Discrimination and classification (I)

- Objectives
- Two populations
- Minimum expected cost of misclassification rule
- Minimum total probability of misclassification rule
- Two multivariate normal populations
  - Equal population variances
  - Unequal population variances

Objectives

- Assume the existence of clusters which are known a priori.
- Construct discrimination rules to allocate new observations into these known groups.

Two populations

Two populations: π_1 and π_2

Classification measurements $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$

An object with measurements $\boldsymbol{x}$, must be assigned into either π_1 or π_2

Let $\boldsymbol{\Omega}$ be the sample space, and

R_1 = set of $\boldsymbol{x}$ for which the object is being classified into π_1

R_2 = set of $\boldsymbol{x}$ for which the object is being classified into π_2

In addition, $R_1 \cup R_2 = \boldsymbol{\Omega}$, $R_1 \cap R_2 = \emptyset$

Classification rule (Region R_1 and R_2): minimizes the chances of making misclassification error

Two populations

Important concepts

1. Prior

   p_1 = prior probability of π_1
   p_2 = prior probability of π_2

2. Misclassification

   $P(2|1) =$ Conditional probability of misclassification of an object as π_2 when, in fact, it is from π_1
   $P(1|2) =$ Conditional probability of misclassification of an object as π_1 when, in fact, it is from π_2

3. Cost

   Cost matrix

True Population	**Classify as**	
	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

Two populations

Expected cost of misclassification (ECM)

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

Minimum Expected Cost of Misclassification

The Regions R_1 and R_2 that minimize the ECM are defined by the values $\boldsymbol{x}$ for which the following inequalities hold.

R_1: $$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

R_2: $$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Two populations

<u>Special Cases of Minimum Expected Cost of Misclassification Rule</u>

1. Equal priors: $p_1 = p_2$

$$R_1:\quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right); \qquad R_2:\quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)$$

2. Equal misclassification costs: $c(1|2) = c(2|1)$

$$R_1:\quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{p_2}{p_1}\right); \qquad R_2:\quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{p_2}{p_1}\right)$$

3. Equal priors and equal misclassification costs : $p_1 = p_2$, $c(1|2) = c(2|1)$

$$R_1:\quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1; \qquad R_2:\quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < 1$$

Two populations

<u>TMP rule: Minimize the total probability of misclassification</u>

TMP $= P$(misclassifying a π_1 observation or misclassifying a π_2 observation)
$= P$(observation from π_1 and is misclassified) + P(observation from π_2 and is misclassified)
$= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$

Allocate observations in order to minimize TPM. Mathematically, equivalent to minimizing the expected cost of misclassification when the costs of misclassification are equal.

Two populations (Normal populations: $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$)

Equal population variances $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

Minimum ECM rule:

Allocate a new observation $\boldsymbol{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 - \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\boldsymbol{x}_0$ to π_2 otherwise.

Two populations (Normal populations: $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$)

Equal population variances $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

Estimated Minimum ECM rule:

Allocate a new observation $\boldsymbol{x}_0$ to π_1 if

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}_0 - \frac{\mathbf{1}}{\mathbf{2}} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \boldsymbol{S}_{pooled}^{-1} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\boldsymbol{x}_0$ to π_2 otherwise.

Two populations (Normal populations: $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$)

Unequal population variances $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Minimum ECM rule:

Let $k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$

Allocate a new observation $\boldsymbol{x}_0$ to π_1 if

$$-\frac{1}{2}\boldsymbol{x}_0'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0 + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\boldsymbol{x}_0$ to π_2 otherwise.

Estimated Minimum ECM rule: same as above except the unknown parameters are replaced by sample estimates. Hence, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_2$ are replaced by $\bar{\boldsymbol{x}}_1$, $\boldsymbol{S}_1$, $\bar{\boldsymbol{x}}_2, \boldsymbol{S}_2$.