Discriminant Analysis

(1) <u>Background:</u>

· For m population $G_i \sim F_i(X)$. or we have samples from it. If we're given another sample x to be tested. How can we discriminate which one population x comes from?

(2) <u>Distance:</u>

· The ideal of discrimination by distance is: choose G_i. where $i = \arg\min_i d(x, G_i)$. and regard x is from G_i which is the nearest.

① <u>Mahalanobis Distance:</u>

<u>Def:</u> If $\vec{x}, \vec{\eta}$ are samples from population G whose mean is $\vec{\mu}$ and covariance $\Sigma = (\sigma_{ij})_{p \times p}$

Define: $d_m(\vec{x}, \vec{\eta}) = d(\vec{x}, \vec{\eta}) = (x-\eta)^T \Sigma^{-1} (x-\eta)$

$d_m(\vec{x}, G_i) = d^2(\vec{x}, G_i) = (x-\mu^i) \Sigma^{-1} (x-\mu^i)$

where μ^i is mean of G_i.

<u>Rmk:</u> d_m normalizes the variables so that it's indept with unit of measure.

② Discrimination:

i) μ^i's are known. $\Sigma^i = \Sigma$ known:

Rule: $\begin{cases} \eta \in G_1, \text{ if } d^2(\eta, G_1) < d^2(\eta, G_2) \\ \eta \in G_2, \text{ if } d^2(\eta, G_2) < d^2(\eta, G_1) \\ \text{undetermined, otherwise.} \end{cases}$

Note that: $d^2(\eta, G_1) - d^2(\eta, G_2)$

$$= 2(\eta - \frac{\mu^1+\mu^2}{2})^T \Sigma^{-1} (\mu^1 - \mu^2)$$

Denote: $\alpha = \Sigma^{-1}(\mu^1 - \mu^2)$, $\bar{\mu} = \mu^1+\mu^2/2$ ($\vec{\alpha}$ is projection direction!)

$\Rightarrow d^2(\eta, G_1) - d^2(\eta, G_2) = 2\alpha^T(\eta - \bar{\mu})$

ii) μ^i, Σ^i known:

$d^2(\eta, G_1) - d^2(\eta, G_2) = (\eta - \mu^1)^T \Sigma_1^{-1} (\eta - \mu^1) - (\eta - \mu^2)^T \Sigma_2^{-1} (\eta - \mu^2)$

iii) μ^i, Σ^i unknown:

Consider we have samples $\mathcal{X}_1 = (x_{ij}^{(1)})_{n_1 \times p}$ and $\mathcal{X}_2 = (x_{ij}^{(2)})_{n_2 \times p}$ from G_1, G_2.

Denote: $\bar{X}_i^{(t)} = \sum_{k=1}^{n_t} x_{ki}^{(t)} / n_t$, $t = 1, 2$

$$s_{ij}^{(t)} = \frac{1}{n_t - 1} \sum_k (x_{ki}^{(t)} - \bar{x}_i^{(t)})(x_{kj}^{(t)} - \bar{x}_j^{(t)})$$

$$S^{(t)} = (s_{ij}^{(t)})_{p \times p}, \quad \bar{X}^{(t)} = (x_1^{(t)} \cdots x_p^{(t)})$$

Def: $d^2(x, G^{(t)}) = (X - \bar{X}^{(t)})^T S^{(t)-1} (X - \bar{X}^{(t)})$

$\Rightarrow$ Find $i = \arg\min_k d^2(x, G_k)$

(3) Bayesian:

- Distance Discriminant doesn't take prior prob. of occurance in G_i into account and consider the loss of incorrect discrimination.

$\Rightarrow$ Suppose we have m populations $G_i \sim f_i(x)$ whose occurance prob is "q_i" based on the past datas.

① Maximum Posterior:

- Note that $P(G_i|x) = q_i f_i(x) / \sum_{k=1}^{m} q_k f_k(x)$

Find $l = \arg\max_k P(G_k|x) \Rightarrow x \in G_l$.

Rmk: When $f_i(x) \sim N_p(\mu^i, \Sigma^i)$.

since $\max_k P(G_k|x) \Leftrightarrow \max_k q_k f_k(x)$

$$q_k f_k(x) = \frac{q_k}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu^k)^T \Sigma_k^{-1} (x-\mu^k)\right)$$

② Minimum ECM:

- Suppose $D = \cup D_i$ a partition, if x falls into D_i. Then regard $x \in G_i$.

Next, find a partition $\cup D_i$ st. minimize the loss of incorrect discrimination.

Def: $p(j|i) = p(x \in D_j | G_i) = \int_{D_j} f_i(x)\,dx$.

$L(j|i)$ is loss coefficient. st. $L(i|i) = 0$.

Expected Cost of Misclassification of D:

$$E(M(D)) = \sum_{i=1}^{m} q_i \left(\sum_{j=1}^{m} L(j|i)\, p(j|i) \right)$$

We call $D^* = \arg\min_D E(M(D))$, the partition is a Bayesian Rule.

<u>Thm</u>. Denote $h_j(x) = \sum_{i=1}^{m} q_i L(j|i) f_i(x)$. Then: the Bayesian Rule is $D_i^* = \{x \mid h_i(x) \le h_j(x),\ \forall j \ne i\}$

<u>Pf</u>: If $D = \bigcup_i^m D_i$ is another partition.

$$\Rightarrow E(M(D^*)) - E(M(D))$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} \int_{D_i^* \cap D_j} [h_i(x) - h_j(x)]\, dx \le 0$$

<u>Rmk</u>: i) It means: find $l = \arg\min_i h_i(x)$.

$\Rightarrow$ Discriminate $x \in G_l$.

ii) Note $h_i(x) = \sum_1^m q_k f_k(x) - q_i f_i(x)$

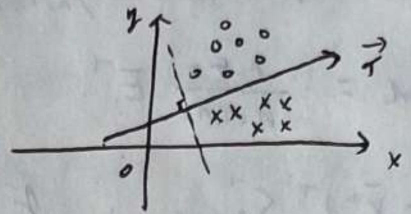
$\Leftrightarrow$ maximize $q_i f_i(x)$. It's equi. with criteria ①. if $L(j|i) = 1 - \delta_{ij}$

(4) <u>Fisher Discrimination</u>:

· The ideal is projecting the datas into several direction to separate each class significantly. Then choose a criteria to classify.

Def:

· We want to project the data to some direction which can separate different kinds of data as most as possible.



Rmk: As we did before. When

$\Sigma^1 = \Sigma^2 = \Sigma$. Then To minimize $E(D)$:

We classify D_1 if $(\mu_1^T - \mu_2^T)\Sigma^{-1}x - k \geq \ln \frac{L(1|2)}{L(2|1)} \frac{p_2}{p_1}$

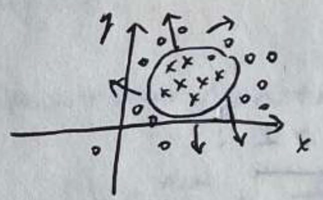
It projects to $(\mu_1 - \mu_2)^T \Sigma^{-1}$. (The direction)

But when $\Sigma^1 \neq \Sigma^2$. The allocation is:

D_1 if $-\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x + \mu_1^T \Sigma_1^{-1} x - \mu_2^T \Sigma_2^{-1} x \geq k + \ln \square$

It's quadratic form:

(i.e. separate by elliptic)



proceed: Set $Y = C^T X$ s.t. minimize the variance in the same group and maximize variance between different groups. i.e.

Set $E_0 = \sum_{i=1}^{m} V_y^i = C^T \sum_i^m V_x^i C = C^T E C$

where $E = \sum V_x^i = \sum_{i,t} (X_t^{(i)} - \bar{X}^{(i)})(X_t^{(i)} - \bar{X}^{(i)})^T$

$B_0 = \sum_i^m n_i (\bar{Y}^{(i)} - \bar{Y})^2 = C^T \sum (\bar{X}^{(i)} - \bar{X})(\bar{X}^{(i)} - \bar{X})^T n_i C$

$= C^T B C$

maximize: $A^2(C) = \frac{C^T B C}{C^T E C}$. i.e. $\max = \max\{\lambda \mid \lambda \in \sigma_{E^{-1}B}\}$

Rmk: If number of group is large. Then we will find $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_k$. eigenvalues of $E^{-1}B$.

i) <u>When $m=2$:</u>

Then $B = \frac{n_1 n_2}{n_1+n_2} (\bar{X}^1 - \bar{X}^2)(\bar{X}^1 - \bar{X}^2)^T$, $r(B) = 1$

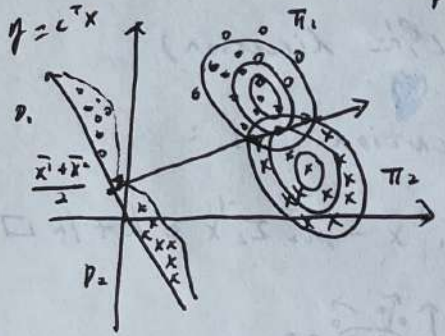
$E^{-1}B$ has max eigenvalue: $\frac{1}{1/n_1 + 1/n_2} (\bar{X}^1 - \bar{X}^2)^T E^{-1} (\bar{X}^1 - \bar{X}^2)$

Denote $S_{pooled}^{-1} = \frac{n_1 n_2}{n_1+n_2} E^{-1}$.

$\Rightarrow c^T = (\bar{X}^1 - \bar{X}^2)^T S_{pooled}^{-1}$, $y = c^T X$. suppose $\bar{X}^1 > \bar{X}^2$

Allocate X to D_1 if $y > \frac{1}{2} c^T (\bar{X}^1 + \bar{X}^2)$

D_2 if $y < \frac{1}{2} c^T (\bar{X}^1 + \bar{X}^2)$



<u>Rmk:</u> When $y = \frac{1}{2} c^T (\bar{X}^1 + \bar{X}^2)$

It doesn't work.

ii) <u>When $m > 2$:</u>

$E^{-1}B$ may have different eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$

suppose each corresponds c_i. the project direction

Criteria Functions: $u_i(x) = c_i^T x$. $\bar{u}_i^t = c_i^T \bar{X}^t$

If $\exists$ unique i_0 s.t.

$i_0 = \arg\min_t |u_1(x) - \bar{u}_1^t| / \hat{\sigma}_t^1$. $\hat{\sigma}_t^i = c_i^T S_t c_i$

Then allocate x to π_{i_0}.

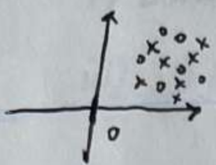
Otherwise. Consider using $u_2(x)$, $u_3(x)$ ...

until exists a unique π_α.

<u>Rmk:</u> We may choose $\{u_i(x)\}_{i=1}^r$ s.t. $\frac{\sum_1^r \lambda_i}{\sum_1^s \lambda_i} > p_0$.

p_0 is the expected efficient.

① Classification sometimes may not be a good ideal for the data:



So. before separating the data.

We can apply T^2's test:

$H_0: \mu_1 = \mu_2$ v.s. $H_1: \mu_1 \neq \mu_2$

Rmk: Siginificant separation $\not\Rightarrow$ good Classification

② Evaluting classification function:

i) Total prob of misclassification $TPM = \sum p_i \int_{\mathbb{R}^n / D_i} f_i(x)\,dx$

ii) Actual error rate $AER = \sum p_i \int_{\mathbb{R}^n / \hat{D}_i} \hat{f}_i(x)\,dx$. $\hat{f}_i = f_i(x \mid \hat{\theta})$

$\hat{D}_i$ is separation based on samples.

iii) Apparent error rate $APER = n_{1m} + n_{2m} / (n_1 + n_2)$

		π_1	π_2	
Actual	π_1	n_{1c}	n_{1m}	$= n_1$
	π_2	n_{2c}	n_{2m}	$= n_2$

prediction

$$= \frac{n_1}{n_1+n_2} \cdot \frac{n_{1m}}{n_1} + \frac{n_2}{n_1+n_2} \cdot \frac{n_{2m}}{n_2}$$

(n_i / n_1+n_2 can estimate the rate of class π_i occurs)

Rmk: i) need the information of populations (D_i. f_i ...)

ii) need the information from samples and pdf's.

iii) doesn't depends on density. But it may under-estimate the error rate. Since it totally bases on the datas —— which separates them as most.

$\Rightarrow$ Cross-Validation method: (Leave one out)

Start with π_1 group. omit one obeservation. $\Rightarrow$ Develop classification on remain $n_1 - 1 + n_2$ ob's. $\Rightarrow$ Classify the omit ob. $\xRightarrow[\text{on } \pi_1]{\text{rep.}}$ Caculate misclassification $n_{1m}^{(H)} / n_1$ $\Rightarrow$ On π_2

$\Rightarrow$ Classify π_2 group. obtain $n_{2m}^{(H)} / n_2$ $\Rightarrow$ $\hat{E}(AER) = \dfrac{n_{1m}^{(H)} + n_{2m}^{(H)}}{n_1 + n_2}$