Chapter 18 Notes

Discriminant Analysis

Discriminant analysis is a classification method that seeks to classify data into known groups. It is sometimes also referred to as **segmentation**. We have seen methods for feature extraction (principal components analysis) and clustering, which in some ways classify data, but these have some fundamental differences from the classification done in discriminant analysis.

- In pca and cluster analysis, we generally do not know the groups or features we wish to identify.
- In pca, we are selecting features based on direction of maximal variation in the explanatory variables.
- In (hierarchical) cluster analysis, we sequentially group data elements based on proximity but we typically do not know how many clusters there should be or what those clusters represent (in the examples we looked at in class, we actually had more knowledge about the data and so could attempt to match up known groups, but this is not fundamental to the technique).

A couple of the fundamental differences in discriminant analysis are:

- We have data from known groups.
- We want to determine one or more functions based on explanatory variables to classify those groups.
- The classification is based on group means.
- Our goal is to be able to classify new data into one of the known groups based on the classification defined by discriminant function(s).

Here are a few terms that are useful to keep in mind:

- Discriminant functions – these are the function(s) that define our dividing lines in terms of the explanatory variables to tell us which group a point should be classified into.
- Prior probabilities – these tell us about the relative sizes of the known groups in the general population (if we have 2 groups, should we expect that the populations for those groups are the same size in general, proportional to the sample sizes, or some other proportion?).
- Posterior probabilities – these tell us the probability that an observation came from a particular group. This is based on the assumptions of the discriminant analysis, the assumed prior probabilities (the assumed proportions in the general population), and the estimated densities for the subpopulations. For each group there is some probability that an observation (with group unknown) was from that group. An observation (with group unknown) will be classified as coming from the group with largest posterior probability.
- Misclassification rate estimates – (usually) some number of observations will be placed in the wrong group. We can estimate what percentage will be incorrectly classified by using a method such as re-substitution, leave-one-out cross-validation, or (if we have a lot of data) we can

randomly separate the data into training and test sets and use the test set to estimate the misclassification rate.

- Training set – the set of observations the discrimination is based on.
- Test set – a set of observations that we do not include in the training set, but for which we know the group. We can classify these points based on the discrimination obtained with the training set to estimate how well the discrimination classifies "new" observations.

We will focus on linear and quadratic discriminant analysis, which assume multivariate normality. Logistic discriminants, nonparametric methods, and support vector machines are more advanced or more general techniques which can be employed if multivariate normality is not a reasonable assumption.

Linear Discriminant Analysis (LDA)

We can think of discriminant analysis as drawing lines in space to (in some sense) optimally separate groups. In the linear case with two groups, we would have a discriminant function

$$z = a_1 x_1 + a_2 x_2 + \ldots + a_p x_p$$

The vector $a = (a_1, a_2, \ldots, a_p)'$ maximizes the ratio of the between-groups variance of z to its within-groups variance. LDA carries the following assumptions:

- The data from each group have multivariate normal distributions.
- The covariances of each group are the same.

For two groups and assuming both groups are equally probable (e.g. have equal prior probabilities), we can compute z_i from the i^{th} observation and compare it with

$$z_c = \frac{\bar{z_1} + \bar{z_2}}{2}$$

If group 1 has smaller mean z value, a new observation is classified as being in group 1 if $z_i < z_c$ and in group 2 otherwise. In the case of more than 2 groups, there would be multiple (linear in the case of LDA) discriminant functions partitioning the space.

Quadratic Discriminant Analysis (QDA)

In quadratic discriminant analysis, we still assume multivariate normality but within-group covariances are no longer assumed to be equal and the discriminant functions are no longer linear, so QDA carries the following assumptions:

- The data from each group have multivariate normal distributions.
- The covariances of each group are not all the same.

The Discrim Procedure

The fundamental procedure of interest in SAS is **proc discrim**. It will allow us to perform a discriminant analysis with whatever variables we choose from the data. We will want to specify the classification variable via a **class** statement.

By default the procedure will give us classification rates based on re-substitution—all the data was used for the discrimination and then classification is done based on that discrimination. Our typical goal, though, is to classify new data based on known data, so a leave-one-out cross-validation (classifying each data point based on a discriminant analysis that does not include that particular data point) will tend to provide a more realistic estimate of misclassification rate, removing the influence of individual data points on their own classification if we have a small amount of data. We can use the **crossvalidate** option to perform this type of misclassification analysis. Using training and test sets will be even better still if we have enough data.

Examples & Exercises

We will start with a linear discriminant, then a quadratic discriminant to introduce the procedure and concept, and then include tests for multivariate differences of means and for equality of covariance to choose an appropriate discrimination method (assuming there exist differences of means across groups… if there is no difference of means, there's not much point in trying to discriminate groups).

Skulls Data Set

We will start with the **skulls** data set from Chapter 18 of the textbook. The goal is to classify skulls based on measurements of those skulls. Type A skulls were from local graves and assumed to be Mongolian or Indian. Type B skulls were from a battlefield and assumed to be Tibetan. The hypothesis was that the groups were of fundamentally different background and could be classified based on skull characteristics.

The variables in the data set are:

- **Length** – measurement of skull length
- **Width** – measurement of skull width
- **Height** – measurement of skull height
- **Faceheight** – measurement of face height
- **Facewidth** – measurement of face width
- **Type** – A (from local graves) or B (from battlefield)

Linear Discrimination with One Explanatory Variable

The simplest case is discrimination of 2 groups based on a single explanatory variable. In the Tibetan skulls example, let's try discriminating between type A and type B skulls based only on **length**.

Because we only have one discriminating variable and two groups it will be easier to visualize what the discrimination is telling us than it would be in general.

Example

Let's start with re-substitution error rates assuming equal prior probabilities and visualize what the posterior probabilities are telling us about discrimination of skull type based on skull **length**. We will look at:

- the classifications
- misclassification rates
- the posterior probabilities (via the **out** option)
- estimated densities (via the **outd** option) for the underlying populations

Cross-validation will be more realistic in general, and we can get cross-validation classification rates using the **crossvalidate** option.

Exercise

In the previous example, we assumed the underlying populations are equally likely. We could make other assumptions about the relative sizes of the populations. The **priors** statement allows us to modify our assumptions about relative population sizes.

- Add a **priors** statement that assumes the underlying populations are proportional to the sample sizes and repeat the previous analysis.
- How do the results change?
- How do they stay the same?

Probabilities for Iris Data with One Explanatory Variable

This is a simplified version of the **The DISCRIM Procedure>> Examples>> Univariate Density Estimates and Posterior Probabilities** example from the documentation. Here we extend to three groups (the iris species) classified by one predictor (petal width), and examine the results for both LDA and QDA (because we will find that the equal variance assumption is not very reasonable).

Probabilities for Iris Data with Two Explanatory Variables

For this example, we will look at **The DISCRIM Procedure>> Examples>> Bivariate Density Estimates and Posterior Probabilities** example from the documentation. We will now classify species by petal length and width. This should at least give a feel for how discriminant analysis generalizes to higher dimensions. Once we get beyond two predictors, it gets much more difficult to visualize the separating surfaces. Again we will look at LDA and QDA because we will see that the equal covariance assumption isn't very reasonable. We will also see an example of kernel density based discrimination.

The Stepdisc Procedure

The **stepdisc** procedure provides a stepwise method for picking out the important explanatory variables. We can think of this procedure as being akin to the **selection=stepwise** option we have seen for regression procedures.

Stepwise Selection of Explanatory Variables

If we use **stepdisc** we can choose a subset of explanatory variables. From this selected subset we can generate a classification with **discrim**. Again, we could set prior probabilities and compare the classification results based on equal and proportional prior probabilities.

For this example, we will go back to the **skulls** data set. Now, we will start with all 5 skull measurements and use stepwise selection to pick the significant predictors of skull **type**.

Validation with a Test Data Set

We can look at the **The DISCRIM Procedure>> Examples>> Linear Discriminant Analysis of Remote-Sensing Data on Crops** example from the documentation to see how training and test data can be incorporated to estimate misclassification rates.

Here we take the following steps:

- perform discrimination on a training data set
- output the results of that discrimination for the training set to a special data set via an **outstat=** option
- use the data set created in the previous step as a new input data set in **proc discrim** via the **data=** option and define a test set to be classified via the **testdata=** option
- a **testout=** option can also be added in the previous step to write the test classification information out like we did with the **out=** statement before