VI Lecture

Discrimination and Classification

6.1. Introduction. Separation and Classification for two populations

Discriminant analysis and classification are widely used multivariate techniques. The goal is either *separating sets of objects* (in discriminant analysis terminology) or *allocating new objects to given groups* (in classification theory terminology).

Basically, discriminant analysis is more exploratory in nature than classification. However, the difference is not significant especially because very often a function that separates may sometimes serve as an allocator, and, conversely, a rule of allocation may suggest a discriminatory procedure. In practice, the goals in the two procedures often overlap.

We will consider the case of two populations (classes of objects) first.

Typical examples include: an anthropologist wants to classify a skull as a male or female; a patient needs to be classified as needing surgery or not needing surgery etc.

Denote the two classes by π_1 and π_2. The separation is to be performed on the basis of measurements of p associated random variables that form a vector $\mathbf{X} \in \mathbf{R^p}$. The observed values of $\mathbf{X}$ belong to different distributions when taken from π_1 and π_2 and we shall denote the densities of these two distributions by $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, respectively.

Allocation or classification is possible due to the fact that one has a *learning sample* at hand, i.e. there are some measurement vectors that are known to have been generated from each of the two populations. These measurements have been generated in earlier similar experiments. The goal is to partition the sample space into 2 mutually exclusive regions, say R_1 and R_2, such that if a *new* observation falls in R_1, it is allocated to π_1 and if it falls in R_2, it is allocated to π_2.

6.2. Classification errors.

There is always a chance of an erroneous classification (misclassification). Our goal will be to develop such classification methods that in a suitably defined sense minimize the chances of misclassification.

It should be noted that one of the two classes may have a greater likelihood of occurrence because one of the two populations might be much larger than the other. For example, there tend to be a lot more financially sound companies than bankrupt companies. These *prior probabilities* of occurrence should also be taken into account when constructing an optimal classification rule if we want to perform optimally.

In a more detailed study of optimal classification rules, cost is also important. If classifying a π_1 object to the class π_2 represents a much more serious error than classifying a π_2 object to the class π_1 then these cost differences should also be taken into account when designing the optimal rule.

The **conditional** probabilities for misclassification are defined naturally as:

$$P(2|1) = P(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x} \tag{6.1}$$

$$P(1|2) = P(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \tag{6.2}$$

6.3. Optimal classification rules

i) rules that minimize the expected cost of misclassification (ECM)

Denote by p_i the **prior** probability of $\pi_i, i = 1, 2 (p_1 + p_2 = 1)$. Then the **overall** probabilities of incorrectly classifying objects will be: $P(\text{misclassified as } \pi_1) = P(1|2)p_2$ and $P(\text{misclassified as } \pi_2) = P(2|1)p_1$. Further, let $c(i|j), i \neq j, i, j = 1, 2$ be the misclassification costs. Then the **expected cost of misclassification** is

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \tag{6.3}$$

Lemma 6.3.1. The regions R_1 and R_2 that minimize ECM are given by

$$R_1 = \{\mathbf{x} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\} \tag{6.4}$$

and

$$R_2 = \{\mathbf{x} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\} \tag{6.5}$$

Proof. It is easy to see that $ECM = \int_{R_1} [c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})]d\mathbf{x} + c(2|1)p_1$. Hence, the ECM will be minimized if R_1 includes those values of $\mathbf{x}$ for which the integrand $[c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})] \leq 0$ and excludes all the complementary values.

Note the significance of the fact that in Lemma 6.3.1 only ratios are involved. Often in practice, one would have a much clearer idea about the cost ratio rather than for the actual costs themselves.

For your own exercise, consider the partial cases of Lemma 6.3.1 when $p_2 = p_1$, $c(1|2) = c(2|1)$ and when both these equalities hold. Comment on the soundness of the classification regions in these cases.

ii) rules that minimize the total probability of misclassification (TPM)

If we ignore the cost of misspecification, we can define the total probability of misspecification as

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

Mathematically, this is a particular case of i) when the costs of misclassification are equal-so nothing new here.

iii) Bayesian approach

Here we try to allocate a new observation $\mathbf{x_0}$ to the population with the larger posterior probability $P(\pi_i|\mathbf{x_0}), i = 1, 2$. According to Bayes formula we have

$$P(\pi_1|\mathbf{x_0}) = \frac{p_1 f_1(\mathbf{x_0})}{p_1 f_1(\mathbf{x_0}) + p_2 f_2(\mathbf{x_0})}, P(\pi_2|\mathbf{x_0}) = \frac{p_2 f_2(\mathbf{x_0})}{p_1 f_1(\mathbf{x_0}) + p_2 f_2(\mathbf{x_0})}$$

Mathematically, the strategy of classifying an observation $\mathbf{x_0}$ as π_1 if $P(\pi_1|\mathbf{x_0}) > P(\pi_2|\mathbf{x_0})$ is again a particular case of i) when the costs of misclassification are equal. (**Why ?**) But note that the calculation of the posterior probabilities themselves is in itself a useful and informative operation.

6.4. Classification with two multivariate normal populations

Until now we did not specify any particular form of the densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$. 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.

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

Now we assume that the two populations π_1 and π_2 are $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$, correspondingly.

Then (6.4) becomes

$$R_1 = \{\mathbf{x} : exp[-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)] \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\}$$

Similarly, from (6.5) we get

$$R_2 = \{\mathbf{x} : exp[-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1) + \frac{1}{2}(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)] < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\}$$

and we arrive at the following result:

Theorem 6.4.1. Under the assumptions in 6.4.1, the allocation rule that minimizes the ECM is given by:

allocate $\mathbf{x_0}$ to π_1 if

$$(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x_0} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \log[\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}]$$

Otherwise, allocate $\mathbf{x_0}$ to π_2.

Proof. Simple exercise (to be discussed at lectures).

Note also that it is unrealistic to assume in most situations that the parameters μ_1, μ_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 $\mathbf{\bar{x}_1}$ and $\mathbf{\bar{x}_2}$ are the sample mean vectors and $\mathbf{S_1}, \mathbf{S_2}$ the corresponding population covariance matrices then under the assumption of $\Sigma_1 = \Sigma_2 = \Sigma$ we can derive the pooled covariance matrix estimator $\mathbf{S_{pooled}} = \frac{\mathbf{(n_1-1)S_1+(n_2-1)S_2}}{\mathbf{n_1+n_2-2}}$ (This is an unbiased estimator of Σ (!)).

Hence the *sample classification rule* becomes: allocate $\mathbf{x_0}$ to π_1 if

$$(\mathbf{\bar{x}_1} - \mathbf{\bar{x}_2})'\mathbf{S_{pooled}^{-1}}\mathbf{x_0} - \frac{1}{2}(\mathbf{\bar{x}_1} - \mathbf{\bar{x}_2})'\mathbf{S_{pooled}^{-1}}(\mathbf{\bar{x}_1} + \mathbf{\bar{x}_2}) \geq \log[\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}] \qquad (6.6)$$

Otherwise, allocate $\mathbf{x_0}$ to π_2. This empirical classification rule is called **an allocation rule based on Fisher's discriminant function** The function

$$(\mathbf{\bar{x}_1} - \mathbf{\bar{x}_2})'\mathbf{S_{pooled}^{-1}}\mathbf{x_0} - \frac{1}{2}(\mathbf{\bar{x}_1} - \mathbf{\bar{x}_2})'\mathbf{S_{pooled}^{-1}}(\mathbf{\bar{x}_1} + \mathbf{\bar{x}_2})$$

itself (which is linear in the vector observation $\mathbf{x_0}$) 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 (6.6) is **linear** in the new observation $\mathbf{x_0}$. The simplicity of its form is a consequence of the multivariate normality assumption.

6.4.2. Case of different covariance matrices ($\Sigma_1 \neq \Sigma_2$)

Now we assume that the two populations π_1 and π_2 are $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$, correspondingly. Repeating the same steps like in 6.4.1 we get

$$R_1 = \{\mathbf{x} : -\frac{1}{2}(\mathbf{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\mathbf{x} - k \geq \log[\frac{c(1|2)}{c(2|1)}.\frac{p_2}{p_1}]\}$$

$$R_2 = \{\mathbf{x} : -\frac{1}{2}(\mathbf{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\mathbf{x} - k < \log[\frac{c(1|2)}{c(2|1)}.\frac{p_2}{p_1}]\}$$

where $k = \frac{1}{2}\log(\frac{|\Sigma_1|}{|\Sigma_2|}) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)$ and we see that the classification regions are **quadratic** functions of the new observation in this case. One obtains the following rule:

allocate $\mathbf{x_0}$ to π_1 if

$$-\frac{1}{2}\mathbf{x_0}'(S_1^{-1} - S_2^{-1})\mathbf{x_0} + (\bar{x}_1'S_1^{-1} - \bar{x}_2'S_2^{-1})\mathbf{x_0} - \hat{k} \geq \log[\frac{c(1|2)}{c(2|1)}.\frac{p_2}{p_1}]$$

where $\hat{k}$ is the empirical analog of k. Allocate $\mathbf{x_0}$ to π_2 otherwise.

When $\Sigma_1 = \Sigma_2$ the quadratic term disappears and we can easily see that the classification regions from 6.4.1 are obtained. Of course, the case considered in 6.4.2 is more general but we should be cautious when applying it in practice. It turns out that in more than two dimensions, classification rules based on quadratic functions do not always perform nicely and can lead to strange results. This is especially true when the data are not quite normal and when the differences in the covariance matrices are significant. The rule is very sensitive (non-robust) towards departures from normality. Therefore, it is advisable to try to first transform the data to more nearly normal by using some classical normality transformations. A detailed discussion of these effects will be provided during the lecture.

6.4.3. Optimum error rate and Mahalanobis distance

We defined the TPM quantity in general terms for any classification rule (see 6.3). 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 characterize 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

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

and OER is obtained by choosing

$$R_1 = \{\mathbf{x} : (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq 0\}$$

$$R_2 = \{\mathbf{x} : (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < 0\}$$

If we introduce the random variable $Y = (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x} = l'\mathbf{x}$ then $Y \sim N_1(\mu_{iY}, \Delta^2), i = 1, 2$ for the two populations π_1 and π_2 where $\mu_{iY} = (\mu_1 - \mu_2)'\Sigma^{-1}\mu_i, i = 1, 2$. The quantity

$\Delta = \sqrt{(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)}$ is the **Mahalanobis distance** between the two normal populations and it has an important role in many applications of Multivariate Analysis. Now

$$P(2|1) = P(Y < \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)) = P(\frac{Y - \mu_{1Y}}{\Delta} < -\frac{\Delta}{2}) = \Phi(-\frac{\Delta}{2})$$

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

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

6.5. Classification with more than 2 normal populations.

Formal generalization of the theory for the case of $g > 2$ groups $\pi_1, \pi_2, \dots, \pi_g$ 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 becomes (compare to (6.4) and (6.5)) now:

Allocate $\mathbf{x_0}$ to π_k if $p_k f_k > p_i f_i$ for all $i \neq k$. Equivalently, one can check if $\log p_k f_k > \log p_i f_i$ for all $i \neq k$.

When applying this classification rule to g normal populations $f_i(\mathbf{x}) \sim N_p(\mu_i, \Sigma_i), i = 1, 2, \dots, g$ it becomes:

Allocate $\mathbf{x_0}$ to π_k if

$$\log p_k f_k(x_0) = \log p_k - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_0 - \mu_k)'\Sigma_k^{-1}(x_0 - \mu_k) = \max_i \log p_i f_i(x_0)$$

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

$$d_i^Q(x) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i) + \log p_i \tag{6.7}$$

and the rule advocates to allocate $\mathbf{x}$ to the population with a largest quadratic discriminant score. It is obvious how one would estimate from the data the unknown quantities involved in (6.7) in order to obtain the *estimated* minimum total probability of misclassification rule. (You 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 $i = 1, 2, \dots, g$ in (6.7) we can define the **linear discriminant score**: $d_i(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \log p_i$. Correspondingly, a **sample version** of the linear discriminant score is obtained by substituting the arithmetic means $\bar{x}_i$ instead of μ_i and $\mathbf{S_{pooled}} = \mathbf{\frac{n_1 - 1}{n_1 + n_2 + \dots n_g - g}S_1} + \dots + \mathbf{\frac{n_g - 1}{n_1 + n_2 + \dots n_g - g}S_g}$ instead of $\mathbf{\Sigma}$ thus arriving at

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

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

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

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