Discriminant Analysis
and/or
Classification

Basic Problem: Classify an individual as belonging to one of several well defined populations.

Examples:

1. Classify insects into one of several sub-species using measurements on external features.

2. Use information on background, family support, psychological test scores, etc... to screen applicants for parole from prison.

3. Use measurements on blood proteins and family history to classify women as carriers or non-carriers of a genetic disorder.

4. Use information on credit history, income, employment history, home ownership, current debt, etc... to classify loan applicants.

5. Use information on sex, age, income, education level, credit card use, marital status, etc..... to classify potential targets of a mail marketing campaign.
Given values of measurements on some traits, what is the "best" way to classify?

Minimize the probability of misclassifying when p=1 trait is measured.

Suppose $X =$ income has a normal distribution for each population

- Population 1 (responders) $X \sim N(\mu_1, \sigma^2)$
- Population 2 (non-responders) $X \sim N(\mu_2, \sigma^2)$

A likelihood ratio rule:

$$
\lambda = \frac{\text{likelihood of } X \text{ for pop. 1}}{\text{likelihood of } X \text{ for pop. 2}}
$$

$$
= \frac{f_1(x)}{f_2(x)} \exp\left[-\frac{1}{2}(x-\mu_1)^2/\sigma^2\right]\frac{1}{\sqrt{2\pi} \sigma} \exp\left[-\frac{1}{2}(x-\mu_2)^2/\sigma^2\right]\frac{1}{\sqrt{2\pi} \sigma}
$$
\[
\lambda = e^{-\frac{1}{2}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2}\right)}
\]

Classify into population 1 if \( \lambda > 1 \) (flip a coin if \( \lambda = 1 \)) otherwise classify the individual into population 2.

Classify into population 1 if the "standardized" distance of \( x \) from \( \mu_1 \) is less than the "standardized" distance of \( x \) from \( \mu_2 \).

Consider

\[
-2 \ln(\lambda) = \frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2}
= -2\frac{(\mu_1-\mu_2)x + \mu_1^2-\mu_2^2}{\sigma_1^2 + \sigma_2^2}
= \beta x + \alpha
\]

Classify into population 1 if this is less than 0. A linear rule.

Suppose \( p \) traits are measured and for each population the distribution of traits across members of the population is multivariate normal.

Population 1: \( X \sim \mathcal{N}_p(\mu_1, \Sigma_1) \)

Population 2: \( X \sim \mathcal{N}_p(\mu_2, \Sigma_2) \)

Consider the natural logarithm of the likelihood ratio for an observed \( x \) for some individual.

\[
-2\ln\left(\frac{\xi_1(x)}{\xi_2(x)}\right) = -2\ln\left(\frac{1}{\sqrt{(2\pi)^p|\Sigma_1|}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)} \right)
\]

\[
-2\ln\left(\frac{\xi_2(x)}{\xi_1(x)}\right) = -2\ln\left(\frac{1}{\sqrt{(2\pi)^p|\Sigma_2|}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2)} \right)
\]
This reduces to a linear rule when the covariance matrices are equal, \( \Sigma_1 = \Sigma_2 = \Sigma \).

\[
\begin{align*}
&= 2m|\Sigma_1| + (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1) \\
&\quad - \left[ 2m|\Sigma_2| + (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) \right]
\end{align*}
\]

Classify into population 1 if this quantity is less than zero; otherwise classify into population 2.

This is a quadratic rule.

For \( p = 2 \),

\[
\begin{bmatrix} X_2 \\ X_1 \end{bmatrix}
\]

Multiply by \(-\frac{1}{2}\) and classify into population 1 if

\[
\begin{bmatrix} 2 \Sigma_1^{-1} x - \frac{1}{2} \Sigma_1^{-1} \Sigma_2^{-1} \mu_1 \\ \Sigma_2^{-1} x - \frac{1}{2} \Sigma_2^{-1} \Sigma_1^{-1} \mu_2 \end{bmatrix}
\]
The linear classification rule can be written as

\[
\left( x_1 - \mu_2 \right) \Sigma^{-1} x - \frac{1}{2} \left( x_1 - \mu_2 \right) \Sigma^{-1} (\mu_1 + \mu_2) > 0
\]

Fisher's linear discriminant function.

When \( \mu_1, \mu_2 \) and \( \Sigma \) are known the probability of misclassification can be determined.

\[
P(x|1) = \Pr \left( \text{classify into pop. 2} \mid \text{x is a measurement on an individual in population 1} \right)
\]

\[
= \Pr \left( (x_1 - \mu_2) \Sigma^{-1} x = \frac{1}{2} (x_1 - \mu_2) \Sigma^{-1} (\mu_1 + \mu_2) \mid x \sim \mathcal{N}(\mu_1, \Sigma) \right)
\]

\[
= \Phi \left( -\frac{1}{2} \delta \right)
\]

where \( \delta^2 = (x_1 - \mu_2) \Sigma^{-1} (\mu_1 - \mu_2) \)

Similarly,

\[
P(1|1) = \Phi \left( -\frac{1}{2} \delta \right)
\]
Is this the best way to classify?

In Section 11.2 classification is considered in terms of decision theory and the criterion is to minimize the expected cost of misclassification (ECM).

Suppose the prior probability that an individual comes from population 1 is \( p_1 \) and the prior probability that an individual comes from population 2 is \( p_2 = 1 - p_1 \).

Specify misclassification costs

<table>
<thead>
<tr>
<th>Classification</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True population</td>
<td>1</td>
<td>( C(2</td>
</tr>
<tr>
<td>2</td>
<td>( C(1</td>
<td>2) )</td>
</tr>
</tbody>
</table>

The expected cost of misclassification is

\[
ECM = C(2|1) P(2|1) p_1 + C(1|2) P(1|2) p_2
\]

This is minimized by classifying an individual with measurements \( x \) into population 1 if

\[
\frac{f_1(x)p_1}{f_2(x)p_2} \geq \frac{C(1|2)}{C(2|1)}
\]
The misclassification costs can be absorbed into the specification of the priors:

Classify into population 1 if

\[
\frac{f_1(x)}{f_2(x)} > \frac{\log(p_2)}{\log(p_1)} = \frac{p_2^*}{p_1^*}
\]

or

Classify into population 1 if

\[
\ln(f_1(x)) + \ln(p_1^*) > \ln(f_2(x)) + \ln(p_2^*)
\]

With equal misclassification costs, ECM is minimized by classifying a unit with observation \( x \) into the population with the largest posterior probability.

Classify into population k if

\[
\ln(f_k(x)) + \ln(p_k^*)
\]

is largest.
For multivariate normal distributions, classify into the population with the largest quadratic score

\[ d_k^Q(x) = -\frac{1}{2} \ln(1 |x_k|^2) \]
\[ -\frac{1}{2} (x-x_k)^T \Sigma_k^{-1} (x-x_k) \]
\[ + \ln(p_k) \]
\[ = \ln(f_k(x)) + \ln(p_k) \]

This is called "quadratic" discriminant analysis.

For multivariate normal distributions with homogeneous covariance matrices

\[ \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k \]

\[ \ln(f_k(x)) + \ln(p_k) \]
\[ = -\frac{1}{2} \ln(1 |x|^2) - \frac{1}{2} (x-x_k)^T \Sigma^{-1} (x-x_k) \]
\[ + \ln(p_k) \]
\[ = -\frac{1}{2} \ln(1 |x|^2) - \frac{1}{2} x^T \Sigma^{-1} x + \frac{1}{2} x_k^T \Sigma^{-1} x_k \]
\[ + \frac{1}{2} x_k^T \Sigma^{-1} x - \frac{1}{2} x_k^T \Sigma^{-1} x_k + \ln(p_k) \]
\[ = -\frac{1}{2} \ln(1 |x|^2) - \frac{1}{2} x^T \Sigma^{-1} x \]
\[ + \frac{1}{2} x_k^T \Sigma^{-1} x - \frac{1}{2} x_k^T \Sigma^{-1} x_k + \ln(p_k) \]

this is the only part that varies across populations
Classify a unit with measurement \( \bar{x} \) into the population with the largest linear discriminant score

\[
d_k(\bar{x}) = \frac{1}{2} \left( \mu_k - \mu \right) ^\prime \Sigma_k ^{-1} \left( \mu_k - \mu \right) + \ln(\pi_k)
\]

This is called "linear" discriminant analysis.

Use information from training samples to estimate unknown parameters.

\[
prior \quad \text{population 1: } \quad \bar{x} \sim N(\mu_1, \Sigma_1) \quad \pi_1
\]

\[
population k: \quad \bar{x} \sim N(\mu_k, \Sigma_k) \quad \pi_k
\]

Independent Training Samples:

Random sample of \( n_j \) units from population \( j \):

\[
X_{j1}, X_{j2}, \ldots, X_{jn_j}
\]

Compute:

\[
\bar{x}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} X_{jk} \\
S_j = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (X_{jk} - \bar{x}_j)(X_{jk} - \bar{x}_j)^\prime
\]
Should a linear or quadratic discriminant rule be used?

Use Bartlett’s test to test

\[ H_0: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_K \]

versus

\[ H_a: \Sigma_i \neq \Sigma_j \text{ for some } i \neq j \]

This can be done with

PROC DISCRIM in SAS

Do not pick a "large" significance level (use \( \alpha \leq .01 \))

If you decide that the covariance matrices are not homogeneous, classify a unit with measurement \( x \) into the population with the largest value of

\[
\widehat{d}_j^2(x) = -\frac{1}{2} \ln|\Sigma_j| - \frac{1}{2} (x - \overline{x}_j)' \Sigma_j^{-1} (x - \overline{x}_j) + \ln(p_j)
\]
A linear classification rule can perform well even when covariance matrices are not the same or data are non-normal.

If you decide that $z' = z_0 = \ldots = z_k$ is not too severely violated, classify a unit with measurements $x$ into the population with the largest value of $d_j(x) = \frac{\bar{x}_j - \bar{x}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}$, where

$$S = \frac{x_j - \bar{x}_j}{\sqrt{(n_j-1) \frac{s_j^2}{n_j}}},$$

and $s_j^2$ is the sample variance.
Estimation of misclassification probabilities:

\[ p(\theta) \] conditional probability of misclassifying a case into population 2 "given" that the case actually belongs to population 1

\[ p(\phi) \] conditional probability of misclassifying a case into population 1 "given" that the case actually belongs to population 2

(1) Resubstitution method (apparent error rate):

Use the classification rule estimated from the training samples to classify each member of the training samples.

Compute proportions of correct and incorrect classifications classified as

<table>
<thead>
<tr>
<th>True status</th>
<th>pop.1</th>
<th>pop.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>pop.1</td>
<td>( \hat{\theta} )</td>
<td>correct</td>
</tr>
<tr>
<td>pop.2</td>
<td>( \hat{\phi} )</td>
<td>proportion correct</td>
</tr>
</tbody>
</table>
This method is biased
\[ \hat{p}(211) \text{ tends to underestimate } p(211) \]
\[ \hat{p}(112) \text{ tends to underestimate } p(112) \]

The bias is smaller for larger training samples.

The estimated classification rule tends to be better at separating the training samples than distinguishing between the underlying populations.

\[ \hat{p}(211) \text{ and } \hat{p}(112) \text{ decrease as more variables are included.} \]

2. Data splitting ("set aside" method)

- Randomly select some cases from the training samples and "set them aside."
- Use the remaining data to build the classification rule.
- Classify the cases that were set aside and compute proportions of misclassifications.

\[ \hat{p}(112) \]
\[ \hat{p}(211) \]
3. Crossvalidation:

- Randomly split the cases in the training samples into $g$ groups.
- Set aside one of the $g$ groups
  - build the classification rule from the cases in the other $g-1$ groups
  - classify the cases in the group that was set aside
  - record the results
- Repeat previous step $g$ times
4. Lachenbruch’s "holdout" method
   • crossvalidation with $g=n$
     groups, each with one case.
   • called "crossvalidation" is SAS
   • sometimes called a "jackknife" method, but it really isn’t
   • biased estimates of misclassification rates
     - less biased than apparent error rates
     - can increase when more variables are included.
   • efficient computation for linear discriminant analysis
     (Bartlett, 1951)
5. Bootstrap estimation

(Step 1) Use simple random sampling with replacement to select new training samples.

(Step 2) Fit the model to the new training samples.

- Compute apparent error rates for new training samples
  \[ \hat{P}_B(112) \text{ and } \hat{P}_B(211) \]

- Compute "apparent" error rates for the original training samples
  \[ \bar{P}_B^*(112) \text{ and } \bar{P}_B^*(211) \]

- Estimate bias
  \[ \text{bias}(112) = \bar{P}_B^*(112) - \hat{P}_B(112) \]
  \[ \text{bias}(211) = \bar{P}_B^*(211) - \hat{P}_B(211) \]

Repeat steps 1 and 2 a large number, say \( B \), times. Compute the mean of the bias estimates
\[ \hat{\text{bias}}(112) \text{ and } \hat{\text{bias}}(211) \]

Compute adjusted error rates
\[ \hat{P}(112) + \hat{\text{bias}}(112) \]
\[ \hat{P}(211) + \hat{\text{bias}}(211) \]

apparent error rate from the original training samples
Reference:
B. Efron and R. J. Tibshirani, 1993,
An Introduction to the Bootstrap
chapman & Hall, New York,
Chapter 17

Variable selection:

- Information on many variables (or traits) is recorded for the training samples.

- Apparent error rates will not increase as more variables are included in the classification rule.

- Probability of misclassifying a "new" case may increase when more variables are included in the classification model.
• **PROC STEPDISC** in SAS
  * will perform stepwise variable selection for linear discriminant analysis (assumes $\Sigma_1 = \ldots = \Sigma_k$)
  * Backward elimination
  * Stepwise procedure

  * use partial F-tests only as guidelines; also check cross-validation or bootstrap estimates of misclassification probabilities.

**Alternative methods:**

- Construct a classification rule from the ratio of likelihoods corresponding to the joint distributions of the traits for the various populations.

- Other methods:
  - Logistic regression
  - Nearest neighbor methods
  - CART: Classification and Regression Trees
  - Neural Nets