DISCRIMINANT ANALYSIS

1. Introduction

Discrimination and classification are concerned with **separating** objects from different populations into different groups and with **allocating** new observations to one of these groups. The goals thus are

- To describe (graphically or algebraically) the difference between objects from several known populations. We construct "discriminants" that have numerical values which separate the different collections as much as possible.

- To assign objects into several labeled classes. We derive a "classification" rule that can be used to assign (new) objects to one of the labeled classes.

Examples

1. Based on historical bank data, separate the good from poor credit risks (based on income, age, family size, etc). Classify new credit applications into one of these two classes to decide to allow or reject a loan.

2. Make a distinction between readers and non-readers of a magazine or newspaper based on e.g. education level, age, income, profession, etc... such that the publishers knows which category of people are potential new readers.

A good procedure should result in as few misclassifications as possible. It should take into account the likelihood of objects to belong to each of the classes (=prior probability of occurrence). One often also takes into account the costs of misclassification. For example the cost of not operating a person needing surgery is much higher than unnecessarily operating a person, so the first type a misclassification has to be avoided as much as possible.

2. Discrimination and Classification of Two Populations

We now focus on separating objects from two classes and assigning new objects to one of these two classes. The classes will be labeled π_1 and π_2. Each object consists of measurements for p random variables $X_1, \ldots, X_p$ such that the observed values differ to some extend from one class to the other. The distributions associated with both populations will be described by their density functions f_1 and f_2 respectively.

Now consider an observed value $x = (x_1, \ldots, x_p)^\tau$ of the random variable $X = (X_1, \ldots, X_p)^\tau$. Then

$$\begin{cases} f_1(x) \text{ is the density in } x & \text{if } x \text{ belongs to population } \pi_1 \\ f_2(x) \text{ is the density in } x & \text{if } x \text{ belongs to population } \pi_2 \end{cases}$$

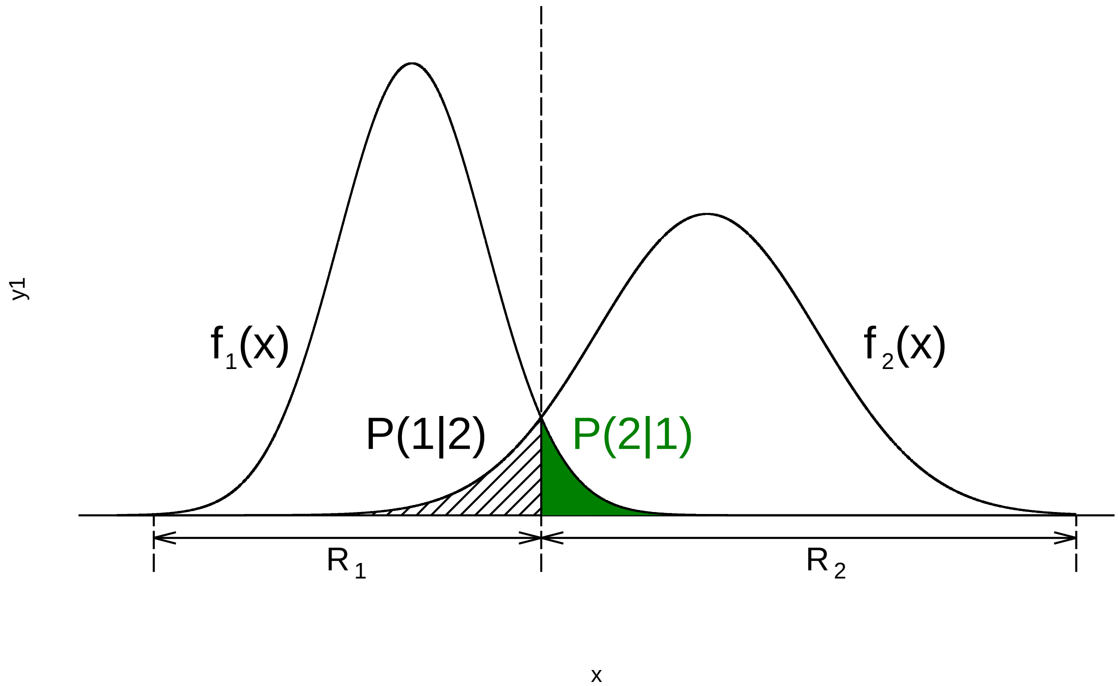
The object x must be assigned to either population π_1 or π_2. Denote Ω the sample space (= collection of all possible outcomes of X) and partition the sample space as $\Omega = R_1 \cup R_2$ where R_1 is the subspace of outcomes which we classify as belonging to population π_1 and $R_2 = \Omega - R_1$ the subspace of outcomes classified as belonging to π_2.

It follows that the (conditional) probability of classifying an object as belonging to π_2 when it is really from π_1 equals

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

and the (conditional) probability of assigning an object to π_1 when it in fact is from π_2 equals

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



Similarly, we define the conditional probabilities $P(1|1)$ and $P(2|2)$.

To obtain the probabilities of correctly and incorrectly classifying objects we also have to take the prior class probabilities into account. Denote

$$\begin{cases} p_1 = P(X \in \pi_1) = & \text{prior probability of } \pi_1 \\ p_2 = P(X \in \pi_2) = & \text{prior probability of } \pi_2 \end{cases}$$

where $p_1 + p_2 = 1$. It follows that the overall probabilities of correctly and incorrectly classifying objects are given by

$$
\begin{aligned}
P(\text{object is correctly classified as } \pi_1) &= P(X \in \pi_1 \text{ and } X \in R_1) \\
&= P(X \in R_1 | X \in \pi_1) P(X \in \pi_1) \\
&= P(1|1) p_1 \\
P(\text{object is misclassified as } \pi_1) &= P(X \in \pi_2 \text{ and } X \in R_1) \\
&= P(X \in R_1 | X \in \pi_2) P(X \in \pi_2) \\
&= P(1|2) p_2 \\
P(\text{object is correctly classified as } \pi_2) &= P(X \in \pi_2 \text{ and } X \in R_2) \\
&= P(X \in R_2 | X \in \pi_2) P(X \in \pi_2) \\
&= P(2|2) p_2 \\
P(\text{object is misclassified as } \pi_2) &= P(X \in \pi_1 \text{ and } X \in R_2) \\
&= P(X \in R_2 | X \in \pi_1) P(X \in \pi_1) \\
&= P(2|1) p_1
\end{aligned}
$$

To consider the cost of misclassification, denote

$$
\begin{cases}
c(2|1) = & \text{the cost of classifying an object from } \pi_1 \text{ as } \pi_2 \\
c(1|2) = & \text{the cost of classifying an object from } \pi_2 \text{ as } \pi_1
\end{cases}
$$

A classification rule is obtained by minimizing the **expected cost of misclassification**:

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

Result 1. The regions R_1 and R_2 that minimize the ECM are given by

$$R_1 = \left\{ x \in \Omega; \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1|2)}{c(2|1)}\right) \left(\frac{p_2}{p_1}\right) \right\}$$
$$R_2 = \left\{ x \in \Omega; \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1|2)}{c(2|1)}\right) \left(\frac{p_2}{p_1}\right) \right\}$$

Proof. Using that $P(1|1) + P(2|1) = 1$ (since $R_1 \cup R_2 = \Omega$) we obtain

$$\begin{aligned} \text{ECM} &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\ &= c(2|1)(1 - P(1|1))p_1 + c(1|2)P(1|2)p_2 \\ &= c(2|1)p_1 + \int_{R_1} [c(1|2)f_2(x)p_2 - c(2|1)f_1(x)p_1]d\,x \end{aligned}$$

Since probabilities and densities, as well as misclassification costs (there is no gain by misclassifying objects) are nonnegative, the ECM is minimal if the integrand $[c(1|2)f_2(x)p_2 - c(2|1)f_1(x)p_1] \leq 0$ for all $x \in R_1$ which yields the regions above. □

Note that these regions only depend on ratios:

- $\frac{f_1(x)}{f_2(x)}$ = density ratio
- $\frac{c(1|2)}{c(2|1)}$ = cost ratio
- $\frac{p_2}{p_1}$ = prior probability ratio

These ratios are often much easier to determine than the exact values of the components.

Special Cases:

1. Equal (or unknown) prior probabilities: Compare density ratio with cost ratio
$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}$$
2. Equal (or undetermined) misclassification cost: Compare density ratio with prior probability ratio:
$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \qquad R_2: \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$$
3. Equal prior probabilities and equal misclassification cost (or $\frac{p_2}{p_1} = (\frac{c(1|2)}{c(2|1)})^{-1}$)
$$R_1: \frac{f_1(x)}{f_2(x)} \geq 1 \qquad R_2: \frac{f_1(x)}{f_2(x)} < 1$$

Example 1. If we set the cost ratio equal to 2 and we know that 20% of all objects belong to π_2, then given that $f_1(x_0) = 0.3$ and $f_2(x_0) = 0.4$, do we classify x_0 as belonging to π_1 or π_2?

We have that $p_2 = 0.2$ so $p_1 = 0.8$, and $p_2/p_1 = 0.25$. Therefore, we obtain
$$R_1: \frac{f_1(x)}{f_2(x)} \geq 2(0.25) = 0.5 \quad \text{and} \quad R_2: \frac{f_1(x)}{f_2(x)} < 2(0.25) = 0.5$$
For x_0 we have
$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{0.3}{0.4} = 0.75 > 0.5$$
so we find $x_0 \in R_1$ and classify x_0 as belonging to π_1.

3. Classification with Two Multivariate Normal Populations

We now assume that f_1 and f_2 are multivariate normal densities with respectively mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2.

3.1. $\boldsymbol{\Sigma_1 = \Sigma_2 = \Sigma}$

The density of population π_i $(i = 1, 2)$ is now given by

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^\tau \Sigma^{-1}(x-\mu_i)}.$$

Result 2. If the populations π_1 and π_2 both have multivariate normal densities with equal covariance matrices, then the classification rule corresponding to minimizing ECM becomes:
Classify x_0 as π_1 if

$$(\mu_1 - \mu_2)^\tau \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)^\tau \Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

and classify x_0 as π_2 otherwise.

Proof. We assign x_0 to π_1 if

$$\frac{f_1(x_0)}{f_2(x_0)} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

which can be rewritten as

$$e^{-\frac{1}{2}(x_0-\mu_1)^\tau \Sigma^{-1}(x_0-\mu_1) + \frac{1}{2}(x_0-\mu_2)^\tau \Sigma^{-1}(x_0-\mu_2)} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

By taking logarithms on both sides and using the equality

$$-\frac{1}{2}(x_0-\mu_1)^\tau\Sigma^{-1}(x_0-\mu_1)+\frac{1}{2}(x_0-\mu_2)^\tau\Sigma^{-1}(x_0-\mu_2)=$$
$$(\mu_1-\mu_2)^\tau\Sigma^{-1}x_0-\frac{1}{2}(\mu_1-\mu_2)^\tau\Sigma^{-1}(\mu_1+\mu_2)$$

we obtain the classification rule. □

In practice, the population parameters μ_1, μ_2 and Σ are unknown and have to be estimated from the data. Suppose we have n_1 objects belonging to π_1 (denoted as $x_1^{(1)},\dots,x_{n_1}^{(1)}$) and n_2 objects from π_2 (denoted as $x_1^{(2)},\dots,x_{n_2}^{(2)}$) with $n_1+n_2=n$ the total sample size.

The sample mean vectors and covariance matrices of both groups are given by

$$\bar{x}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x_j^{(1)} \qquad S_1=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_j^{(1)}-\bar{x}_1)(x_j^{(1)}-\bar{x}_1)^\tau$$
$$\bar{x}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}x_j^{(2)} \qquad S_2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_j^{(2)}-\bar{x}_2)(x_j^{(2)}-\bar{x}_2)^\tau$$

Since both populations have the same covariance matrix Σ we combine the two sample covariance matrices S_1 and S_2 to obtain a more precise estimate of Σ given by

$$S_{\text{pooled}}=\left(\frac{n_1-1}{(n_1-1)+(n_2-1)}\right)S_1+\left(\frac{n_2-1}{(n_1-1)+(n_2-1)}\right)S_2$$

By replacing μ_1, μ_2 and Σ with $\bar{x}_1$, $\bar{x}_2$ and S_{pooled} in Result 2 we obtain the sample classification rule:

Classify x_0 as π_1 if

$$(\bar{x}_1-\bar{x}_2)^\tau S_{\text{pooled}}^{-1}x_0-\frac{1}{2}(\bar{x}_1-\bar{x}_2)^\tau S_{\text{pooled}}^{-1}(\bar{x}_1+\bar{x}_2)\geq\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

and classify x_0 as π_2 otherwise.

Special Case: Equal prior probabilities and equal misclassification cost:

$$\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] = \ln(1) = 0$$

such that we assign x_0 to π_1 if

$$(\bar{x}_1 - \bar{x}_2)^\tau S_{\text{pooled}}^{-1} x_0 \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^\tau S_{\text{pooled}}^{-1}(\bar{x}_1 + \bar{x}_2)$$

Denote $a = S_{\text{pooled}}^{-1}(\bar{x}_1 - \bar{x}_2) \in I\!R^p$, then this can be rewritten as

$$a^\tau x_0 \geq \frac{1}{2}(a^\tau \bar{x}_1 + a^\tau \bar{x}_2)$$

That is, we have to compare the scalar $\hat{y}_0 = a^\tau x_0$ with the midpoint $\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(a^\tau \bar{x}_1 + a^\tau \bar{x}_2)$. We thus have created to univariate populations (determined by the y-values) by projecting the original data on the direction determined by a. This direction is the (estimated) direction in which the two populations are best separated.

Remark. By replacing the unknown parameters with their estimates, there is no guarantee anymore that the resulting classification rule minimizes the expected cost of misclassification. However, we expect that we obtain a good estimate of the optimal rule.

Example 2. To develop a test for potential hemophilia carriers, blood samples were taken from two groups of patients. The two variables measured are AHF activity and AHF-like antigen where AHF means AntiHemophilic Factor. For both variables we take the logarithm (base 10). The first group of $n_1 = 30$ patients did not carry the hemophilia gene. The second group consisted of known hemophilia carriers. From these samples the following statistics have been derived

$$\bar{x}_1 = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix}, \quad \bar{x}_2 = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix}, \text{and} \quad S_{\text{pooled}} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix}.$$

Assuming equal costs and equal priors, we compute

$a = S_{\text{pooled}}^{-1}(\bar{x}_1 - \bar{x}_2) = \begin{pmatrix} 37.61 \\ -28.92 \end{pmatrix}$ and

$\bar{y}_1 = a^\tau \bar{x}_1 = 0.88$, $\bar{y}_2 = a^\tau \bar{x}_2 = -10.10$.

The corresponding midpoint thus becomes $\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = -4.61$. A new object $x = (x_1, x_2)^\tau$ is classified as non-carrier if $\hat{y} = 37.61x_1 - 28.92x_2 \geq \hat{m} = -4.61$ and is a carrier otherwise.

A potential hemophilia carrier has values $x_1 = -0.210$ and $x_2 = -0.044$. Should this patient be classified as carrier?

We obtain $\hat{y} = -6.62 < -4.61$ so we indeed assign this patient to the population of carriers.

Suppose now that it is known that the prior probability of being a hemophilia carrier is 25%, then a new patient is classified as non-carrier if $\hat{y} - \hat{m} \geq \ln\left(\frac{p_2}{p_1}\right)$.

We find $\hat{y} - \hat{m} = -6.62 + 4.61 = -2.01$ and

$\ln\left(\frac{p_2}{p_1}\right) = \ln\left(\frac{0.25}{0.75}\right) = \ln\left(\frac{1}{3}\right) = -1.10$

so we still classify this patient as carrier.

3.2. $\Sigma_1 \neq \Sigma_2$

The density of population π_i $(i = 1, 2)$ is now given by

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma_i)^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^\tau \Sigma_i^{-1}(x-\mu_i)}.$$

Result 3. If the populations π_1 and π_2 both have multivariate normal densities with mean vectors and covariance matrices μ_1, Σ_1 and μ_2, Σ_2 respectively, then the classification rule corresponding to minimizing ECM becomes:

Classify x_0 as π_1 if

$$-\frac{1}{2}x_0^\tau(\Sigma_1^{-1} - \Sigma_2^{-1})x_0 + (\mu_1^\tau\Sigma_1^{-1} - \mu_2^\tau\Sigma_2^{-1})x_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

and classify x_0 as π_2 otherwise.

The constant k is given by

$$k = \frac{1}{2}\ln\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right) + \frac{1}{2}(\mu_1^\tau\Sigma_1^{-1}\mu_1 - \mu_2^\tau\Sigma_2^{-1}\mu_2)$$

Proof. We assign x_0 to π_1 if

$$\ln\left(\frac{f_1(x_0)}{f_2(x_0)}\right) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \quad \text{and}$$

$$\ln\left(\frac{f_1(x_0)}{f_2(x_0)}\right) = -\frac{1}{2}\ln\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right) + \frac{1}{2}(x_0-\mu_2)^\tau\Sigma_2^{-1}(x_0-\mu_2) - \frac{1}{2}(x_0-\mu_1)^\tau\Sigma_1^{-1}(x_0-\mu_1)$$

$$= -\frac{1}{2}x_0^\tau(\Sigma_1^{-1} - \Sigma_2^{-1})x_0 + (\mu_1^\tau\Sigma_1^{-1} - \mu_2^\tau\Sigma_2^{-1})x_0 - k$$

□

In practice, the parameters μ_1, μ_2, Σ_1 and Σ_2 are unknown and replaced by the estimates $\bar{x}_1$, $\bar{x}_2$, S_1 and S_2 which yields the following sample classification rule:

Classify x_0 as π_1 if

$$-\frac{1}{2}x_0^\tau(S_1^{-1} - S_2^{-1})x_0 + (\bar{x}_1^\tau S_1^{-1} - \bar{x}_2^\tau S_2^{-1})x_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

and classify x_0 as π_2 otherwise.

The constant k is given by

$$k = \frac{1}{2}\ln\left(\frac{\det(S_1)}{\det(S_2)}\right) + \frac{1}{2}(\bar{x}_1^\tau S_1^{-1}\bar{x}_1 - \bar{x}_2^\tau S_2^{-1}\bar{x}_2)$$

4. Evaluating Classification Rules

To judge the performance of a sample classification procedure, we want to calculate its misclassification probability or **error rate**.

A measure of performance that can be calculated for any classification procedure is the **apparent error rate** (APER) which is defined as the fraction of observations in the sample that are misclassified by the classification procedure. Denote n_{1M} and n_{2M} the number of objects misclassified as π_1 respectively π_2 objects, then

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

The APER is intuitively appealing and easy to calculate. Unfortunately, it tends to underestimate the **actual error rate** (AER) when classifying new objects. This underestimation occurs because we used the sample to “build” the classification rule (therefore we call this the “training sample”) as well as to evaluate it. To obtain a reliable estimate of the AER we ideally consider an independent “test sample” of new objects from which we know the true class label. This means that we split the original sample in a training sample and test sample. The AER is then estimated by the proportion of misclassified objects in the test sample while the training sample was used to construct the classification rule. However, there are two drawbacks with this approach

- It requires large samples.
- The classification rule is less precise because we do not use the information from the test sample to build the classifier.

An alternative is the (leave-one-out) **cross-validation** or **jackknife** procedure which works as follows.

1. Leave one object out of the sample and construct a classification rule based on the remaining $n-1$ objects in the sample.

2. Classify the left-out observation using the classification rule constructed in step 1.

3. Repeat the two previous steps for each of the objects in the sample. Denote n_{1M}^{CV} and n_{2M}^{CV} the number of left-out observations misclassified in class 1 and 2 respectively.

Then a good estimate of the actual error rate is given by

$$A\hat{E}R = \frac{n_{1M}^{CV} + n_{2M}^{CV}}{n_1 + n_2}$$

Example 3. We consider a sample of size $n = 98$ containing the response to visual stimuli for both eyes measured for patients suffering from multiple-sclerosis and for controls (healthy patients). Based on these measured responses and age we want to develop a rule that will allow to classify potential patients. Estimate the actual error rate as well. The assumption of equal covariances is acceptable. Prior probabilities and cost of misclassification are undetermined and thus considered to be equal.

Analyzing the data in S-Plus we find the group means

$$\bar{x}_1 = \begin{pmatrix} 37.98551 \\ 1.562319 \\ 1.62029 \end{pmatrix} \text{ and } \bar{x}_2 = \begin{pmatrix} 42.06897 \\ 12.275862 \\ 13.08276 \end{pmatrix} \text{ and}$$

the pooled covariance matrix $S_{\text{pooled}} = \begin{pmatrix} 231.9880 & -2.09989 & -6.4015 \\ -2.09989 & 93.81391 & 87.0732 \\ -6.4015 & 87.0732 & 104.0572 \end{pmatrix}$

The classification rule becomes: Classify patient as suffering from multiple-sclerosis if $\hat{y} - \hat{m} = -0.012x_1 + 0.019x_2 + 0.147x_3 + 1.657 \geq 0$ and otherwise the patient is healthy. Based on the training sample we obtain the following misclassifications: $n_{1M} = 14$ and $n_{2M} = 3$ which yields the apparent error rate APER $= \frac{14+3}{98} = 17.3\%$.

On the other hand, by using cross-validation we obtain the misclassifications $n_{1M}^{CV} = 15$ and $n_{2M}^{CV} = 5$ which yields the estimated actual error rate $A\hat{E}R = \frac{15+5}{98} = 20.4\%$ which is 3% higher!

Note that three times as many persons are misclassified as MS patients than as healthy even while misclassification cost was assumed equal.

5. Classification with Several Populations

We now consider the more general situation of separating objects from g ($g \geq 2$) classes and assigning new objects to one of these g classes.
For $i = 1, \ldots, g$ denote

- f_i the density associated with population π_i
- p_i the prior probability of population π_i
- R_i the subspace of outcomes assigned to π_i
- $c(j|i)$ the cost of misclassifying an object to π_j when it is from π_i
- $P(j|i)$ the conditional probability of assigning an object of π_i to π_j.

The (conditional) expected cost of misclassifying an object of population π_1 is given by

$$\begin{aligned} \mathrm{ECM}(1) &= P(2|1)c(2|1) + \cdots + P(g|1)c(g|1) \\ &= \sum_{i=2}^{g} P(i|1)c(i|1) \end{aligned}$$

and similarly we can determine the expected cost of misclassifying objects of population $\pi_2, \ldots, \pi_g$. It follows that the overall ECM equals

$$\mathrm{ECM} = \sum_{j=1}^{g} p_j \,\mathrm{ECM}(j) = \sum_{j=1}^{g} p_j \sum_{\substack{i=1 \\ i \neq j}}^{g} P(i|j)c(i|j)$$

Result 4. The classification rule that minimizes the ECM assigns each object x to the population π_i for which

$$\sum_{\substack{j=1 \\ j \neq i}}^{g} p_j f_j(x) c(i|j)$$

is smallest. If the minimum is not unique then x can be assigned to any of the populations for which the minimum is attained.

(without proof)

Special Case: If all misclassification costs are equal (or unknown) we assign x to the population π_i for which $\sum_{\substack{j=1 \\ j \neq i}}^{g} p_j f_j(x)$ is smallest, or equivalently for which $p_i f_i(x)$ is largest. We thus obtain

$$\text{Classify } x \text{ as } \pi_i \text{ if } p_i f_i(x) > p_j f_j(x) \quad \forall j \neq i$$

5.1. Classification with Normal Populations

The density of population π_i $(i = 1, \ldots, g)$ is now given by

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma_i)^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^\tau \Sigma_i^{-1}(x-\mu_i)}.$$

Result 5. If all misclassification costs are equal (or unknown) we assign x to the population π_i if the (quadratic) score $d_i(x) = \max_{j=1}^{g} d_j(x)$ where the scores are given by

$$d_j(x) = -\frac{1}{2}\ln(\det(\Sigma_j)) - \frac{1}{2}(x-\mu_j)^\tau \Sigma_j^{-1}(x-\mu_j) + \ln(p_j) \quad j = 1, \ldots, g$$

Proof. We assign x to π_i if $\ln(p_i f_i(x)) = \max \ln(p_j f_j(x))$ and
$\ln(p_j f_j(x)) = \ln(p_j) - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln(\det(\Sigma_j)) - \frac{1}{2}(x-\mu_j)^\tau \Sigma_j^{-1}(x-\mu_j)$.
Dropping the second term which is constant yields the result. □

In practice, the parameters μ_j and Σ_j are unknown and will be replaced by the sample means $\bar{x}_j$ and covariances S_j which yields the sample classification rule

Classify x as π_i if the (quadratic) score $\hat{d}_i(x) = \max_{j=1}^{g} \hat{d}_j(x)$ where the scores are given by

$$\hat{d}_j(x) = -\frac{1}{2}\ln(\det(S_j)) - \frac{1}{2}(x-\bar{x}_j)^\tau S_j^{-1}(x-\bar{x}_j) + \ln(p_j) \quad j = 1, \ldots, g$$

If all covariance matrices are equal: $\Sigma_j = \Sigma$ for $j = 1, \ldots, g$, then the quadratic scores $d_j(x)$ become

$$d_j(x) = -\frac{1}{2}\ln(\det(\Sigma)) - \frac{1}{2}x^\tau\Sigma^{-1}x + \mu_j^\tau\Sigma^{-1}x - \frac{1}{2}\mu_j^\tau\Sigma^{-1}\mu_j + \ln(p_j).$$

The first two terms are the same for all $d_j(x)$ so they can be left out, which yields the (linear) scores $d_j(x) = \mu_j^\tau\Sigma^{-1}x - \frac{1}{2}\mu_j^\tau\Sigma^{-1}\mu_j + \ln(p_j)$.

To estimate these scores in practice we use the sample means $\bar{x}_j$ and the pooled estimate of Σ given by

$$S_{\text{pooled}} = \frac{1}{(n_1 - 1) + \cdots + (n_g - 1)}[(n_1 - 1)S_1 + \cdots + (n_g - 1)S_g]$$

which yields the sample classification rule

Classify x as π_i if the (linear) score $\hat{d}_i(x) = \max_{j=1}^{g} \hat{d}_j(x)$ where the scores are given by

$$\hat{d}_j(x) = \bar{x}_j^\tau S_{\text{pooled}}^{-1}x - \frac{1}{2}\bar{x}_j^\tau S_{\text{pooled}}^{-1}\bar{x}_j + \ln(p_j) \quad j = 1, \ldots, g$$

Remark. In the case of equal covariance matrices, the scores $d_j(x)$ can also be reduced to

$$d_j(x) = -\frac{1}{2}(x - \mu_j)^\tau\Sigma^{-1}(x - \mu_j) + \ln(p_j) = -\frac{1}{2}d_\Sigma^2(x, \mu_j) + \ln(p_j)$$

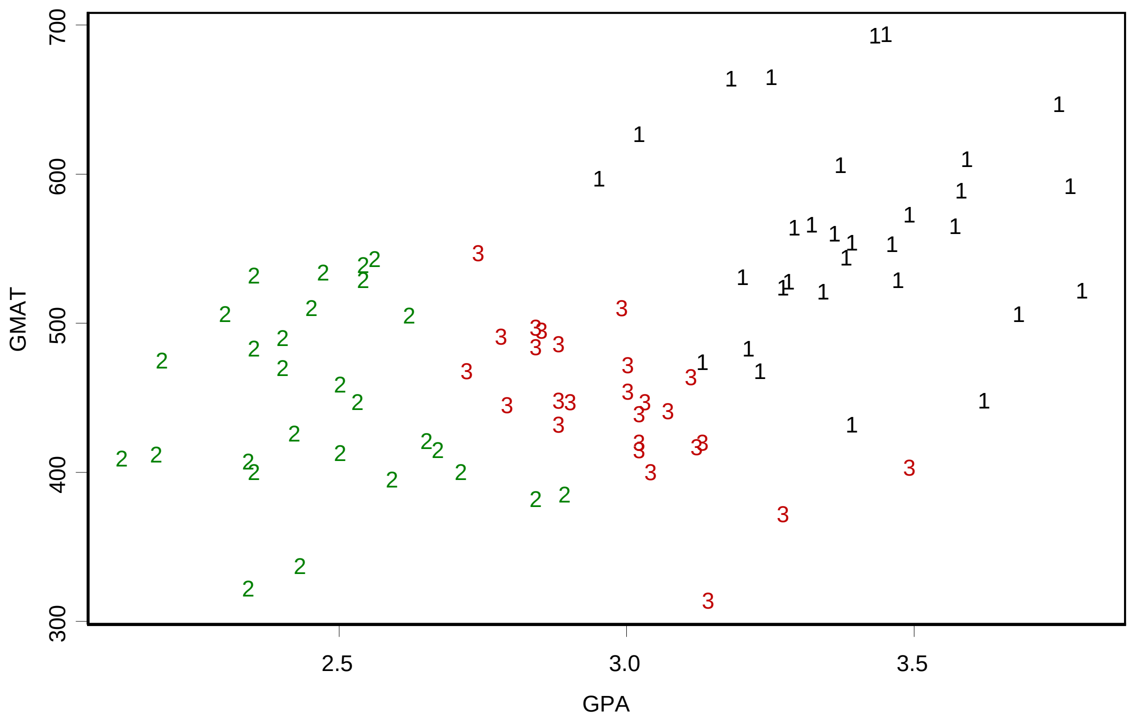
Which can be estimated by $\hat{d}_j(x) = -\frac{1}{2}d_{S_{\text{pooled}}}^2(x, \bar{x}_j) + \ln(p_j)$.

If the prior probabilities are all equal (or unknown) we thus assign an object x to the closest population.

Example 4. The admission board of a business school uses two measures to decide on admittance of applicants:

- GPA= undergraduate grade point average
- GMAT=graduate management aptitude test score

Based on these measures applicants are categorized as: admit (π_1), do not admit (π_2), and borderline (π_3). The training set is shown below.



Based on the training sample with group sizes $n_1 = 31$, $n_2 = 28$, and $n_3 = 26$ we calculate the group means

$$\bar{x}_1 = \begin{pmatrix} 3.40 \\ 561.23 \end{pmatrix}, \; \bar{x}_2 = \begin{pmatrix} 2.48 \\ 447.07 \end{pmatrix}, \text{ and } \bar{x}_3 = \begin{pmatrix} 2.99 \\ 446.23 \end{pmatrix},$$

and their pooled covariance matrix $S_{\text{pooled}} = \begin{pmatrix} 0.0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{pmatrix}$

With equal prior probabilities we assign a new applicant $x = (x_1, x_2)^\tau$ to the closest class, so we compute its (quadratic) distance to each of the three classes:

$$\begin{aligned}
d^2_{S_{\text{pooled}}}(x, \bar{x}_1) &= (x - \bar{x}_1)^\tau S_{\text{pooled}}(x - \bar{x}_1) \\
d^2_{S_{\text{pooled}}}(x, \bar{x}_2) &= (x - \bar{x}_2)^\tau S_{\text{pooled}}(x - \bar{x}_2) \\
d^2_{S_{\text{pooled}}}(x, \bar{x}_3) &= (x - \bar{x}_3)^\tau S_{\text{pooled}}(x - \bar{x}_3)
\end{aligned}$$

Suppose a new applicant has test scores $x^\tau = (3.21, 497)$ then we obtain

$$\begin{aligned}
d^2_{S_{\text{pooled}}}(x, \bar{x}_1) &= \begin{pmatrix} 3.21 - 3.40 & 497 - 561.23 \end{pmatrix} \begin{pmatrix} 28.61 & 0.016 \\ 0.016 & 0.0003 \end{pmatrix} \begin{pmatrix} 3.21 - 3.40 \\ 497 - 561.23 \end{pmatrix} \\
&= 2.58 \\
d^2_{S_{\text{pooled}}}(x, \bar{x}_2) &= \begin{pmatrix} 3.21 - 2.48 & 497 - 447.07 \end{pmatrix} \begin{pmatrix} 28.61 & 0.016 \\ 0.016 & 0.0003 \end{pmatrix} \begin{pmatrix} 3.21 - 2.48 \\ 497 - 447.07 \end{pmatrix} \\
&= 17.10 \\
d^2_{S_{\text{pooled}}}(x, \bar{x}_3) &= \begin{pmatrix} 3.21 - 2.99 & 497 - 446.23 \end{pmatrix} \begin{pmatrix} 28.61 & 0.016 \\ 0.016 & 0.0003 \end{pmatrix} \begin{pmatrix} 3.21 - 2.99 \\ 497 - 446.23 \end{pmatrix} \\
&= 2.47
\end{aligned}$$

The distance from x to π_3 is thus smallest such that this applicant is a borderline case.