5. Classification & Clustering

5.1 Discriminant Analysis

$C_1, \dots, C_g$ populations, groups, classes

$R_1, \dots, R_g$, $\bigcup_{i=1}^{g} R_i = \mathbb{R}^p$ discriminant rule by

$x \in C_i$ if $x \in R_i$

5.1.1. Fisher's linear discriminant rule

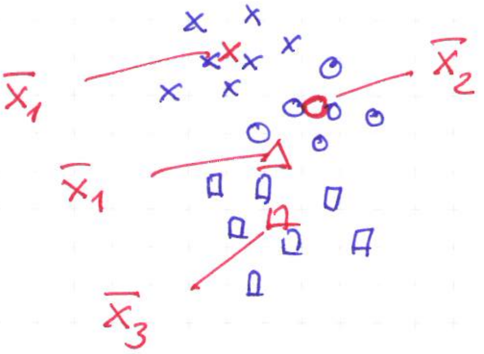
Given training set $x_1, \dots, x_n$ with known class affiliation.

x some observation (unknown class affiliation)

Objective: Find a linear discriminant rule

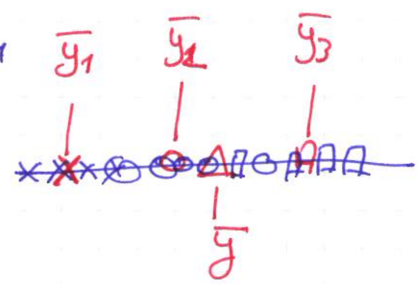
$a^T x$ such that x is allocated to some class optimally.

Training set



linear transformation

$$a^T x_i = y_i$$



Determine a s.t.

$$\frac{\text{SSQ between groups}}{\text{SSQ within groups}} \to \max_a$$

$$\frac{\sum_{\ell=1}^{g} n_\ell (\bar{y}_\ell - \bar{y})^2}{\sum_{\ell=1}^{g} \sum_{j \in C_\ell} (y_j - \bar{y}_\ell)^2} \to \max_a$$

RWTH AACHEN UNIVERSITY

Notation:

Training set: $X = [x_1, \dots, x_n]^T$ samples from g groups $C_1, \dots, C_g$.

$$X_\ell = [x_j]_{j \in C_\ell}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^p \quad \text{(total average)}$$

$$\bar{x}_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} x_i \in \mathbb{R}^p \quad \text{(average over class } C_\ell\text{)}, \; n_\ell = \# C_\ell$$

$a \in \mathbb{R}^p$ linear transformation

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = Xa \quad \in \mathbb{R}^n$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \bar{y}_\ell = \frac{1}{n_\ell} \sum_{j \in C_\ell} y_j$$

Note that:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{\ell=1}^{g} \sum_{j \in C_\ell} (y_j - \bar{y}_\ell + \bar{y}_\ell - \bar{y})^2$$

$$= \sum_{\ell=1}^{g} \left[\sum_{j \in C_\ell} (y_j - \bar{y}_\ell)^2 + \sum_{j \in C_\ell} (\bar{y}_\ell - \bar{y})^2 \right]$$

$$= \underbrace{\sum_{\ell=1}^{g} \sum_{j \in C_\ell} (y_j - \bar{y}_\ell)^2}_{\text{(1)}} + \underbrace{\sum_{\ell=1}^{g} n_\ell (\bar{y}_\ell - \bar{y})^2}_{\text{(2)}}$$

(1) sum of within-groups SSQ (2) between groups SSQ

Thinking the Future
Zukunft denken

In matrix notation:

$E_n = I_n - \frac{1}{n} 1_{n\times n}, \quad E_{n_\ell} = E_\ell, \ \ell = 1, \dots, g$

centering operators.

$$① = \sum_{\ell=1}^{g} y_\ell^T E_\ell y_\ell = \sum_{\ell=1}^{g} a^T X_\ell^T E_\ell X_\ell a$$

$$= a^T \underbrace{\left[\sum_{\ell=1}^{g} X_\ell^T E_\ell X_\ell\right]}_{=W} a = a^T W a \quad \text{(within)}$$

$y_\ell = (y_j)_{j \in C_\ell}$

$$② = \sum_{\ell=1}^{g} n_\ell \left(a^T(\bar{x}_\ell - \bar{x})\right)^2 \overset{(*)}{=} \sum_{\ell=1}^{g} n_\ell a^T (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})^T a$$

$$= a^T \underbrace{\left[\sum_{\ell=1}^{g} n_\ell (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})^T\right]}_{=B} a$$

$$= a^T B a \quad \text{(between)}$$

$(*)\ (a^T b)^2 = a^T b (a^T b)^T = a^T b b^T a, \quad a, b \in \mathbb{R}^p$

Linear discriminant analysis:

$$\max_{a \in \mathbb{R}^p} \frac{a^T B a}{a^T W a} \qquad (\Delta)$$

T. 5.1.1 The maximum of (4) is attained at the eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue.

Proof. $$\max_{a \in \mathbb{R}^p} \frac{a^T B a}{a^T W a} = \max_{b \in \mathbb{R}^p} \frac{b^T W^{-1/2} B W^{-1/2} b}{b^T b} \underset{\text{Th. 2.4}}{=} \lambda_{max}\left(W^{-1/2} B W^{-1/2}\right)$$

. . . .

Application: Given a training set $x_1, \dots, x_n \in \mathbb{R}^p$ with known group allocation.

Compute the optimum a by solving (Δ).

New observation x. Compute $a^T x$.

Allocate x to the group with the closest value $a^T \bar{x}_\ell = \bar{y}_\ell$

- Discrimination rule

Allocate x to group ℓ, if

$$|a^T x - a^T \bar{x}_\ell| \leq |a^T x - a^T \bar{x}_j| \quad \forall j = 1, \dots, g.$$

(python code: sebastianraschka.com/Articles/2014-python-lda.html)

- The case of $g=2$ (2 groups of size n_1, n_2, $n_1 + n_2 = n$)

$$B = n_1 (\bar{x}_1 - \bar{x})(\bar{x}_1 - \bar{x})^T + n_2 (\bar{x}_2 - \bar{x})(\bar{x}_2 - \bar{x})^T$$

$$n_1 \bar{x}_1 + n_2 \bar{x}_2 = (n_1 + n_2)\bar{x}$$

$$= n_1 \left(\bar{x}_1 - \frac{n_1}{n}\bar{x}_1 - \frac{n_2}{n}\bar{x}_2\right)(\dots)^T + n_2 \left(\bar{x}_2 - \frac{n_1}{n}\bar{x}_1 - \frac{n_2}{n}\bar{x}_2\right)(\dots)^T$$

$$= n_1 \left(\frac{n_2}{n}\bar{x}_1 - \frac{n_2}{n}\bar{x}_2\right)(\dots)^T + n_2 \left(\frac{n_1}{n}\bar{x}_2 - \frac{n_1}{n}\bar{x}_1\right)(\dots)^T$$

$$= \frac{n_1 n_2^2}{n^2}(\bar{x}_1 - \bar{x}_2)(\dots)^T + \frac{n_2 n_1^2}{n^2}(\bar{x}_1 - \bar{x}_2)(\dots)^T$$

$$= \frac{n_1 n_2}{n}(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^T$$

$$= \frac{n_1 n_2}{n} d d^T, \quad d = \bar{x}_1 - \bar{x}_2$$

B has rank 1, only 1 eigenvalue is $\neq 0$.

Positive eigenvalue of $W^{-1}B$ is

$$\operatorname{tr} W^{-1}B = \frac{n_1 n_2}{n} d^T W^{-1} d$$

with eigenvector $a = W^{-1}d$

Since: $W^{-1}B \underbrace{W^{-1}d}$

$$= \frac{n_1 n_2}{n} W^{-1} d d^T W^{-1} d$$

$$= \underbrace{\frac{n_1 n_2}{n} d^T W^{-1} d}_{\text{eigenvalue}} \underbrace{W^{-1} d}_{\text{eigenvector}}$$

Discriminant rule becomes

allocate x to C_1 if $d^T W^{-1}\left(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right) > 0$.

Ex

- $a = W^{-1}d$ is normal to the discriminating hyperplane between the classes.
- Note: Fisher's LDA is distribution free.

RWTH AACHEN UNIVERSITY

5.1.2. Gaussian ML discrimination rule

Maximum likelihood rule: allocate observation x to the class C_ℓ which max. the likelihood $L_\ell(x) = \max_j L_j(x)$.

Assume that the class distributions are Gaussian and known.

$N_p(\mu_\ell, \Sigma_\ell)$, μ_ℓ, Σ_ℓ fixed and known with densities

$$f_\ell(u) = \frac{1}{(2\pi)^{p/2} |\Sigma_\ell|^{1/2}} \exp\left\{-\frac{1}{2}(u-\mu_\ell)^T \Sigma_\ell^{-1}(u-\mu_\ell)\right\}, \; u \in \mathbb{R}^p$$

Objective: given, $\max_\ell f_\ell(x)$

Th. 5.1.2 a) The ML discrimination rule allocates x to the class C_ℓ which max. $f_\ell(x)$ over $\ell = 1, \dots, g$.

b) If $\Sigma_\ell = \Sigma \; \forall \ell$ then the ML rule allocates x to the C_ℓ which ~~max~~ minimizes the Mahalanobis distance

$$(x-\mu_\ell)^T \Sigma^{-1}(x-\mu_\ell).$$

c) If $g=2$ and $\Sigma_1 = \Sigma_2 = \Sigma$ then the ML rule allocates x to the class C_1 if

$$\alpha^T(x-\mu) > 0$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$

Thinking the Future
Zukunft denken

Proof c) It holds

$$f_1(x) > f_2(x) \iff (x-\mu_1)^T \Sigma^{-1} (x-\mu_1) < (x-\mu_2)^T \Sigma^{-1} (x-\mu_2)$$

$$\overset{\text{Ex}}{\iff} (\mu_1-\mu_2)^T \Sigma^{-1} \left(x - \frac{1}{2}(\mu_1+\mu_2)\right) > 0 \quad \blacksquare$$

- c) is analog to Fisher's LDA for $g=2$.
- If μ_ℓ, Σ_ℓ are unknown use estimators from a training set.

Graphical example: