Topic 2: Linear discriminant analysis

Classification with two multivariate normal populations

Until now we did not specify any particular form of the densities $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$. Essential simplification occurs under normality assumption and we are going over to a more detailed discussion of this particular case now. Two different cases will be considered- of equal and of non-equal covariance matrices.

Case of equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$

Now we assume that the two populations π_1 and π_2 are $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$, respectively. Then, (5.4) becomes

$$R_1 = \left\{ \boldsymbol{x} : \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \right] \geq \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1} \right\}.$$

Similarly, from (5.5) we get

$$R_2 = \left\{ \boldsymbol{x} : \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \right] < \frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1} \right\},$$

and we arrive at the following result

Theorem 5.1. *Under the above assumptions, the allocation rule that minimises the ECM is given by:*

1. *Allocate $\boldsymbol{x}_0$ to π_1 if*

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

2. *Otherwise, allocate $\boldsymbol{x}_0$ to π_2.*

Note, also that it is unrealistic to assume in most situations that the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and Σ are known. They will need to be estimated by the data instead. Assume, n_1 and n_2 observations are available from the first and from the second population, respectively. If $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$ are the sample mean vectors and $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ the corresponding sample covariance matrices, then under the assumption of $\Sigma_1 = \Sigma_2 = \Sigma$ we can derive the pooled covariance matrix estimator $\boldsymbol{S}_{\text{pooled}} = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$ (This is an unbiased estimator of Σ (!)).

Hence the *Sample classification rule* becomes:

1. Allocate $\boldsymbol{x}_0$ to π_1 if

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geq \log\left[\frac{c(1|2)}{c(2|1)} \times \frac{p_2}{p_1}\right] \quad (5.6)$$

2. Otherwise, allocate $\boldsymbol{x}_0$ to π_2.

This empirical classification rule is called **an allocation rule based on Fisher's discriminant function**. The function

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)$$

itself (which is linear in the vector observation $\boldsymbol{x}_0$) is called **Fisher's linear discriminant function.**

Of course, the latter rule is only an *estimate* of the optimal rule since the parameters in the latter have been replaced by estimated quantities. But we are expecting this rule to perform well when n_1 and n_2 are large. It is to be pointed out that the allocation rule in (5.6) is **linear** in the new observation $\boldsymbol{x}_0$. The simplicity of its form is a consequence of the multivariate normality assumption.

Optimum error rate and Mahalanobis distance

We defined the TPM quantity in general terms for any classification in (5.3). When the regions R_1 and R_2 are selected in an optimal way, one obtains the minimal value of TPM which is called **optimum error rate (OER)** and is being used to characterise the difficulty of the classification problem at hand. Hereby we shall illustrate the calculation of the OER for the simple case of two normal populations with $\Sigma_1 = \Sigma_2 = \Sigma$ and prior probabilities $p_1 = p_2 = \frac{1}{2}$. In this case

$$\mathrm{TPM} = \frac{1}{2}\int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x},$$

and OER is obtained by choosing

$$R_1 = \left\{\boldsymbol{x} : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0\right\}$$

and

$$R_2 = \left\{\boldsymbol{x} : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0\right\}.$$

If we introduce the random variable $Y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{X} = \boldsymbol{l}^\top \boldsymbol{X}$ then $Y|i \sim N_1(\mu_{iY}, \Delta^2), i = 1, 2$ for the two populations π_1 and π_2 where $\mu_{iY} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\boldsymbol{\mu}_i, i = 1, 2$. The quantity $\Delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ is the **Mahalanobis distance** between the two normal populations and it has an important role in many applications of Multivariate Analysis. Now

$$\Pr(2|1) = \Pr\left(Y < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) = \Pr\left(\frac{Y - \mu_{1Y}}{\Delta} < -\frac{\Delta}{2}\right) = \Phi\left(-\frac{\Delta}{2}\right),$$

$\Phi(\cdot)$ denoting the cumulative distribution function of the standard normal. Along the same lines we can get (**do it (!)**) : $\Pr(1|2) = \Phi\left(-\frac{\Delta}{2}\right)$ to that finally $\mathrm{OER} = \mathrm{minim}\ \mathrm{TPM} = \Phi\left(-\frac{\Delta}{2}\right)$.

In practice, Δ is replaced by its estimated value $\hat{\Delta} = \sqrt{\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\mathrm{pooled}}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}$.

Classification with more than 2 normal populations

Formal generalisation of the theory for the case of $g > 2$ groups $\pi_1, \pi_2, \ldots, \pi_g$ is straightforward but optimal error rate analysis is difficult when $g > 2$. It is easy to see that the ECM classification rule with **equal** misclassification costs (compare to (5.4) and (5.5)) becomes now:

1. Allocate $\boldsymbol{x}_0$ to π_k if $p_k \boldsymbol{f}_k > p_i \boldsymbol{f}_i$ for all $i \neq k$.

Equivalently, one can check if $\log p_k \boldsymbol{f}_k > \log p_i \boldsymbol{f}_i$ for all $i \neq k$.

When applying this classification rule to g normal populations $\boldsymbol{f}_i(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}_i, \Sigma_i), i = 1, 2, \ldots, g$ it becomes:

1. Allocate $\boldsymbol{x}_0$ to π_k if

$$\begin{aligned} \log p_k \boldsymbol{f}_k(\boldsymbol{x}_0) &= \log p_k - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_k) \\ &= \max_i \log p_i \boldsymbol{f}_i(\boldsymbol{x}_0). \end{aligned}$$

Ignoring the constant $\frac{p}{2}\log(2\pi)$ we get the **quadratic discriminant score for the ith population:**

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \log p_i \tag{5.7}$$

and the rule advocates to allocate $\boldsymbol{x}$ to the population with a largest quadratic discriminant score. It is obvious how one would estimate from the data the unknown quantities involved in (5.7) in order to obtain the *estimated* minimum total probability of misclassification rule. (You should formulate the precise statement (!)).

In the case we are justified to assume that **all covariance matrices** for the g populations are equal, a simplification is possible (like in the case $g = 2$). Looking only at the terms that vary with $i = 1, 2, \ldots, g$ in (5.7) we can define the **linear discriminant score:** $d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \Sigma^{-1}\boldsymbol{\mu}_i + \log p_i$. Correspondingly, a **sample version** of the linear discriminant score is obtained by substituting the arithmetic means $\bar{\boldsymbol{x}}_i$ instead of $\boldsymbol{\mu}_i$ and $\boldsymbol{S}\text{pooled} = \frac{n_1 - 1}{n_1 + n_2 + \ldots n_g - g}\boldsymbol{S}_1 + \cdots + \frac{n_g - 1}{n_1 + n_2 + \cdots + n_g - g}\boldsymbol{S}_g$ instead of Σ thus arriving at

$$\hat{d}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{\text{pooled}}^{-1}\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{\text{pooled}}^{-1}\bar{\boldsymbol{x}}_i + \log p_i$$

Therefore the **Estimated Minimum TPM Rule for Equal Covariance Normal Populations** is the following:

1. Allocate $\boldsymbol{x}$ to π_k if $\hat{d}_k(\boldsymbol{x})$ is the largest of the g values $\hat{d}_i(\boldsymbol{x}), i = 1, 2, \ldots, g$.

In this form, the classification rule has been implemented in many computer packages.

Check your understanding

Complete the below exercises to check your understanding of concepts presented so far.

Three bivariate normal populations, labelled $i = 1, 2, 3$ have same covariance matrix given by $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and means $\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ respectively.

a) Suggest a classification rule for an observation $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ that corresponds to one of the three populations. You may assume equal priors for the three populations and equal misclassification costs.

b) Classify the following observations to one of the three distributions:

$$\begin{pmatrix} 0.2 \\ 0.6 \end{pmatrix}, \begin{pmatrix} 2 \\ 0.8 \end{pmatrix}, \begin{pmatrix} 0.75 \\ 1 \end{pmatrix}$$

c) Show that in $\mathbb{R}^2$, the 3 classification regions are bounded by straight lines and draw a graph of these three regions.