Published on *STAT 505* (https://onlinecourses.science.psu.edu/stat505)

Lesson 10: Discriminant Analysis

Introduction

Discriminant analysis is a classification problem, where two or more groups or clusters or populations are known *a priori* and one or more new observations are classified into one of the known populations based on the measured characteristics. Let us look at three different examples.

Example 1 - Swiss Bank Notes:

We have two populations of bank notes, genuine, and counterfeit. Six measures are taken on each note:

- Length
- Right-Hand Width
- Left-Hand Width
- Top Margin
- Bottom Margin
- Diagonal across the printed area

Take a bank note of unknown origin and determine just from these six measurements whether or not it is real or counterfeit. Perhaps this is not as impractical as it might sound. A more modern equivalent is a scanner that would measure the notes automatically and makes a decision.

Example 2 - Pottery Data:

Pottery shards are sampled from four sites: L) Llanedyrn, C) Caldicot, I) Ilse Thornes, and A) Ashley Rails and the concentrations of the following chemical constituents were measured at a laboratory

- Al: Aluminum
- Fe: Iron
- Mg: Magnesium
- Ca: Calcium
- Na: Sodium

An archaeologist encounters a pottery specimen of unknown origin. To determine possible trade routes, the archaeologist may wish to classify its site of origin.

Example 3 - Insect Data:

Data were collected on two species of insects in the genus *Chaetocnema*, (a) *Ch. concinna* and (b) *Ch. heikertlingeri*. Three variables were measured on each insect:

- width of the 1^{st} joint of the tarsus (legs)
- width of the 2^{nd} joint of the tarsus
- width of the aedeagus (reproductive organ)

Our objective is to obtain a classification rule for identifying the insect species based on these three variables. An entomologist can identify these two closely related species, but the differences are so subtle that one has to have considerable experience to be able to tell the difference. If a classification rule may be developed, then this might be a more accurate way to help differentiate between these two different species.

Learning objectives & outcomes

Upon completion of this lesson, you should be able to do the following:

- Determine whether linear or quadratic discriminant analysis should be applied to a given data set;
- Be able to carry out both types of discriminant analyses using SAS/Minitab;
- Be able to apply the linear discriminant function to classify a subject by its measurements;
- Understand how to assess the efficacy of a discriminant analysis.

10.1 - Bayes Rule and Classification Problem

Bayes' Rule

Consider any two events *A* and *B*. To find *P*(*B*|*A*), the probability that *B* occurs given that *A* has occurred, Bayes' Rule states the following:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

This says that the conditional probability is the probability that both *A* and *B* occur divided by the unconditional probability that *A* occurs. This is a simple algebraic restatement of a rule for finding the probability that two events occur together, which is *P*(*A* and *B*) = *P*(*A*)*P*(*B*|*A*).

Bayes' Rule Applied to the Classification Problem

We are interested in $P(\pi_i \mid \boldsymbol{x})$, the conditional probability that an observation came from population π_i given that the observed values of the multivariate vector of variables ***x***. **We will classify an observation to the population for which the value of** $P(\pi_i \mid \boldsymbol{x})$ **is greatest**. This is the most probable group given the observed values of ***x***.

- Suppose that we have *g* populations (groups) and that the *i*th population is denoted as π_i .
- Let $p_i = P(\pi_i)$, be the probability that a randomly selected observation is in population π_i .
- Let $f(\boldsymbol{x} \mid \pi_i)$ be the conditional probability density function of the multivariate set of variables ***x***, given that the observation came from population π_i .

Technical Note: We have to be careful about the word probability in conjunction with our observed vector **x**. A probability density function for continuous variables does not give a probability, but instead gives a measure of "likelihood."

Using the notation of Bayes' Rule above, event *A* = observing the vector ***x*** and event *B* = observation came from population π_i . Thus our probability of interest can be found as

$$P(\text{ member of } \pi_i | \text{ we observed } \mathbf{x}) = \frac{P(\text{ member of } \pi_i \text{ and we observe } \mathbf{x})}{P(\text{ we observe } \mathbf{x})}$$

- The numerator of the expression just given is the likelihood that a randomly selected observation is both from population π_i and has the value ***x***. This likelihood = $p_i f(\boldsymbol{x} \mid \pi_i)$.

- The denominator is the unconditional likelihood (over all populations) that we could observe ***x***. This likelihood = $\sum_{j=1}^{g} p_j f(\mathbf{x}|\pi_j)$

Thus the posterior probability that an observation is a member of population π_i is

$$p(\pi_i|\mathbf{x}) = \frac{p_i f(\mathbf{x}|\pi_i)}{\sum_{j=1}^{g} p_j f(\mathbf{x}|\pi_j)}$$

The ***classification rule*** is to assign observation ***x*** to the population for which the posterior probability is the greatest.

The denominator is the same for all posterior probabilities (for the various populations) so it is equivalent to say that we will classify an observation to the population for which $p_i f(\mathbf{x} | \pi_i)$ is greatest.

Two Populations

With only two populations we can express a classification rule in terms of the ratio of the two posterior probabilities. Specifically we would classify to population 1 when

$$\frac{p_1 f(\mathbf{x}|\pi_1)}{p_2 f(\mathbf{x}|\pi_2)} > 1$$

This can be rewritten to say the we classify to population 1 when

$$\frac{f(\mathbf{x}|\pi_1)}{f(\mathbf{x}|\pi_2)} > \frac{p_2}{p_1}$$

Decision Rule

We are going to classify the sample unit or subject into the population π_i that maximizes the posterior probability $p(\pi_i)$. that is the population that maximizes

$$f(\mathbf{x}|\pi_\mathbf{i})p_i$$

We are going to calculate the posterior probabilities for each of the populations. Then we are going to assign the subject or sample unit to that population that has the highest posterior probability. Ideally that posterior probability is going to be greater than a half, the closer to 100% the better!

Equivalently we are going to assign it to the population that maximizes this product:

$$\log f(\mathbf{x}|\pi_\mathbf{i})p_i$$

The denominator that appears above does not depend on the population because it involves summing over all the populations. Equivalently all we really need to do is to assign it to the population that has the largest for this product, or equivalently we can maximize the log of that product. A lot of times it is easier to write the log.

10.2 - Discriminant Analysis Procedure

Discriminant analysis is a 7 step procedure:

- **Step 1**: Collect training data.

Training data are data with known group memberships. Here, we actually know which population contains each subject. For example, in the Swiss Bank Notes, we actually know which of these are genuine notes and which others are counterfeit examples.

- **Step 2**: Prior Probabilities:

The prior probability p_i represents the expected portion of the community that belongs to population π_i. There are three common choices:

1) Equal priors:

$$\hat{p}_i = \frac{1}{g}$$

This is useful if we believe that all of the population sizes are equal.

2) Arbitrary priors selected according to the investigators beliefs regarding the relative population sizes. Note that we require:

$$\hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_g = 1$$

3) Estimated priors:

$$\hat{p}_i = \frac{n_i}{N}$$

where n_i is the number observations from population π_i in the training data, and $N = n_1 + n_2 + ... + n_g$

- **Step 3**: Use Bartlett's test to determine if the variance-covariance matrices are homogeneous for all populations involved. The result of this test will determine whether to use Linear or Quadratic Discriminant Analysis.

Case 1: Linear discriminant analysis is for homogeneous variance-covariance matrices:

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$$

In this case the variance-covariance matrix does not depend on the population.

Case 2: Quadratic discriminant analysis is used for heterogeneous variance-covariance matrices:

$$\Sigma_i \neq \Sigma_j \text{ for some } i \neq j$$

This allows the variance-covariance matrices to depend on the population.

(We do not discuss testing whether the means of the populations are different. If they are not, there is no case for DA)

- **Step 4**: Estimate the parameters of the conditional probability density functions $f(\mathbf{X} \mid \pi_i)$. Here, we shall make the following standard assumptions:

1. The data from group *i* has common mean vector μ_i
2. The data from group *i* has common variance-covariance matrix Σ.
3. Independence: The subjects are independently sampled.
4. Normality: The data are multivariate normally distributed.

- **Step 5**: Compute discriminant functions. This is the rule to classify the new object into one of the known populations.

- **Step 6:** Use cross validation to estimate misclassification probabilities.

  As in all statistical procedures it is helpful to use diagnostic procedures to asses the efficacy of the discriminant analysis. We use cross-validation to assess the classification probability. Typically you are going to have some prior rule as to what is an acceptable misclassification rate. Those rules might involve things like, "what is the cost of misclassification?" This could come up in a medical study where you might be able to diagnose cancer. There are really two alternative costs. The cost of misclassifying someone as having cancer when they don't. This could cause a certain amount of emotional grief! There is also the alternative cost of misclassifying someone as not having cancer when in fact they do have it. The cost here is obviously greater if early diagnosis improves cure rates.

- **Step 7**: Classify observations with unknown group memberships.

The procedure described above assumes that the unit or subject being classified actually belongs to one of the considered populations. If you have a study where you look at two species of insects, A and B, and the insect to classify actually belongs to species C, then it will obviously be misclassified as to belonging to either A or B.

10.3 - Linear Discriminant Analysis

We assume that in population π_i the probability density function of $\boldsymbol{x}$ is multivariate normal with mean vector $\boldsymbol{\mu_i}$ and variance-covariance matrix Σ (same for all populations). As a formula, this is

$$f(\mathbf{x}|\pi_i) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu_\mathbf{i})\right]$$

We classify to the population for which $p_i f(\boldsymbol{x} \mid \pi_i)$ is largest.

Because a log transform is monotonic, this equivalent to classifying an observation to the population for which $\log[p_i f(\boldsymbol{x} \mid \pi_i)]$ is largest.

Linear discriminant analysis is used when the variance-covariance matrix does not depend on the population. In this case, our decision rule is based on the Linear Score Function, a function of the population means for each of our *g* populations, $\boldsymbol{\mu_i}$, as well as the pooled variance-covariance matrix.

The *Linear Score Function* is:

$$s_i^L(\mathbf{X}) = -\frac{1}{2}\mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mu_\mathbf{i} + \mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mathbf{x} + \log p_i = d_{i0} + \sum_{j=1}^{p} d_{ij}x_j + \log p_i$$

where

$$d_{i0} = -\frac{1}{2}\mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mu_\mathbf{i}$$

$$d_{ij} = j\text{th element of } \mu_i'\Sigma^{-1}$$

The far left-hand expression resembles a linear regression with intercept term d_{i0} and regression coefficients d_{ij}.

Linear Discriminant Function:

$$d_i^L(\mathbf{x}) = -\frac{1}{2}\mu_{\mathbf{i}}'\mathbf{\Sigma^{-1}}\mu_{\mathbf{i}} + \mu_{\mathbf{i}}'\mathbf{\Sigma^{-1}}\mathbf{x} = d_{i0} + \sum_{j=1}^{p} d_{ij}x_j$$

$$d_{i0} = -\frac{1}{2}\mu_{\mathbf{i}}'\mathbf{\Sigma^{-1}}\mu_{\mathbf{i}}$$

Given a sample unit with measurements x_1, x_2, ... , x_p, we classify the sample unit into the population that has the largest Linear Score Function. This is equivalent to classifying to the population for which the posterior probability of membership is largest. The linear score function is computed for each population, then we plug in our observation values and assign the unit to the population with the largest score.

However, this is a function of unknown parameters, $\boldsymbol{\mu_i}$ and Σ. So, these must be estimated from the data.

Discriminant analysis requires estimates of:

Prior probabilities:

$$p_i = \Pr(\pi_i); i = 1, 2, \ldots, g$$

The population means are estimated by the sample mean vectors:

$$\mu_{\mathbf{i}} = E(\mathbf{X}|\pi_i); i = 1, 2, \ldots, g$$

The variance-covariance matrix is estimated by using the pooled variance-covariance matrix

$$\Sigma = \text{var}(\mathbf{X}|\pi_i); i = 1, 2, \ldots, g$$

Typically, these parameters are estimated from training data, in which the population membership is known.

Conditional Density Function Parameters:

Population Means: $\boldsymbol{\mu_i}$ is estimated by substituting in the sample means $\bar{\mathbf{x}}_i$.

Variance-Covariance matrix: Let S_i denote the sample variance-covariance matrix for population *i*. Then the variance-covariance matrix Σ is estimated by substituting in the pooled variance-covariance matrix into the Linear Score Function as shown below:

$$\mathbf{S}_p = \frac{\sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i}{\sum_{i=1}^{g}(n_i - 1)}$$

to obtain the estimated linear score function:

$$\hat{s}_i^L(\mathbf{x}) = -\frac{1}{2}\bar{\mathbf{x}}_{\mathbf{i}}'\mathbf{S_p^{-1}}\bar{\mathbf{x}}_{\mathbf{i}} + \bar{\mathbf{x}}_{\mathbf{i}}'\mathbf{S_p^{-1}}\mathbf{x} + \log\hat{p}_i = \hat{d}_{i0} + \sum_{j=1}^{p}\hat{d}_{ij}x_j + \log p_i$$

where

$$\hat{d}_{i0} = -\frac{1}{2}\bar{\mathbf{x}}_{\mathbf{i}}'\mathbf{S}_{\mathbf{p}}^{-\mathbf{1}}\bar{\mathbf{x}}_{\mathbf{i}}$$

and

$$\hat{d}_{ij} = j\text{th element of } \bar{\mathbf{x}}_{\mathbf{i}}'\mathbf{S}_{\mathbf{p}}^{-\mathbf{1}}$$

This is a function of the sample mean vectors, the pooled variance-covariance matrix, and prior probabilities for *g* different populations. This is written in a form that looks like a linear regression formula with an intercept term plus a linear combination of response variables, plus the natural log of the prior probabilities.

Decision Rule: Classify the sample unit into the population that has the largest estimated linear score function.

10.4 - Example: Insect Data

Data were collected on two species of insects in the genus *Chaetocnema*, (species a) *Ch. concinna* and (species b) *Ch. heikertlingeri*. Three variables were measured on each insect:

- X_1 = Width of the 1st joint of the tarsus (legs)
- X_2 = Width of the 2nd joint of the tarsus
- X_3 = Width of the aedeagus (reproductive organ)

We have ten individuals of each species to make up training data. Data on these ten individuals of each species is used to estimate the model parameters which we will use in linear score function.

Our objective is to obtain a classification rule for identifying the insect species from these three variables.

Let's begin...

Step 1: Collect the training data. (described above)

Step 2: Specify the prior probabilities. In this case we do not have any information regarding the relative abundances of the two species. Without any information in order to help specify prior probabilities, equal priors are selected:

$$\hat{p}_1 = \hat{p}_2 = \frac{1}{2}$$

Step 3: Test for homogeneity of the variance-covariance matrices using Bartlett's test.

- Using SAS
- Using Minitab

Here we will use the SAS program insect.sas [2] as shown below:

```
options ls=78;
title "Discriminant Analysis - Insect Data";
data insect;
  infile "D:\Statistics\STAT 505\data\insect.txt";
  input species $ jointl joint2 aedeagus;
  run;
data test;
  input jointl joint2 aedeagus;
  cards;
194 124 49
;
proc discrim data=insect pool=test crossvalidate testdata=test testout=a;
  class species;
  var jointl joint2 aedeagus;
  run;
proc print;
  run;
```

Inspect! Launch SAS [3]

No significant difference between the variance-covariance matrices for the two species (L' = 9.83; d.f. = 6; p = 0.132) is found. Thus linear discriminant analysis is appropriate for the data.

Step 4: Estimate the parameters of the conditional probability density functions, i.e., the population mean vectors and the population variance-covariance matrices involved. It turns out that all of this is done automatically in the discriminant analysis procedure.

Step 5: The linear discriminant functions for the two species can be obtained directly from the SAS or Minitab output.

Species	Function
Ch. concinna	$\hat{d}_a^L(\mathbf{x}) = -247.276 - 1.417x_1 + 1.520x_2 + 10.954x_3$
Ch. heikertlingeri	$\hat{d}_b^L(\mathbf{x}) = -193.178 - 0.738x_1 + 1.113x_2 + 8.250x_3$

Step 6: We will discuss this step in Lesson 10.5.

Step 7: Now, consider an insect with the following measurements. Which species does this belong to?

Variable	Measurement
Joint 1	194
Joint 2	124
Aedeagus	49

These are responses for the first three variables. The linear discriminant function for species a is obtained by plugging in the values for these three measurements into the equation for species (a):

$$\hat{d}_a^L(\mathbf{x}) = -247.276 - 1.417 \times 194 + 1.520 \times 124 + 10.954 \times 49 = 203.052$$

and then for species (b):

$$\hat{d}_b^L(\mathbf{x}) = -193.178 - 0.738 \times 194 + 1.113 \times 124 + 8.250 \times 49 = 205.912$$

Add in a log of .5 to obtain the linear score function for species (a):

$$\hat{s}_a^L(\mathbf{x}) = \hat{d}_a^L(\mathbf{x}) + \log \hat{p}_a = 203.052 + \log 0.5 = 202.359$$

and then for species (b):

$$\hat{s}_b^L(\mathbf{x}) = \hat{d}_b^L(\mathbf{x}) + \log \hat{p}_b = 205.912 + \log 0.5 = 205.219$$

Conclusion

According to the classificaqtion rule the insect is classified into the species with the highest linear discriminant function. Because $\hat{s}_b^L(\mathbf{x}) > \hat{s}_a^L(\mathbf{x})$, we conclude that the insect belongs to species (b) *Ch. heikertlingeri.*

Of course, the addition of the log of .5 does not make any difference. Whether we classify on the basis of $\hat{d}_b^L(\mathbf{x})$ or on the basis of the score function, the decision will remain the same. In case the priors are not equal, this would not hold.

You can think of the priors as a 'penalty' in some sense. If you have a higher prior probability of a given species you will give it very little 'penalty' because you will be taking the log of a number close to one which is not going to subtract much. On the other hand, if there is a low prior probability, then the log of a very small number results in a larger reduction.

Note: SAS by default assumes equal priors. Later on we will look at an example where we do not assume equal priors - the Swiss Banks Notes example.

Posterior Probabilities

You can also calculate the posterior probabilities. These are used to measure uncertainty regarding the classification of a unit from an unknown group. They will give us some indication of our confidence in our classification of individual subjects.

In this case, the estimated posterior probability that the insect belongs to species (a) Ch. concinna given the observed measurements is:

$$\begin{aligned} p(\pi_a|\mathbf{x}) &= \frac{\exp\{\hat{s}_a^L(\mathbf{x})\}}{\exp\{\hat{s}_a^L(\mathbf{x})\}+\exp\{\hat{s}_b^L(\mathbf{x})\}} \\ &= \frac{\exp\{202.359\}}{\exp\{202.359\}+\exp\{205.219\}} \\ &= 0.05 \end{aligned}$$

This is a function of the linear score functions for the two species. Here we are looking at the exponential function of the linear score function for species (a) divided by the sum of the exponential functions of the score functions for species (a) and species (b). Using the numbers obtained earlier, this equals 0.05.

Similarly for species (b), the estimated posterior probability that the insect belongs to *Ch. heikertlingeri* is:

$$\begin{aligned} p(\pi_b|\mathbf{x}) &= \frac{\exp\{\hat{s}_b^L(\mathbf{x})\}}{\exp\{\hat{s}_a^L(\mathbf{x})\}+\exp\{\hat{s}_b^L(\mathbf{x})\}} \\ &= \frac{\exp\{205.219\}}{\exp\{202.359\}+\exp\{205.219\}} \\ &= 0.95 \end{aligned}$$

In this case we are 95% confident that the insect belongs to species (b). This is a pretty high level of confidence with a 5% chance that we might be in error in this classification. You would have to decide what is an acceptable error rate here. For classification of insects this might be perfectly acceptable, however, in some situations it might not be acceptable. For example, looking at the cancer case that we talked about earlier where we were trying to classify someone as having cancer or not having cancer, it may not be acceptable to have a 5% error rate. This is an ethical decision. It is a decision that has nothing to do with statistics and must be tailored to the situation at hand.

10.5 - Estimating Misclassification Probabilities

When an unknown specimen is classified according to any decision rule, there is always a possibility that the specimen is wrongly classified. This is unavoidable. This is part of the inherent uncertainty in any statistical procedure. One procedure to evaluate the discriminant rule is to classify the *training data* according to the developed discrimination rule. Because we know which unit comes from which population among the training data, this will give us some idea of the validity of the discrimination procedure.

Method 1. The *confusion table* describes how the discriminant function will classify each observation in the data set. In general, the confusion table takes the form:

	Classified As				
Truth	1	2	⋯	g	Total
1	n_{11}	n_{12}	⋯	n_{1g}	$n_{1.}$
2	n_{21}	n_{22}	⋯	n_{2g}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
g	n_{g1}	n_{g2}	⋯	n_{gg}	$n_{g.}$
Total	$n_{.1}$	$n_{.2}$	⋯	$n_{.g}$	$n_{..}$

Rows 1 through *g* are *g* populations to which the items truly belong. Across the columns we are looking at how they are classified. n_{11} is the number of insects correctly classified in species (1). But n_{12} is the number of insects incorrectly classified into species (2). In this case n_{ij} = the number belonging to population *i* classified into population *j*. Ideally this matrix will be a diagonal matrix; in practice we hope to see very small off-diagonal elements.

The row totals provide the number of individuals belonging to each of our populations or species in our training dataset. The column totals are the number classified into each of these species. The total number of observations in the dataset is *n...* The dot notation is used here in the row totals for summing over the second subscript, whereas in the column totals we are summing over the first subscript.

We will let:

$$p(i|j)$$

denote the probability that a unit from population π_j is classified into population π_i. These misclassification probabilities are estimated by taking the number of insects from population *j* that are misclassified into population *i* divided by the total number of insects in the sample from population *j* as shown here:

$$\hat{p}(i|j) = \frac{n_{ji}}{n_{j.}}$$

These are the misclassification probabilities.

Example - Insect Data:

From the SAS output, we obtain the following confusion table.

Classified As			
Truth	**a**	**b**	**Total**
a	10	0	10
b	0	10	10
Total	10	10	20

Here, none of the insects were misclassified! The misclassification probabilities are all estimated equal to zero.

Method 2: Set Aside Method

Step 1: Randomly partition the observations into two ”halves”

Step 2: Use one ”half” to obtain the discriminant function.

Step 3: Use the discriminant function from Step 2 to classify all members of the second ”half” of the data, from which the proportion of misclassified observations is computed.

Advantage: This method yields unbiased estimates of the misclassification probabilities.

Problem: This does not make optimum use of the data, and so, estimated misclassification probabilities are not as precise as possible.

Method 3: Cross validation

Step1: Delete one observation from the data.

Step 2: Use the remaining observations to compute a discriminant function.

Step 3: Use the discriminant function from Step 2 to classify the observation removed in Step 1. Steps 1-3 are repeated for all observations; compute the proportions of observations that are misclassified.

Example: Insect Data

The confusion table for the cross validation is

Classified As			

Truth	a	b	Total
a	10	0	10
b	2	8	10
Total	12	8	20

Here, the estimated misclassification probabilities are:

$$\hat{p}(b|a) = \frac{0}{10} = 0.0$$

for insects belonging to species A, and

$$\hat{p}(a|b) = \frac{2}{10} = 0.2$$

for insects belonging to species B.

Specifying Unequal Priors

Suppose that we have information (from prior experience or from another study) that suggests that 90% of the insects belong to *Ch. concinna*. Then the score functions for the unidentified specimen are

$$\hat{s}_a^L(\mathbf{x}) = \hat{d}_a^L(\mathbf{x}) + \log \hat{p}_a = 203.052 + \log 0.9 = 202.946$$

and

$$\hat{s}_b^L(\mathbf{x}) = \hat{d}_b^L(\mathbf{x}) + \log \hat{p}_b = 205.912 + \log 0.1 = 203.609$$

In this case, we would still classify this specimen into *Ch. heikertlingeri* with posterior probabilities

$$p(\pi_a|\mathbf{x}) = 0.36 \text{ and } p(\pi_b|\mathbf{x}) = 0.64$$

These priors can be specified in SAS by adding the ”priors” statement: priors ”a” = 0.9 ”b” = 0.1; following the var statement. However, it should be noted that when the "priors" statement is added, SAS will include log p_i as part of the constant term. In other words, SAS outputs the estimated linear score function, not the estimated linear discriminant function.

10.6 - Quadratic Discriminant Analysis

Linear Discriminant Analysis is for homogeneous variance-covariance matrices. However not all cases come from such simplified situations. Quadratic Discriminant Analysis is used for heterogeneous variance-covariance matrices:

$$\Sigma_i \neq \Sigma_j \text{ for some } i \neq j$$

Again, this allows the variance-covariance matrices to depend on the population.

Quadratic discriminant analysis calculates a Quadratic Score Function:

$$s_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{\Sigma_i}| - \frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})'\mathbf{\Sigma_i^{-1}}(\mathbf{x}-\mu_\mathbf{i}) + \log p_i$$

This is a function of population mean vectors and the variance-covariance matrices for the *i*th group. Similarly we will determine a separate quadratic score function for each of the groups.

This is of course a function of the unknown population mean vector for group i and the variance-covariance matrix for group i. These will have to be estimated from the training data. As before, we replace the unknown values of $\boldsymbol{\mu_i}$, $\boldsymbol{\Sigma_i}$,and p_i by their estimates to obtain the estimated quadratic score function as shown below:

$$\hat{s}_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S_i}| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_\mathbf{i})'\mathbf{S_i^{-1}}(\mathbf{x}-\bar{\mathbf{x}}_\mathbf{i}) + \log p_i$$

All natural logs are used in this function.

Decision Rule: Our decision rule remains the same as well. We will classify the sample unit into the population that has the largest quadratic score function.

$$\hat{s}_i^Q(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S_i}| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S_i^{-1}}(\mathbf{x}-\bar{\mathbf{x}}) + \log p_i$$

Let's illustrate this using the Swiss Bank Notes example...

10.7 - Example: Swiss Bank Notes

Recall that we have two populations of notes, genuine and counterfeit, and that six measurements were taken on each note:

- Length
- Right-Hand Width
- Left-Hand Width
- Top Margin
- Bottom Margin
- Diagonal

Priors

In this case it would not be reasonable to consider equal priors for the two types of banknotes. Equal priors would assume that half the banknotes in circulation are counterfeit and half are genuine. This is a very high counterfeit rate and if it was that bad the Swiss government would probably be bankrupt! We need to consider unequal priors in which the vast majority of banknotes are thought to be genuine. For this example let us assume that no more than 1% of bank notes in circulation are counterfeit and 99% of the notes are genuine. The prior probabilities can then be expressed as:

$$\hat{p}_1 = 0.99 \text{ and } \hat{p}_2 = 0.01$$

The first step in the analysis is going to carry out Bartlett's test to check for homogeneity of the variance-covariance matrices.

- Using SAS
- Using Minitab

To do this we will use the SAS program swiss9.sas [4] - shown below:

```
options ls=78;
title "Discriminant - Swiss Bank Notes";
data real;
  infile "D:\Statistics\STAT 505\data\swiss1.txt";
  input length left right bottom top diag;
  type="real";
  run;
data fake;
  infile "D:\Statistics\STAT 505\data\swiss2.txt";
  input length left right bottom top diag;
  type="fake";
  run;
data combine;
  set real fake;
  run;
data test;
  input length left right bottom top diag;
  cards;
214.9 130.1 129.9 9 10.6 140.5
;
  run;
proc discrim data=combine pool=test crossvalidate testdata=test testout=a;
  class type;
  var length left right bottom top diag;
  priors "real"=0.99 "fake"=0.01;
  run;
proc print;
  run;
```

[4]

SAS Notes

By default, SAS will make this decision for you. Let's look at the proc descrim procedure in the SAS Program swiss9.sas [4] that we just used.

```
proc discrim data=combine pool=test crossvalidate testdata=test testout=a;
  class type;
  var length left right bottom top diag;
  priors "real"=0.99 "fake"=0.01;
  run;
```

By including pool=test, SAS will decide what kind of discriminant analysis to carry out based on the results of this test.

If the test fails to reject, then SAS will automatically do a linear discriminant analysis. If the test rejects, then SAS will do a quadratic discriminant analysis.

There are two other options here. If we put pool=yes then SAS will conduct a linear discriminant analysis whether it is warranted or not. It will pool the variance-covariance matrices and do a linear discriminant analysis without reporting Bartlett's test.

If pool=no then SAS will not pool the variance-covariance matrices and perform the quadratic discriminant analysis.

SAS does not actually print out the quadratic discriminant function, but it will use quadratic discriminant analysis to classify sample units into populations.

Click on the arrow in the window below to see how discriminant analysis is performed using the Minitab statistical software application.

Performing discriminant analysis (Swiss bank notes data)

Bartlett's Test finds a significant difference between the variance-covariance matrices of the genuine and counterfeit bank notes (L' = 121.90; $d.f.$ = 21; $p < 0.0001$). The variance-covariance matrix for the genuine notes is not equal to the variance-covariance matrix for the counterfeit notes. Because we reject the null hypothesis of equal variance-covariance matrices, this suggests that a linear discriminant analysis is not appropriate for these data. A quadratic discriminant analysis is necessary.

Let us consider a bank note with the following measurements:

Variable	Measurement
Length	214.9
Left Width	130.1
Right Width	129.9
Bottom Margin	9.0
Top Margin	10.6
Diagonal	140.5

Any number of lines of measurements may be considered. Here we are just interested in one set of measurements. It is requested that this bank note be classified as real or genuine. The posterior probability that it is fake or counterfeit is only 0.000002526. So, the posterior probability that it is genuine is very close to one (actually, this posterior probability is 1 - 0.000002526 = 0.999997474). We are nearly 100% confident that this is a real note and not counterfeit.

Next consider the results of crossvalidation. Note that crossvalidation yields estimates of the probability that a randomly selected note is correctly classified. The resulting confusion table is as follows:

Classified As			
Truth	**Counterfeit**	**Genuine**	**Total**
Counterfeit	98	2	100
Genuine	1	99	100
Total	99	101	200

Here, we can see that 98 out of 100 counterfeit notes are expected to be correctly classified, while 99 out of 100 genuine notes are expected to be correctly classified.Thus, the estimated misclassification probabilities are estimated to be:

$$\hat{p}(\text{real} \mid \text{fake}) = 0.02 \text{ and } \hat{p}(\text{fake} \mid \text{real}) = 0.01$$

The question remains: Are these acceptable misclassification rates?

A decision should be made in advance as to what would be the acceptable levels of error. Here again, you need to think about the consequences of making a mistake. In terms of classifying a genuine note as a counterfeit, one might put an innocent person in jail. If you make the opposite error you might let a criminal go free. What are the costs of these types of errors? And, are the above error rates acceptable? This decision should be made in advance. You should have some prior notion of what you would consider reasonable.

10.8 - Summary

In this lesson we learned:

- How to determine which type of discriminant analysis is appropriate, linear or quadratic;
- How the linear discriminant function is used to classify a subject into the appropriate population;
- About issues to consider when selecting prior probabilities that a randomly selected subject belongs to a particular population;
- How to use posterior probabilities to assess the uncertainty of the classification of a particular subject;
- How to use crossvalidation and confusion tables to assess the efficacy of discriminant analysis.

Complete the homework problems that will give you a chance to put what you have learned to use.

Source URL: https://onlinecourses.science.psu.edu/stat505/node/89

Links:
[1] https://www.dynamicdrive.com
[2] https://onlinecourses.science.psu.edu/stat505/sites/onlinecourses.science.psu.edu.stat505/files/sas/insect.sas
[3] https://onlinecourses.science.psu.edu/stat505/../sas/insect.sas
[4] https://onlinecourses.science.psu.edu/stat505/sites/onlinecourses.science.psu.edu.stat505/files/sas/swiss9.sas