = \frac{f_j(x)\cdot \pi_j}{f_1(x)\cdot\pi_1 + \dots + f_J(x)\cdot\pi_J}$$ $(j=1,\dots,J)$

(konstanta (nezávisí na j))

aposteriorní pravděpodobnost

Bayesovské rozhodovací pravidlo – pozorování x zařaď do třídy $\underset{j=1,\dots,J}{\operatorname{argmax}}\, P(Y=j \mid X=x) = \underset{j=1,\dots,\pi}{\operatorname{argmax}}\, f_j(x)\cdot\pi_j = \underset{j=1,\dots,J}{\operatorname{argmax}} \{\log f_j(x) + \log \pi_j\}$

– minimalizuje pravděpodobnost špatné klasifikace

a) Lineární diskriminační analýza (LDA)

$$X|Y=j \sim N_p(\mu_j, \Sigma)$$

$$f_j(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma^{-1} (x-\mu_j)\right\}, \; x \in \mathbb{R}^p$$

Bayesovské rozhodovací pravidlo:

$$\log \pi_j + \log f_j(x) = \log \pi_j \underbrace{- \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma|}_{\text{nezávisí na } j} - \frac{1}{2} x^T \Sigma^{-1} x + \frac{1}{2} x^T \Sigma^{-1} \mu_j + \frac{1}{2} \mu_j^T \Sigma^{-1} x - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j = \underbrace{\log \pi_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \mu_j^T \Sigma^{-1} x}_{L_j(x) \dots \text{lineární diskriminační funkce}} + K$$

neznámé parametry π_j, μ_j a Σ je třeba odhadnout z dat

- $\hat{\pi}_j = \frac{n_j}{n}$ ... pokud věříme zastoupení skupin v datech
  $= \frac{1}{J}$ -- pokud jsou všechny skupiny zastoupeny rovnoměrně
- $\hat{\mu}_j$ odhadneme pomocí výběrového průměru v j-té skupině
- S_j ... výběrová kovariační matice v j-té skupině

$$\hat{\Sigma} = \frac{1}{n-J}\left((n_1-1)S_1 + \dots + (n_J-1)S_J\right)$$

b) Kvadratická diskriminační analýza (QDA)

$$X|Y=j \sim N_p(\mu_j, \Sigma_j)$$

$$f_j(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_j|}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)\right\}, \quad x \in \mathbb{R}^p$$

Bayesovské rozhodovací pravidlo:

$$\log \pi_j + \log f_j(x) = \log \pi_j - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\mu_j^T \Sigma_j^{-1}\mu_j + \mu_j^T \Sigma_j^{-1} x - \frac{1}{2} x^T \Sigma_j^{-1} x$$

nezávisí na j

$Q_j(x)$ ... kvadratická diskriminační funkce