5. Classification and Clustering

→ given a set of data points

→ Goal: to put the points into subgroups which express closeness or similarity of the points

→ cluster head.

5.1. discriminant analysis

Suppose that g populations/groups/classes $C_1, \dots, C_g$ are given, each represented by a p.d.f. $f_i(x)$ on $\mathbb{R}^p$, $i = 1, \dots, p$

Often the densities $f_i(x)$ are completely unknown or its parameters such as its mean and variance must be estimated from data.

A discriminant rule divides $\mathbb{R}^p$ into disjoint regions $R_1, \dots, R_g$: $\bigcup_{i=1}^{p} R_i = \mathbb{R}^p$. The discriminant rule is defined by:

allocate some observation x to C_i if $x \in R_i$

5.1.1. Fisher's linear discriminant analysis

→ a training set with known class allocation is given → supervised learning

Suppose $x_1, \ldots, x_n$ with known labels are given.

When a new observation x with unknown label → a linear discriminant rule $\underline{\underline{a^T x}}$ is calculated such that x is allocated to some class in an optimal way.

$a \in \mathbb{R}^p$ $\quad X = [x_1 \cdots x_n]^T$ from $\underline{\underline{g}}$ groups $C_1, \ldots, C_g$

$X_j = [x_i]^T_{i \in C_j}$ $\quad n_\ell = |\{j : 1 \le j \le n, j \in C_\ell\}|$

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^p$ $\quad \bar{x}_\ell = \frac{1}{n_\ell} \sum_{j \in C_\ell} x_j$

$a^T x$: choose a such that the ratio of

the between group sum of squares and
?
The within group sum of squares is maximize
?

$$y_i = a^T x_i \implies y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = X^T a \in \mathbb{R}^p$$

$$X = [x_1 \cdots x_n]$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{y}_\ell = \frac{1}{n_\ell}\sum_{i \in C_\ell} y_i$$

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

Steiner's rule (*)

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

$$= \underbrace{\sum_{\ell=1}^{g} \sum_{j \in C_\ell} (y_i - \bar{y}_\ell)^2}_{\text{within group sum of squares}} + \underbrace{\sum_{\ell=1}^{g} n_\ell (\bar{y}_\ell - \bar{y})^2}_{\text{between group sum of squares}}$$

$E_n \circ E_{n_\ell} = E_\ell \qquad y_\ell = (y_i)_{i \in C_\ell}$

$$\sum_{\ell=1}^{g} \sum_{j \in C_\ell} (y_j - \bar{y}_\ell)^2 = \sum_{\ell=1}^{g} y_\ell^T E_\ell y_\ell \qquad \to E_\ell E_\ell = E_\ell$$

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

$$= a^T \left(\sum_{\ell=1}^{g} X_\ell^T E_\ell X_\ell \right) a$$

$$\left(W = \sum_{\ell=1}^{g} X_\ell^T E_\ell X_\ell \right) \qquad = a^T W a$$

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

$$= \sum_{\ell=1}^{g} n_\ell a^T (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})^T a$$

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

$$= a^T B a$$

$$a^T W a \downarrow \qquad a^T B a \uparrow \qquad \Rightarrow \max_{a \in \mathbb{R}^p} \frac{a^T B a}{a^T W a}$$

Thm 5.1. The ~~minimum~~ maximum value of

$$\max_{a \in \mathbb{R}^p} \frac{a^T B a}{a^T W a}$$

is attained at the eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue.

Proof.

$$\frac{a^T B a}{a^T W a} = \frac{(W^{1/2}a)^T W^{-1/2} B W^{-1/2} (W^{1/2}a)}{(W^{1/2}a)^T (W^{1/2}a)}$$

$$W \text{ n.n.d } \quad W^{1/2} W^{1/2} = W$$

Let $b = W^{1/2}a \Longrightarrow \quad \max_{b \in \mathbb{R}^p} \dfrac{b^T \overbrace{(W^{-1/2} B W^{-1/2})}^{\tilde{B}} b}{\underbrace{b^T b}_{= \|b\|_2^2}}$

$$\max_b \frac{b^T \tilde{B} b}{b^T b} = \max_b \left(\frac{b}{\|b\|}\right)^T \tilde{B} \left(\frac{b}{\|b\|}\right)$$

$$= \max_{\substack{b \in \mathbb{R}^p \\ \|b\|_2 = 1}} b^T (W^{-1/2} B W^{-1/2}) b = \lambda_{max}(W^{-1/2} B W^{-1/2})$$

v_{max} eigenvector corresponding to λ_{max}

$$W^{-1/2} B W^{-1/2} v_{max} = \lambda_{max} v_{max}.$$

$$(W^{-1/2} B W^{-1/2}) v_{max} = \lambda_{max} v_{max}$$

$$b = v_{max} \qquad b = W^{+1/2} a$$

$$W^{-1/2} B W^{-1/2} W^{1/2} a = \lambda_{max} W^{+1/2} a$$

$$\Rightarrow \quad \boxed{W^{-1} B a = \lambda_{max} a}$$

$\Rightarrow$ 1) Eigenvalues of $W^{-1}B$ and $W^{-1/2} B W^{-1/2}$ are the same.

2) a is the top eigenvector of $W^{-1}B$ $\square$.

The linear function $a^T x$ is called Fisher's linear discriminant function.

→ Given $x_1, \dots, x_n \in \mathbb{R}^p$ find a at the top eigenvector of $W^{-1}B$

→ for a new observation x, find $a^T x$

$$|a^T(x - \bar{x}_l)| < |a^T(x - \bar{x}_j)| \; \forall j \neq l$$

Discriminant rule: Allocate x to the group l

if $|a^T x - a^T \bar{x}_l| < |a^T x - a^T \bar{x}_j|$ for

all $j \neq l$.

Special case $g=2$

$$W = \sum_{l=1}^{2} X_l^T E_l X_l \qquad B = \sum_{l=1}^{2} n_l (\bar{x}_l - \bar{x})(\bar{x}_l - \bar{x})^T$$

$$n_1 \to C_1 \qquad n_2 \to C_2 \Rightarrow \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n}$$

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

$$= n_1 (\bar{x}_1 - \tfrac{n_1}{n}\bar{x}_1 - \tfrac{n_2}{n}\bar{x}_2)(\bar{x}_1 - \tfrac{n_1}{n}\bar{x}_1 - \tfrac{n_2}{n}\bar{x}_2)$$

$$+ n_2 (\bar{x}_2 - \tfrac{n_1}{n}\bar{x}_1 - \tfrac{n_2}{n}\bar{x}_2)(\bar{x}_2 - \tfrac{n_1}{n}\bar{x}_1 - \tfrac{n_2}{n}\bar{x}_2)$$

$n = n_1 + n_2$

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

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

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

$$= \left(\frac{n_1 n_2^2 + n_2 n_1^2}{n^2}\right) d d^T = \frac{n_1 n_2}{n} d d^T \quad * \quad 7$$

$d = \bar{x}_1 - \bar{x}_2$

$B = \frac{n_1 n_2}{n} d d^T \qquad rk(B) = 1 \Rightarrow rk(W^{-1}B) = 1$

$W^{-1}B$ n.n.d. $\Rightarrow \lambda_{max} \geq 0 \quad \lambda \neq \lambda_{max} \quad \lambda = 0$

- $Tr(W^{-1}B) = \sum_{i=1}^{n} \lambda_i = \lambda_{max}$

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

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

$$(W^{-1}B) v_{max} = \frac{n_1 n_2}{n} d^T \underline{W^{-1} d} \; v_{max}$$

$$\frac{n_1 n_2}{n} W^{-1} d d^T v_{max} = \quad ''$$

Let's see $(W^{-1}B)(W^{-1}d)$

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

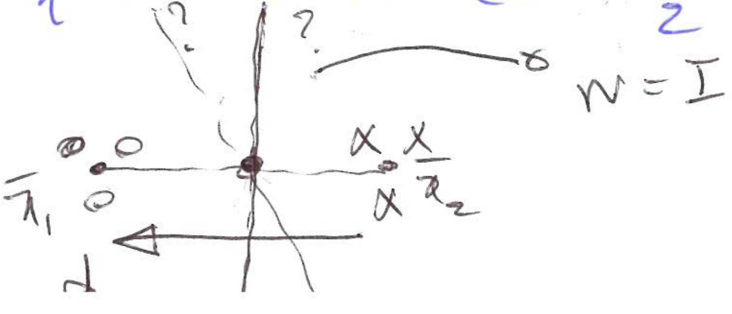
$$= \underbrace{\left(\frac{n_1 n_2}{n} d^T W^{-1} d\right)}_{\lambda_{max}} \underline{W^{-1} d}$$

$v_{max} = a$

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

(exercise)

$d = \bar{x}_1 - \bar{x}_2$



$$W = \sum_{l=1}^{g} X_l^T E_l X_l$$

5.1.2. Gaussian Maximum Likelihood (ML)

discriminant rule.

$$f_l(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1} (x-\mu_l)\right)$$

$l = 1, \dots, g$ the class distribution is

Gaussian as $N_p(\mu_l, \Sigma_l)$

$$L_l(x) = \max_j L_j(x)$$

Thm 5.2. The ML discriminant allocates

x to class C_l which ~~minimizes~~ maximizes $f_l(x)$ over

$l = 1, \dots, g$

a) If $\Sigma_l = \Sigma$ for all l, then the ML

rule allocates x to C_l which minimizes

the Mahalanobis distance

$$(x-\mu_l)^T \Sigma^{-1} (x-\mu_l)$$

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

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

$$\mu = \frac{1}{2}(\mu_1+\mu_2), \quad \alpha = \Sigma^{-1}(\mu_1-\mu_2).$$

Proof. part (a) follows from the definition.

part (b) an exercise. $\square$

If you do not know μ_ℓ, Σ_ℓ then

$$\hat{\mu}_\ell = \bar{x}_\ell \quad \text{and} \quad \hat{\Sigma}_\ell = \frac{1}{n_\ell} X_\ell^T E_\ell X_\ell$$