

**Lecture 9: Discrimination and Classification**

## 1 Basic concept

*Discrimination* is concerned with separating distinct sets of observations whereas *classification* is to allocate new objects to previously defined groups. The goals of these two multivariate techniques are

1. Discrimination: To describe the differential features of objects (observations) from several known collections (populations). To find *discriminants* whose numerical values are such that the collections are separated as much as possible.
2. Classification: To sort objects (observations) into two or more labeled classes. To derive a rule that can be used to optimally assign *new* objects to the labeled classes.

## 2 The case of two populations

Denote the two populations by  $\pi_1$  and  $\pi_2$ , respectively. Denote by  $\mathbf{X} = (X_1, \dots, X_p)'$ , the  $p$ -dimensional random vector of measurements from the object of the populations. Let  $f_i(\mathbf{x})$  be the probability density function of  $\mathbf{X}$  under the  $i$ th population.

An object with associated measurements  $\mathbf{x}$  must be assigned to either  $\pi_1$  or  $\pi_2$ . Let  $\Omega$  be the sample space, i.e. the collection of all possible values of  $\mathbf{x}$ . Let  $R_1$  be the set of  $\mathbf{x}$  values for which we classify objects as  $\pi_1$  and  $R_2 = \Omega - R_1$  be the remaining  $\mathbf{x}$  values for which we classify objects as  $\pi_2$ . Since every object must be assigned to one and only one of the two populations, the sets  $R_1$  and  $R_2$  are mutually exclusive and exhaustive.

The conditional probability,  $P(2|1)$ , of classifying an object as  $\pi_2$  when, in fact, it is from  $\pi_1$  is

$$P(2|1) = P(\mathbf{X} \in R_2 | \pi_1) = \int_{R_2} f_1(\mathbf{x}) d\mathbf{x}.$$

Similarly, the conditional probability,  $P(1|2)$ , of classifying an object as  $\pi_1$  when it is really from  $\pi_2$  is

$$P(1|2) = P(\mathbf{X} \in R_1 | \pi_2) = \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}.$$

Suppose that the prior distribution for the populations is  $P(\pi_i) = p_i$ ,  $i = 1$  and  $2$ , where  $p_1 + p_2 = 1$ . Then, we have

- P(obs correctly classified as  $\pi_1$ ) = P(obs comes from  $\pi_1$  and is correctly classified as  $\pi_1$ ) =  $P(\mathbf{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1$
- P(obs incorrectly classified as  $\pi_1$ ) = P(obs comes from  $\pi_2$  and is incorrectly classified as  $\pi_1$ ) =  $P(\mathbf{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2$
- P(obs correctly classified as  $\pi_2$ ) = P(obs comes from  $\pi_2$  and is correctly classified as  $\pi_2$ ) =  $P(\mathbf{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2$
- P(obs incorrectly classified as  $\pi_2$ ) = P(obs comes from  $\pi_1$  and is incorrectly classified as  $\pi_2$ ) =  $P(\mathbf{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1$

The cost of misclassification is defined by a cost matrix:

|                  |         |              |          |
|------------------|---------|--------------|----------|
|                  |         | Classify as: |          |
|                  |         | $\pi_1$      | $\pi_2$  |
| True population: | $\pi_1$ | 0            | $c(2 1)$ |
|                  | $\pi_2$ | $c(1 2)$     | 0        |

Thus, the costs are (1) zero for correct classification, (2)  $c(1|2)$  when an observation from  $\pi_2$  is incorrectly classified as  $\pi_1$ , and (3)  $c(2|1)$  when a  $\pi_1$  observation is incorrectly classified as  $\pi_2$ .

The *expected cost of misclassification* is

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2.$$

A reasonable classification rule should have an ECM as small as possible.

**Result 11.1.** The regions  $R_1$  and  $R_2$  that minimize the ECM are defined by the values  $\mathbf{x}$  for which the following inequalities hold:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}.$$

Thus,  $R_1$  consists of  $\mathbf{x}$  such that

$$\left( \begin{array}{c} \text{density} \\ \text{ratio} \end{array} \right) \geq \left( \begin{array}{c} \text{cost} \\ \text{ratio} \end{array} \right) \left( \begin{array}{c} \text{prior prob.} \\ \text{ratio} \end{array} \right).$$

**Proof:** See the hint at Exercise 11.3.

**Another criterion:** Total probability of misclassification (TPM)

$$TPM = p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

This criterion is the same as the ECM if the costs of misclassification are the same.

Finally, a new observation  $\mathbf{x}_o$  can be allocated to the population with the largest posterior probability  $P(\pi_i|\mathbf{x}_o)$ . Specifically, by Bayes's rule, the posterior probabilities are

$$\begin{aligned} P(\pi_1|\mathbf{x}_o) &= \frac{P(\pi_1 \text{ occurs and we observe } \mathbf{x}_o)}{P(\text{we observe } \mathbf{x}_o)} \\ &= \frac{P(\text{we observe } \mathbf{x}_o|\pi_1)P(\pi_1)}{P(\text{we observe } \mathbf{x}_o|\pi_1)P(\pi_1) + P(\text{we observe } \mathbf{x}_o|\pi_2)P(\pi_2)} \\ &= \frac{p_1 f_1(\mathbf{x}_o)}{p_1 f_1(\mathbf{x}_o) + p_2 f_2(\mathbf{x}_o)}. \\ P(\pi_2|\mathbf{x}_o) &= \frac{p_2 f_2(\mathbf{x}_o)}{p_1 f_1(\mathbf{x}_o) + p_2 f_2(\mathbf{x}_o)}. \end{aligned}$$

Classifying an observation  $\mathbf{x}_o$  as  $\pi_1$  if  $P(\pi_1|\mathbf{x}_o) > P(\pi_2|\mathbf{x}_o)$ .

### 3 Classification with two multivariate normal populations

**Case 1:** equal covariance matrices

**Result 11.2.** Let the populations  $\pi_i$  ( $i = 1$  and  $2$ ) be described by the density functions  $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ . Then, the allocation rule that minimizes the ECM is as follows:

Allocate  $\mathbf{x}_o$  to  $\pi_1$  if

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

Allocate  $\mathbf{x}_o$  to  $\pi_2$  otherwise.

**Proof:** Use the identity

$$\begin{aligned} -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} \\ &\quad - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \end{aligned}$$

Also, the minimum ECM regions are

$$\begin{aligned} R_1 &: \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \right] \geq \frac{c(1|2)p_2}{c(2|1)p_1}. \\ R_2 &: \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \right] < \frac{c(1|2)p_2}{c(2|1)p_1}. \end{aligned}$$

In practice,  $\Sigma$  is estimated by the pooled covariance matrix  $\mathbf{S}_{pool}$  and  $\boldsymbol{\mu}_i$  by sample mean of the population  $\pi_1$ . Consequently, the *sample* classification rule is as follows:

The estimated minimum ECM rule for two normal populations:

Allocate  $\mathbf{x}_o$  to  $\pi_1$  if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pool}^{-1} \mathbf{x}_o - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pool}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln \left[ \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right].$$

Allocate  $\mathbf{x}_o$  to  $\pi_2$  otherwise.

*Fisher's approach to classification with two populations:*

Fisher arrived at prior minimum ECM rule using an alternative argument. He considered linear combinations of the random vector so as to transform multivariate distributions into univariate ones. The transformation is chosen to achieve maximum separation of the transformed sample means. Let  $y_{1i} = \mathbf{a}' \mathbf{x}_{1i}$  and  $y_{2i} = \mathbf{a}' \mathbf{x}_{2i}$  be the transformed samples. Also, let  $\bar{y}_j$  be the sample mean of  $\{y_{ji}\}$ ,  $j = 1$  and  $2$ , and

$$s_y^2 = \frac{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2}.$$

The separation is measured by  $(\bar{y}_1 - \bar{y}_2)^2 / s_y^2$ .

**Result 11.3.** The linear transformation  $\hat{y} = \hat{\mathbf{a}}' \mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pool}^{-1} \mathbf{x}$  maximizes the ratio

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{\hat{\mathbf{a}}' (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\hat{\mathbf{a}}' \mathbf{S}_{pool} \hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}' \mathbf{d})^2}{\hat{\mathbf{a}}' \mathbf{S}_{pool} \hat{\mathbf{a}}}$$

over all possible coefficient vectors  $\hat{\mathbf{a}}$ , where  $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ . The maximum of the above ratio is  $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pool}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ .

**Proof:** Use the inequality in Eq. (2-50) of the textbook.

An allocation rule based on Fisher's discriminant function:

Allocate  $\mathbf{x}_o$  to  $\pi_1$  if

$$\hat{y}_o = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pool}^{-1} \mathbf{x}_o \geq \hat{m}$$

where  $\hat{m} = \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pool}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$ . Allocate  $\mathbf{x}_o$  to  $\pi_2$  otherwise.

This rule is known as the Fisher linear discriminant function.

**Case 2:** Unequal covariance matrices: Quadratic classification rule.

The Result 11.1 leads to the following classification regions when the covariance matrices are not equal.

$$R_1 : -\frac{1}{2} \mathbf{x}' (\Sigma_1^{-1} - \Sigma_2^{-1}) \mathbf{x} + (\boldsymbol{\mu}'_1 \Sigma_1^{-1} - \boldsymbol{\mu}'_2 \Sigma_2^{-1}) \mathbf{x} - k \geq \ln \left[ \frac{c(1|2)p_2}{c(2|1)p_1} \right],$$

$$R_2 : -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}'_1\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}'_2\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln \left[ \frac{c(1|2)p_2}{c(2|1)p_1} \right],$$

where

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

For a new observation  $\mathbf{x}_o$ , we may use sample statistics to obtain the following quadratic classification rule:

Allocate  $\mathbf{x}_o$  to  $\pi_1$  if

$$\frac{1}{2}\mathbf{x}'_o(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_o + (\bar{\mathbf{x}}'_1\mathbf{S}_1^{-1} - \bar{\mathbf{x}}'_2\mathbf{S}_2^{-1})\mathbf{x}_o - k \geq \ln \left[ \frac{c(1|2)p_2}{c(2|1)p_1} \right].$$

Allocate  $\mathbf{x}_o$  to  $\pi_2$  otherwise.

## 4 Evaluating classification functions

If the population distributions are known, the total probability of misclassification is

$$TPM = p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

The choices of  $R_1$  and  $R_2$  that minimizes TPM give rise to the *optimum error rate* (OER). When  $R_1$  and  $R_2$  are chosen based on Result 11.1 with  $c(1|2) = c(2|1)$ , we have OER. Thus, OER is the error rate determined by the minimum TPM classification rule.

In practice, the population densities are unknown, the actual error rate (AER) is

$$AER = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})d\mathbf{x},$$

where  $\hat{R}_i$  are the classification regions determined by the sample statistics.

A measure that does not depend on the form of populations is called *apparent error rate* (APER), which is defined as the fraction of observations in the *training* sample that are misclassified by the sample classification function.

|                   |         | Predicted membership    |                         |
|-------------------|---------|-------------------------|-------------------------|
|                   |         | $\pi_1$                 | $\pi_2$                 |
| Actual membership | $\pi_1$ | $n_{1c}$                | $n_{1m} = n_1 - n_{1c}$ |
|                   | $\pi_2$ | $n_{2m} = n_2 - n_{2c}$ | $n_{2c}$                |

where

$n_{1c}$  = number of  $\pi_1$  items correctly classified as  $\pi_1$  items

$n_{1m}$  = number of  $\pi_1$  items incorrectly classified as  $\pi_2$  items

$n_{2c}$  = number of  $\pi_2$  items correctly classified as  $\pi_2$  items

$n_{2m}$  = number of  $\pi_2$  items incorrectly classified as  $\pi_1$  items

The apparent error rate is then

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

**Example.** Example 11.8 about classification of Alaskan and Canadian salmon. Page 603 of the textbook.

## 5 More than two populations

### 5.1 The minimum expected cost of misclassification method

Let  $f_i(\mathbf{x})$  be the density function of the population  $\pi_i$ ,  $i = 1, \dots, g$ . Define

- $p_i$  = prior probability of population  $\pi_i$
- $c(k|i)$  = the cost of allocating an item to  $\pi_k$  when, in fact, it belongs to  $\pi_i$ , for  $k, i = 1, \dots, g$  (with  $c(i|i) = 0$ )
- $R_i$  = the set of  $\mathbf{x}$ 's classified as  $\pi_i$
- $P(k|i) = P(\text{classifying item as } \pi_k | \pi_i) = \int_{R_k} f_i(\mathbf{x}) d\mathbf{x}$ .

Note that  $\{R_i\}$  are mutually exclusive and exhaustive.

The cost of misclassifying an item  $\mathbf{x}$  from  $\pi_i$  is

$$ECM(i) = P(1|i)c(1|i) + P(2|i)c(2|i) + \dots + P(g|i)c(g|i) = \sum_{j=1}^g P(j|i)c(j|i),$$

where it is understood that  $c(i|i) = 0$  and  $P(i|i) = 1 - \sum_{j \neq i} P(j|i)$ .

The overall ECM is

$$\begin{aligned} ECM &= p_1 ECM(1) + p_2 ECM(2) + \dots + p_g ECM(g) \\ &= \sum_{i=1}^g p_i \left( \sum_{j=1, j \neq i}^g P(j|i)c(j|i) \right). \end{aligned}$$

It turns out that we have the following result.

**Result 11.5.** The classification regions that minimize the ECM are defined by allocating  $\mathbf{x}$  to that population  $\pi_k$  for which

$$\sum_{i=1, i \neq k}^g p_i f_i(\mathbf{x}) c(k|i)$$

is smallest. If a tie occurs,  $\mathbf{x}$  can be assigned to any of the tied populations. If the misclassification costs are the same, then the classification rule is:

Allocate  $\mathbf{x}$  to  $\pi_k$  if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}), \quad \text{for all } i \neq k$$

or, equivalently, Allocate  $\mathbf{x}$  to  $\pi_k$  if

$$\ln(p_k f_k(\mathbf{x})) > \ln(p_i f_i(\mathbf{x})), \quad \text{for all } i \neq k.$$

This is equivalent to the one that maximizes the posterior probability  $P(\pi_k|\mathbf{x})$ , which is

$$P(\pi_k|\mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{\sum_{i=1}^g p_i f_i(\mathbf{x})}.$$

**Normal populations:** Assume that  $f_i(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$  and the costs of misclassification are the same. Then, we have

$$\ln(p_i f_i(\mathbf{x})) = \ln(p_i) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln(|\boldsymbol{\Sigma}_i|) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i).$$

This leads to the definition of *quadratic discrimination score* as

$$d_i^Q(\mathbf{x}) = -\frac{1}{2} \ln(|\boldsymbol{\Sigma}_i|) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p_i)$$

for  $i = 1, \dots, g$ . In practice, replace the theoretical values by their sample estimates, we have the estimated minimum total probability misclassification rule for several normal populations as

Allocate  $\mathbf{x}$  to  $\pi_i$  if

$$\hat{d}_i^Q(\mathbf{x}) = \max\{\hat{d}_1^Q(\mathbf{x}), \dots, \hat{d}_g^Q(\mathbf{x})\},$$

where  $\hat{d}_j^Q(\mathbf{x})$  is the sample estimate of  $d_j^Q(\mathbf{x})$ .

When  $\boldsymbol{\Sigma}_i$  are equal, then the function  $d_i^Q(\mathbf{x})$  reduces to the linear discriminant score defined as

$$d_i(\mathbf{x}) = \boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln(p_i)$$

for  $i = 1, \dots, g$ .

**Example:** See Examples 11.11 and 11.12.