Discrimination & Classification

Discriminant analysis : optimal way to separate heterogeneous populations

Classification: Classification of new observation vector in one of the possible populations (Classes) using discriminant function.

Examples / Applications

Medical diagnostics, alert systems for adverse events, credit card fraud detection, loan classification, - - - - .

Data structure

Learning sample

$$\mathcal{L} = \{(\underset{\sim}{x}_1, j_1), \ldots, (\underset{\sim}{x}_n, j_n)\}$$

n preclassified examples

$\underset{\sim}{x}_i \in \mathcal{X}$, feature space $\forall\ i = 1(1)n$

$j_i \in \{1, 2, \ldots J\}$ $\forall\ i = 1(1)n$

J : possible number of classes.

$C = \{1, 2, \ldots, J\}$ - set of classes

aim : Given $\underset{\sim}{x} \in \mathcal{X}$; to find a systematic way of predicting class membership

i.e. assigns one of the classes in $\{1, 2, \ldots J\}$ to $\underset{\sim}{x} \in \mathcal{X}$.

Def^n: Classifier

A classifier or a classification rule is a f^n $d(.)$ defined on $\mathcal{X}$ $\ni$ $\forall$ $\underset{\sim}{x}$, $d(\underset{\sim}{x})$ is ~~one~~ equal to one of the numbers $1, 2, \ldots J$.

Note: An alternate way to look at a classifier is through partitions

Let $A_j = \{ \underset{\sim}{x} : d(\underset{\sim}{x}) = j \}$

$A_1, \ldots, A_J$ are $\ni$

$$A_i \cap A_j = \phi \quad \forall\ i \neq j$$

$$\& \quad \bigcup_i A_i = \mathcal{X}$$

Alt def^n: A classifier is a partition of $\mathcal{X}$ into J disjoint set $A_1, \ldots, A_J$ $\ni$ $\forall$ $\underset{\sim}{x} \in A_j$ the assigned class membership is j.

Fisher Linear Discriminant Function (FLDF)

Consider a 2-class problem

i.e. 2 pop^{n}s π_1 & π_2

$$\underset{\sim}{X} \mid \pi_1 \sim \underset{\sim}{\mu}_1, \Sigma \;;\; \underset{\sim}{\mu}_1 = E(\underset{\sim}{X} \mid \pi_1),\ \Sigma = Cov(\underset{\sim}{X} \mid \pi_1)$$

$\Sigma > 0$;

$$\underset{\sim}{X} \mid \pi_2 \sim \underset{\sim}{\mu}_2, \Sigma \;;\; \underset{\sim}{\mu}_2 = E(\underset{\sim}{X} \mid \pi_2),\ \Sigma = Cov(\underset{\sim}{X} \mid \pi_2)$$

Aim:
- To find a f^n $g(.)$ $\ni$ If $\underset{\sim}{X}_1$ is from π_1 & $\underset{\sim}{X}_2$ is from π_2, then $g(\underset{\sim}{X}_1)$ & $g(\underset{\sim}{X}_2)$ should look as "different as possible". Such a $g(.)$ will be the constructed discriminant f^n.
- Given a new obsn $\underset{\sim}{x}$, use $g(.)$ to classify it to one of π_1 & π_2.

FLDF approach: change π_1 & π_2 into univariate pop^n by transforming $\underset{\sim}{X}$ to $\underset{\sim}{l}'\underset{\sim}{X}$

$$\Rightarrow \quad \underset{\sim}{l}'\underset{\sim}{X} \mid \pi_1 \sim \underset{\sim}{l}'\underset{\sim}{\mu}_1,\ \underset{\sim}{l}'\Sigma\underset{\sim}{l}$$

$$\& \ \underset{\sim}{l}'\underset{\sim}{X} \mid \pi_2 \sim \underset{\sim}{l}'\underset{\sim}{\mu}_2,\ \underset{\sim}{l}'\Sigma\underset{\sim}{l}$$

Note that

Separate out the 2 univariate pop^n as much as possible

$\Leftrightarrow$ maximization of distance bet^n the 2 pop^{n}s w.r.t. $\underset{\sim}{l}$

Take the distance betn the 2 univariate popns as

$$\frac{(\underline{l}'\underset{\sim}{\mu}_1 - \underline{l}'\underline{\mu}_2)^2}{\underline{l}'\Sigma\underline{l}} \quad - (*)$$

Maximize $(*)$ w.r.t. $\underline{l}$

i.e. find $\underset{\sim}{l}^* = \arg\max_{\underset{\sim}{l}} \dfrac{(\underline{l}'(\underset{\sim}{\mu}_1 - \underline{\mu}_2))^2}{\underline{l}'\Sigma\underline{l}}$

Define $\underset{\sim}{a}' = \underline{l}'\Sigma^{1/2}$

then
$$\frac{(\underline{l}'(\underset{\sim}{\mu}_1 - \underline{\mu}_2))^2}{\underline{l}'\Sigma\underset{\sim}{l}}$$
$$= \frac{(\underset{\sim}{a}'\Sigma^{-1/2}(\underset{\sim}{\mu}_1 - \underline{\mu}_2))^2}{\underline{a}'\underset{\sim}{a}} \quad - (*^2)$$

By C-S inequality

$$(*)^2 \leq \frac{(\underline{a}'\underline{a})(\underset{\sim}{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underline{\mu}_2)}{\underline{a}'\underset{\sim}{a}}$$
$$= (\underset{\sim}{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underline{\mu}_2)$$

square of Mahalonobis distance

Note that ~~in~~ equality in the above is attained

at $\qquad \underset{\sim}{a}' = (\underset{\sim}{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1/2}$

i.e. $\quad \underline{l}'\Sigma^{1/2} = (\underset{\sim}{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1/2}$

i.e. $\underset{\sim}{l}' = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1}$

So we get the FLDF as

$g(\underset{\sim}{X}) = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X}$ as the discriminant f^n.

Next task is to frame the classification rule based on FLDF

i.e. given a new obsn $\underset{\sim}{x}_0$, to assign it to π_1 or π_2

Realize that

$E\left((\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X} \mid \pi_1 \right) = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{\mu}_1 = m_1$, say

& $E\left((\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X} \mid \pi_2 \right) = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{\mu}_2 = m_2$, say

and $m_1 - m_2 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$

$\Rightarrow m_1 - m_2 \geqslant 0$; equal to 0 only if $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$

i.e. $m_1 \geqslant m_2$

Let $y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{x}_0$

we will assign $\underset{\sim}{x}_0$ to π_1 if

y_0 is closer to m_1 than to m_2

i.e. If $y_0 \geq \frac{m_1+m_2}{2}$; we assign $\underset{\sim}{x_0}$ to π_1

i.e If $y_0 \geq \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)$

and assign $\underset{\sim}{x_0}$ to π_2 if

$$y_0 < \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)$$

Thus the classification partition associated with FLDF is (R_1, R_2)

$$R_1 = \left\{ \underset{\sim}{x} : (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{x} \geq \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2) \right\}$$

$$R_2 = \left\{ \underset{\sim}{x} : (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{x} < \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2) \right\}$$

<u>Remark</u>: Usually $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$, Σ are unknown. We use the learning sample $\mathcal{L}$, of preclassified examples, to estimate $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$, Σ and get sample analogue of $(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{x}$ as

$$\underline{(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} \underset{\sim}{x}}$$

This is called the Fisher sample linear discriminant function.

where, $(\bar{x}_1, S_1)$ are sample mean vector and sample variance covariance matrix computed from pre-classified examples of π_1 class.

Sly, $(\bar{x}_2, S_2) \rightarrow$ computed from pre-classified examples of π_2 class.

$$S = \frac{(n_1-1) S_1 + (n_2-1) S_2}{n_1+n_2-2}$$

n_1 : # of pre-classified π_1 cases in $\mathcal{L}$

n_2 : # of pre-classified π_2 cases in $\mathcal{L}$

$$\hat{R}_1 = \{ x : (\bar{x}_1 - \bar{x}_2)' S^{-1} x \geq \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2) \}$$

$$\hat{R}_2 = \{ x : (\bar{x}_1 - \bar{x}_2)' S^{-1} x < \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2) \}$$

General Classification problem

Two-class problem

π_1 & π_2 : two classes

Suppose,

$\underset{\sim}{X} \mid \pi_1$ has support $\mathcal{X}_1$ $\qquad \mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$

$\underset{\sim}{X} \mid \pi_2$ has support $\mathcal{X}_2$

if $\mathcal{X}_1 \cap \mathcal{X}_2 = \phi$, then there is no misclassification

But usually $\mathcal{X}_1 \cap \mathcal{X}_2 \neq \phi$ and there is a chance that observations can be misclassified.

Let

$c(i|j)$ denote the cost when an obsn from π_j is misclassified as from π_i

$c(1|1) = c(2|2) = 0$

Suppose the class conditional densities are

$f_i(\underline{x})$; $i = 1, 2$

i.e. $\underset{\sim}{X} \mid \pi_i$ has p.d.f. $f_i(\underline{x})$ $\quad i = 1, 2$

and the prior probabilities for π_1 and π_2 be p_1 and p_2, respy.

Let

$P(i|j)$: probability of misclassifying an observation from π_j into π_i

Suppose the classification partition is $\{R_1, R_2\}$, (i.e. if $\underset{\sim}{x} \in R_i$, we classify it to π_i). then

$$P(i|j) = \int_{R_i} f_j(\underline{x})\, d\underline{x}$$

e.g. $$P(2|1) = \int_{R_2} f_1(\underline{x})\, d\underline{x}$$

& $$P(1|2) = \int_{R_1} f_2(\underline{x})\, d\underline{x}$$

Also P(an obsn comes from π_i and is misclassified to π_j) $= p_i\, P(j|i)$

Define

- Total Probability of misclassification (TPM)

$$= p_1\, P(2|1) + p_2\, P(1|2)$$

- Expected cost of misclassification (ECM)

$$= C(2|1)\left(p_1 P(2|1)\right) + C(1|2)\left(p_2 P(1|2)\right)$$

Optimal Strategies

To find the partition $\{R_1^*, R_2^*\}$, say, $\ni$

TPM or ECM is minimised.

(I) TPM minimizing partition

$$TPM = p_2 \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x} + p_1 \int_{R_2} f_1(\underset{\sim}{x})\, d\underset{\sim}{x}$$

$$= p_2 \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x} + p_1 \int_{\mathcal{X} - R_1} f_1(\underset{\sim}{x})\, d\underset{\sim}{x}$$

$$= p_2 \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x} + \int_{\mathcal{X}} p_1 f_1(\underset{\sim}{x})\, d\underset{\sim}{x} - \int_{R_1} p_1 f_1(\underset{\sim}{x})\, d\underset{\sim}{x}$$

i.e.

$$TPM = \int_{R_1} \left(p_2 f_2(\underset{\sim}{x}) - p_1 f_1(\underset{\sim}{x})\right) d\underset{\sim}{x} + \underbrace{\int_{\mathcal{X}} p_1 f_1(\underset{\sim}{x})\, d\underset{\sim}{x}}_{\text{indep of partition}}$$

Minimization of TPM w.r.t. $\{R_1, R_2\}$

$\Leftrightarrow$

minimization of $\displaystyle\int_{R_1} \left(p_2 f_2(\underset{\sim}{x}) - p_1 f_1(\underset{\sim}{x})\right) d\underset{\sim}{x}$

Note that $\forall\, \underset{\sim}{x} \in \mathcal{X}$; $\left(p_2 f_2(\underset{\sim}{x}) - p_1 f_1(\underset{\sim}{x})\right)$ is either ≤ 0

or > 0.

Thus, the TPM is minimized by choosing

$$R_1^* = \{ \underset{\sim}{x} : p_2 f_2(\underline{x}) \le p_1 f_1(\underline{x}) \}$$

i.e. $$R_1^* = \left\{ \underset{\sim}{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{p_2}{p_1} \right\}$$

& $$R_2^* = \left\{ \underset{\sim}{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} < \frac{p_2}{p_1} \right\}$$

In other words,

TPM minimizing classification rule is

assign $\underset{\sim}{x}$ to π_1 if $\dfrac{f_1(\underline{x})}{f_2(\underset{\sim}{x})} \geq \dfrac{p_2}{p_1}$

& to π_2 if o/w.

<u>Remark</u>: Consider testing of

$$H_0 : \underset{\sim}{X} \sim f_1(\underline{x}) \text{ ag } H_A : \underset{\sim}{X} \sim f_2(\underline{x})$$

By Neyman-Pearson lemma, the MP(α) test

has critical region

$$W = \left\{ \underset{\sim}{x} : \frac{f_2(\underline{x})}{f_1(\underline{x})} > K \right\}$$

Where K is $\ni$

$$P_{f_1}\left(\frac{f_2(\underline{X})}{f_1(\underline{X})} > K \right) = \alpha$$

Thus we observe that

$$R_2^* = \left\{ \underset{\sim}{x} : \frac{f_2(\underset{\sim}{x})}{f_1(\underset{\sim}{x})} > \frac{p_1}{p_2} \right\}$$

corresponds to the critical region of an MP test of a fixed size (fixed by $\frac{p_1}{p_2}$)

$$P_{f_1}\left(\frac{f_2(\underset{\sim}{x})}{f_1(\underset{\sim}{x})} > \frac{p_1}{p_2} \right) = \alpha^*, \text{ say.}$$

Example :

Π_1

$f_1(\underset{\sim}{x})$

$x_2 \backslash x_1$	1	2	3
1	.1	.05	.15
2	.25	.2	.25

Π_2

$f_2(\underset{\sim}{x})$

$x_2 \backslash x_1$	1	2	3
1	.2	.2	.2
2	.2	.1	.1

$$p_1 = p_2 = \frac{1}{2}$$

TPM minimizing rule is

assign $\underset{\sim}{x}$ to Π_1 If $f_1(\underset{\sim}{x}) \geq f_2(\underset{\sim}{x})$

i.e. $R_1^* = \{ \underset{\sim}{x} : f_1(\underset{\sim}{x}) \geq f_2(\underset{\sim}{x}) \}$

$\Rightarrow \quad R_1^* = \{ (1,2), (2,2), (3,2) \}$

$R_2^* = \{ (1,1), (2,1), (3,1) \}$

$$P(1|2) = \sum_{\underset{\sim}{x} \in R_1^*} f_2(\underline{x})$$

$$= 0.2 + 0.1 + 0.1 = 0.4$$

$$P(2|1) = \sum_{\underset{\sim}{x} \in R_2^*} f_1(\underline{x})$$

$$= 0.1 + 0.05 + 0.15 = 0.3$$

$$TPM(R_1^*, R_2^*) = \frac{1}{2}\left(P(1|2) + P(2|1)\right)$$

$$= 0.5(0.4 + 0.3)$$

Note: $p_1 \neq p_2$

If $p_1 = 0.4$ & $p_2 = 0.6$ then

$$R_1^* = \{\underset{\sim}{x} : 0.4 f_1(\underline{x}) \geq 0.6 f_2(\underline{x})\}$$

$$R_2^* = \{\underset{\sim}{x} : 0.4 f_1(x) < 0.6 f_2(\underline{x})\}$$

. . . .

Example 2 :

$$\pi_1 \equiv N_p(\underset{\sim}{\mu}_1, \Sigma)$$

$$\pi_2 = N_p(\underline{\mu}_2, \Sigma) \quad ; \quad \Sigma > 0$$

$$p_1 = p_2$$

Optimum TPM rule is

$$R_1^* = \left\{\underset{\sim}{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq 1\right\}$$

$$R_2^* = \left\{\underline{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} < 1\right\}$$

Now,
$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq 1.$$

$\Leftrightarrow$
$$(2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)\right)$$
$$\geq (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\right)$$

i.e. $(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1) \leq (\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)$

i.e. $\cancel{\underline{x}'\Sigma^{-1}}\underline{x} + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - 2\underline{\mu}_1'\Sigma^{-1}\underline{x}$
$$\leq \cancel{\underline{x}'\Sigma^{-1}}\underline{x} + \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2 - 2\underline{\mu}_2'\Sigma^{-1}\underline{x}$$

i.e. $-2(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} \leq \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2 - \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1$

i.e. $-2(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} \leq \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2 - \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1$
$$+ \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_2$$

i.e. $-2(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} \leq -(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2)$

i.e. $(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underline{x} \geq \frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2)$

Hence the TPM minimizing rule is

Assign $\underset{\sim}{x}$ to π_1 if

$$(\underset{\sim}{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{x} \geq \frac{1}{2} (\underset{\sim}{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underline{\mu}_2)$$

to π_2 if o/w

<u>Remark</u>:

The above rule is same as FLDF

<u>Remark</u>: If $\Pi_1 \equiv N_p(\underset{\sim}{\mu}_1, \Sigma)$

$\Pi_2 \equiv N_p(\underline{\mu}_2, \Sigma)$ $\quad\quad \Sigma > 0$

then

$$R_1^* = \left\{ \underset{\sim}{x} : (\underset{\sim}{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{x} \geq \frac{1}{2} (\underset{\sim}{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underline{\mu}_2) + \log\left(\frac{p_2}{p_1}\right) \right\}$$

$$R_2^* = \left\{ \underset{\sim}{x} : (\underset{\sim}{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{x} < \frac{1}{2} (\underset{\sim}{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underline{\mu}_2) + \log\left(\frac{p_2}{p_1}\right) \right\}$$

<u>Remark</u>: $\mathcal{L}$, the learning sample is to be used to obtain estimated partition $(\hat{R}_1^*, \hat{R}_2^*)$

$$\hat{R}_1^* = \left\{ \underset{\sim}{x} : (\bar{\underset{\sim}{x}}_1 - \bar{\underline{x}}_2)' S^{-1} \underset{\sim}{x} \geq \frac{1}{2} (\bar{\underset{\sim}{x}}_1 - \bar{\underline{x}}_2)' S^{-1} (\bar{\underset{\sim}{x}}_1 + \bar{\underline{x}}_2) + \log\left(\frac{p_2}{p_1}\right) \right\}$$

<u>Remark</u>:

$(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} \underset{\sim}{x} - \frac{1}{2} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} (\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)$ is called Anderson's classification statistic

Optimum classification rules (ECM)

Recall that we have earlier derived optimum classification rule minimizing "Total Probability of Misclassification (TPM)".
We now consider the problem of deriving optimum classification rule minimizing "Expected Cost of Misclassification (ECM)".

You may recall that for a 2-class problem, ECM is given by

$$ECM = p_1 P(2|1)\, C(2|1) + p_2 P(1|2)\, C(1|2)$$

p_1 & p_2 are prior probabilities of classes π_1 & π_2

$C(i|j)$: Cost of misclassifying an obsn from π_j into π_i

$P(i|j)$: prob of misclassifying an obsn from π_j into π_i

$$P(1|2) = \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x} \;;\; P(2|1) = \int_{R_2} f_1(\underset{\sim}{x})\, d\underset{\sim}{x}$$

$f_i(\underset{\sim}{x})$: class conditional density of class π_i

$\{R_1, R_2\}$ is the classification partition

$$ECM = p_1 C(2|1) \int_{R_2} f_1(\underset{\sim}{x})\, d\underset{\sim}{x} + p_2 C(1|2) \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x}$$

$$= p_1 C(2|1) \int_{\mathcal{X} - R_1} f_1(\underset{\sim}{x})\, d\underset{\sim}{x} + p_2 C(1|2) \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x}$$

$$= \int_{R_1} \left(p_2 C(1|2) f_2(\underset{\sim}{x}) - p_1 C(2|1) f_1(\underset{\sim}{x}) \right) d\underset{\sim}{x}$$

$$+ \left(p_1 C(2|1) \int_{\mathcal{X}} f_1(\underset{\sim}{x})\, dx \right) \leftarrow \text{indep of partition } \{R_1, R_2\}.$$

The above ECM is minimized w.r.t. $\{R_1, R_2\}$ by choosing

$$R_1^* = \{ \underset{\sim}{x} : p_2 C(1|2) f_2(\underset{\sim}{x}) \le p_1 C(2|1) f_1(\underset{\sim}{x}) \}$$

$$R_2^* = \mathcal{X} - R_1^* = \{ \underset{\sim}{x} : p_2 C(1|2) f_2(\underset{\sim}{x}) > p_1 C(2|1) f_1(\underset{\sim}{x}) \}$$

(logic is same as used for deriving opt TPM rule in class)

Remark: If $C(1|2) = C(2|1)$, then opt ECM rule is same as TPM rule.

Remark: The opt ECM partition $\{R_1^*, R_2^*\}$ correspond to a Most Powerful (MP) test for testing

$H_0: \underset{\sim}{X} \sim f_1(\underset{\sim}{x})$ ag $H_A: \underset{\sim}{X} \sim f_2(\underset{\sim}{x})$

This MP test is of fixed size.

(Recall the logic given in class during discussion on connection betn opt TPM partition and MP test using Neyman-Pearson lemma)

Example 1: $\Pi_1 : N_p(\underset{\sim}{\mu}_1, \Sigma)$

$\Pi_2 : N_p(\underset{\sim}{\mu}_2, \Sigma) \qquad \Sigma > 0$

$$p_1 f_1(\underset{\sim}{x}) C(2|1) \geq p_2 f_2(\underset{\sim}{x}) C(1|2)$$

$$\Leftrightarrow \quad \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \frac{p_2 C(1|2)}{p_1 C(2|1)}$$

$$\Leftrightarrow \quad \frac{(2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)' \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu}_1))}{(2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu}_2))} \geq \frac{p_2 C(1|2)}{p_1 C(2|1)}$$

$\Leftrightarrow$

$$-\frac{1}{2}\left((\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1) - (\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right) \geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)}\right).$$

$\Leftrightarrow$

$$-\frac{1}{2}\left(\cancel{\underset{\sim}{x}'\Sigma^{-1}}\underset{\sim}{x} + \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1 - 2\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x} - \cancel{\underset{\sim}{x}'\Sigma^{-1}}\underset{\sim}{x} - \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2 + 2\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{x}\right) \geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)}\right)$$

$\Leftrightarrow$

$$-\frac{1}{2}\left(-2\underbrace{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}}_{\text{FLDF}} - \left(\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1 + \underbrace{\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_2}_{=0}\right)\right) \geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)}\right)$$

$\Leftrightarrow$

$$-\frac{1}{2}\left(-2(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} - (\underset{\sim}{\mu}_2-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)\right) \geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)}\right)$$

i.e. $R_1^* = \left\{\underset{\sim}{x} : +\frac{1}{2}\left(2(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} - (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)\right) \geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)}\right)\right\}$

$R_2^* = \mathcal{X} - R_1^*$

<u>Sp. Case</u> : If $p_2\, C(1|2) = p_1\, C(2|1)$, then

$$\tilde{R}_1^* = \left\{\underset{\sim}{x} : (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} \geq \frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)\right\}$$

which is same as classification partition based on FLDF.

<u>Remark</u> : R_1^* given above estimated from learning sample $\mathcal{L}$ of ~~pre~~ preclassified examples.

Example 2

$$\pi_1 : N_p(\underset{\sim}{\mu}_1, \Sigma_1)$$

$$\pi_2 : N_p(\underline{\mu}_2, \Sigma_2) \; ; \; \Sigma_i > 0$$

$$R_1^* = \{ \underset{\sim}{x} : p_1 f_1(\underline{x}) C(2|1) \geq p_2 f_2(\underline{x}) C(1|2) \}$$

Proceeding as in example 1, we get

$$p_1 f_1(\underline{x}) C(2|1) \geq p_2 f_2(\underline{x}) C(1|2)$$

$\Leftrightarrow$

$$-\frac{1}{2}\left((\underset{\sim}{x} - \underset{\sim}{\mu}_1)' \Sigma_1^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_1) - (\underset{\sim}{x} - \underline{\mu}_2)' \Sigma_2^{-1} (\underline{x} - \underline{\mu}_2) \right)$$
$$- \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) \geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)} \right)$$

$$\Leftrightarrow -\frac{1}{2}\left(\underset{\sim}{x}' (\Sigma_1^{-1} - \Sigma_2^{-1}) \underset{\sim}{x} \right) + (\underset{\sim}{\mu}_1' \Sigma_1^{-1} - \underline{\mu}_2' \Sigma_2^{-1}) \underset{\sim}{x}$$
$$\geq \underbrace{\log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)} \right) + \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2}(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 - \underline{\mu}_2' \Sigma_2^{-1} \underline{\mu}_2)}_{\text{indep of } \underset{\sim}{x}}$$

The L.h.s. in above expression is called the quadratic discriminant f^n (for obvious reason).

Assignment rule:

For a new $\underset{\sim}{x}_0$;

Assign $\underset{\sim}{x}_0$ to π_1 if

$$-\frac{1}{2} \underset{\sim}{x}_0' (\Sigma_1^{-1} - \Sigma_2^{-1}) \underset{\sim}{x}_0 + (\underset{\sim}{\mu}_1' \Sigma_1^{-1} - \underline{\mu}_2' \Sigma_2^{-1}) \underset{\sim}{x}_0$$
$$\geq \log\left(\frac{p_2\, C(1|2)}{p_1\, C(2|1)} \right) + \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2}\left(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 - \underline{\mu}_2' \Sigma_2^{-1} \underline{\mu}_2 \right)$$

Remark: Note that we have to use the learning sample $\mathcal{L}$ to have sample counterpart $\{\hat{R}_1, \hat{R}_2\}$

If in $\mathcal{L}$ there are n_1 preclassified obsns from Π_1 & n_2 preclassified cases from Π_2 ($n_1 + n_2 = n \leftarrow$ total preclassified cases in $\mathcal{L}$), then use n_1 obsns of preclassified Π_1 examples to calculate $\hat{\underset{\sim}{\mu}}_1 = \bar{\underset{\sim}{x}}_{(1)}$ & $\hat{\Sigma}_1 = S_{(1)}$

Sly $\hat{\underline{\mu}}_2 = \bar{\underset{\sim}{x}}_{(2)}$ & $\hat{\Sigma}_2 = S_{(2)}$ from n_2 preclassified cases from Π_2.

Note that in this case no pooling is required as would be necessary in example 1 for estimation of common covariance matrix.

Example 3 : Discrete distⁿ case

Π_1 p.m.f.

$x_2 \backslash x_1$	0	1	2
0	.1	.05	.15
1	.25	.2	.25

Prior prob $p_1 = 0.4$

Π_2 p.m.f

$x_2 \backslash x_1$	0	1	2
0	.2	.2	.2
1	.2	.1	.1

Prior prob $p_2 = 0.6$

Misclassification cost

$$C(1|2) = 5 \qquad C(2|1) = 10$$

$$C(1|1) = C(2|2) = 0 \text{ (always)}$$

$\forall$ possible pairs $(0,0), (0,1), (1,0), (1,1)$
$(2,0), (2,1)$;

Check the assignment

e.g. $(0,0) = \underset{\sim}{x}$

$$p_1 f_1(\underset{\sim}{x}) C(2|1) = 0.4 \times 0.1 \times 10 = 0.4$$

$$\& \quad p_2 f_2(\underset{\sim}{x}) C(1|2) = 0.6 \times 0.2 \times 5 = 0.6$$

Recall that $R_1^* = \{\underset{\sim}{x} : p_1 f_1(\underset{\sim}{x}) C(2|1) \geq p_2 f_2(\underset{\sim}{x}) C(1|2)\}$

$\& \; R_2^* = \{\underset{\sim}{x} :$ " $<$ " $\}$

Hence assign rule for $(0,0)$ is Π_2

Similarly assignment rules for all other possible (x_1, x_2) can be computed.