Chapter 11: Discrimination and Classification

from Applied Multivariate Statistical Analysis by Johnson and Wichern

Greg Miller

April 9, 2007

Goals

- Goal 1: To describe(graphically or algebraically) the differential features of objects from several populations $\longrightarrow$ Discrimination
- Goal 2: To sort objects into two or more labeled classes, emphasizing a rule that can be used to assign new objects to labeled classes $\longrightarrow$ Classification

Example Scenarios

- Good/Poor Credit Risks: measured by income, age, family size, number of credit cards
- *Federalist Papers* written by Madison and those written by Hamilton: measured by freequencies of different words and sentence length.

Classification: Why are there unknown objects?

How can we be certain about classification of some objects, and uncertain about others?
When one of these conditions is true:

- Incomplete Knowledge of Future performance
- "Perfect" information requires destroying the object
- Unavailable or expensive information

The Probability of Misclassification

π_1, π_2: Two populations, or 'classes'
$f_1(x), f_2(x)$: Probability density functions associated with the $p \times 1$ vector rv X for populations π_1, π_2
R_1, R_2: Set of x values where we classify objects in $\boldsymbol{\pi_1, \pi_2}$, respectively.
Now, the conditional probability of classifying an object in $\boldsymbol{\pi_2}$ when it is part of $\boldsymbol{\pi}$ is:

$$\mathbf{P(2 \mid 1) = P(X \in R_2 \mid \pi_1) = \int_{R_2 = \Omega - R_1} f_1(x) dx} \qquad (1)$$

Overall Probabilities

Let p_1, p_2 be the prior probabilities of π_1, π_2 respectively. $p_1 + p_2 = 1$ The overall probabilities are as follows:

- $P(\text{observation correctly classified as } \pi_1) = P(1 \mid 1)p_1$
- $P(\text{observation incorrectly classified as } \pi_1) = P(1 \mid 2)p_2$
- $P(\text{observation incorrectly classified as } \pi_2) = P(2 \mid 2)p_2$

Expected Cost of Misclassification (ECM)

If we evaluate a classification scheme by its misclassification probabilities, we would not be considering the cost of incorrectly classifying an object.
Let $c(2 \mid 1)$ be the cost for misclassifying object 2 in population 1.
Let $c(1 \mid 2)$ be the cost for misclassifying object 1 in population 2.
The Expected Cost of Misclassification is given by:

$$ECM = c(2 \mid 1)P(2 \mid 1)p_1 + c(1 \mid 2)P(1 \mid 2)p_2 \qquad (2)$$

Minimizing ECM

The regions R_1 and R_2 that minimize ECM are the values of x such that the following inequalities hold:

$$\mathbf{R1}: \frac{\mathbf{f_1(x)}}{\mathbf{f_2(x)}} \geq \frac{\mathbf{c(1 \mid 2)}}{\mathbf{c(2 \mid 1)}} \frac{\mathbf{p_2}}{\mathbf{p_1}} \tag{3}$$

$$\mathbf{R2}: \frac{\mathbf{f_1(x)}}{\mathbf{f_2(x)}} < \frac{\mathbf{c(1 \mid 2)}}{\mathbf{c(2 \mid 1)}} \frac{\mathbf{p_2}}{\mathbf{p_1}} \tag{4}$$

Special Cases of Minimum Expected Cost Regions

- When $p_1/p_2 = 1$ (equal priors)

$$R1 : \frac{f_1(\mathbf{x})}{f_2(x)} \geq \frac{c(1 \mid 2)}{c(2 \mid 1)} \qquad R2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1 \mid 2)}{c(2 \mid 1)}$$

- When $c(1 \mid 2)/c(2 \mid 1) = 1$ (equal misclassification costs)

$$R1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \qquad R2 : \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1} \tag{5}$$

- When priors and misclassification costs are both equal

$$R1 : \frac{f_1(x)}{f_2(x)} \geq 1 \qquad R2 : \frac{f_1(x)}{f_2(x)} < 1$$

Example 11.2, page 589

A researcher has enough data available to estimate the density functions $f_1(x), f_2(x)$ associated with the population parameters π_1, π_2, respectively. Suppose that:

- $c(2 \mid 1) = 5$ units
- $c(1 \mid 2) = 10$ unit
- Twenty of all objects belong to π_2 (this means priors are $p_1 = .80, p_2 = .20$)

What are the classification regions R_1, R_2? If $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$, in which population do we classify $\mathbf{x}_0$?

Example 11.2, continued

$$R1 : \frac{f_1(\mathbf{x})}{f_2(x)} \geq (\frac{10}{5})(\frac{.2}{.8}) = .5$$

$$R2 : \frac{f_1(\mathbf{x})}{f_2(x)} < (\frac{10}{5})(\frac{.2}{.8}) = .5$$

And to classify the object:

$$\frac{f_1(\mathbf{x})}{f_2(x)} = \frac{.3}{.4} = .75$$

This result is greater than .5, so we classify it as belonging to π_1.

Total Probability of Misclassification (TPM)

If costs are not a concern, you can derive a classification procedure based entirely off of the total probability of misclassification(TPM):

$$= \mathrm{P}(\text{misclassifying a } \pi_1 \text{ observation or a } \pi_2 \text{ observation})$$

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

When $\Sigma_1 = \Sigma_2 = \Sigma$

Let

$$f_i(x) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} exp[(-1/2)(\mathbf{x} - \mu_\mathbf{i})^T \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{i})]$$

be the joint densities of $\mathbf{X}' = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}]$
Then (Result 11.2) allocate $\mathbf{X_0}$ to π_1 if the following holds:

$$(\mu_\mathbf{1} - \mu_\mathbf{2})^T \Sigma^{-1} \mathbf{x_0} - (1/2)(\mu_\mathbf{1} - \mu_\mathbf{2})^T \Sigma^{-1} (\mu_\mathbf{1} + \mu_\mathbf{i}) \geq ln[\frac{c(1 \mid 2)}{c(2 \mid 1)} \frac{p_2}{p_1}]$$

Allocate it to π_2 otherwise

Minimum ECM Classification Rule

Let

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

Using the statistics instead of parameters($\overline{x}_i$ for μ_i, and S-pooled for Σ), //we arrive at the following rule for minimizing ECM for Two Normal Populations. We allocate $\mathbf{x_0}$ to π_1 if:

$$(\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2})^T S_{pooled}^{-1}\mathbf{x_0} - \frac{1}{2}(\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2})^T S_{pooled}^{-1}(\overline{\mathbf{x}}_\mathbf{1} + \overline{\mathbf{x}}_\mathbf{1}) \geq ln[\frac{c(1 \mid 2)}{c(2 \mid 1)}\frac{p_2}{p_1}]$$

When $\Sigma_1 \neq \Sigma_2$

When covariance matrices are not equal, we use a slightly different formula for classification.

- We no longer used S_{pooled}, now we use S_1, S_2
- Classification rule calculates to the following. Allocate $\mathbf{x}_0$ to π_1 if:

$$-\frac{1}{2}\mathbf{x}_0^T(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1^T\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2^T\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq ln[\frac{c(1 \mid 2)}{c(2 \mid 1)}\frac{p_2}{p_1}]$$

$$k = \frac{1}{2}ln(\frac{\mid \Sigma_1 \mid}{\mid \Sigma_2 \mid}) + \frac{1}{2}(\mu_1^T\mathbf{\Sigma_1^{-1}}\mu_1 - \mu_2^T\mathbf{\Sigma_2^{-1}}\mu_2)$$

Evaluation

We are concerned with the future performance of a classification function. There are multiple ways of measuring the effectiveness of a classification system (as seen in the last section). There are also multiple ways of evaluating **future performance.**

- **OER**
- **AER**
- **APER**

OER

The Optimum Error Rate results from minimizing the Total Probability of Misclassification.

- Assumes costs of misclassification are equal.
- OER is given by:

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

- R_1, R_2 are the regions determined from the minimizing the ECM when costs are equal

AER

- Actual Error Rate (AER) is similiar to OER, but it deals the *sample* classification function
- AER is given by:

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

- The Regions are given by

$$\hat{R}_1 : (\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2})^T S_{pooled}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2})^T S_{pooled}^{-1}(\bar{\mathbf{x}}_\mathbf{1} + \bar{\mathbf{x}}_\mathbf{1}) \geq ln[\frac{c(1 \mid 2)}{c(2 \mid 1)}\frac{p_2}{p_1}]$$

$$\hat{R}_2 : (\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2})^T S_{pooled}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2})^T S_{pooled}^{-1}(\bar{\mathbf{x}}_\mathbf{1} + \bar{\mathbf{x}}_\mathbf{1}) < ln[\frac{c(1 \mid 2)}{c(2 \mid 1)}\frac{p_2}{p_1}]$$

APER

Note that AER and OER cannot be usually calculated, since they depend on unknown density functions $f_1(\mathbf{x}), f_2(\mathbf{x})$.

- We can use the Apparent Error Rate(APER), which is the fractions of observations in the training set that are misclassified.
- We can use a confusion matrix to calculate this:

		Predicted Membership	
		π_1	π_2
Actual membership	π_1	n_{1c}	$n_{1M} = n_1 - n_{1c}$
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}

APER

- n_{1C}, n_{2C}: number of items correctly classified in groups π_1, π_2 respectively
- n_{1M}, n_{2M}: number of items incorrectly classified in groups π_1, π_2 respectively

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \tag{6}$$

Lachenbruch's "Holdout" Procedure

The APER tends to underestimate the AER, unless n_1, n_2 are both very large. To calculate an error-rate estimate, we can use the Lachenbruch's "holdout" procedure". AKA: Jacknifing

1. Start with the π_1 group of observations, omit one observation, develop functions
2. Classify the "holdout" observation
3. Repeat Steps 1 and 2 until all observations are classified. Let $n_{1M}^{(H)}$ be number of misclassifications
4. Repeat Steps 1 through 3 with the π_2 observations. Let $n_{2M}^{(H)}$ be number of misclassifications

Now we can estimate conditional misclassification probabilities:

$$\hat{P}(2 \mid 1) = \frac{n_{1M}^{(H)}}{n_1} \qquad \hat{P}(1 \mid 2) = \frac{n_{2M}^{(H)}}{n_2}$$

$\hat{E}(AER)$

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \tag{7}$$

Fisher's Idea

- To use Linear combinations of $\mathbf{X}$ to create y's. $\hat{\mathbf{a}}^T x$
- Assumes equal variance, but does not assume normal populations

Allocation Rule Based on Fisher

Allocate $\mathbf{x}_0$ to π_1 if

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{pooled}^{-1} \mathbf{x}_0 \geq \hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$