SAS Workshop - Multivariate Procedures
Handout # 5

Statistical Programs
College of Agriculture

PROC DISCRIM

In cluster analysis, the goal was to use the data to define unknown groups. In contrast, discriminant analysis is designed to classify data into known groups. Discriminant analysis is useful in automated processes such as computerized classification programs including those used in remote sensing. In order to conduct a discriminant analysis, a training data set is required. This data set is used to determine the combination (discriminant function) of the responses which best describes each group. Each observation is assigned a probability of belonging to a given group or class based on the distance of its discriminant function from that of each class mean.

In SAS the discriminant procedure is PROC DISCRIM. The general form is:

```
PROC DISCRIM <options>;
      CLASS <class var>;
      VAR <var1 var2 var3 ... var n>;
      PRIORS <options>;
```

As with other multivariate procedures, there are many possible procedure options. The more important ones are:

OUTSTAT = - saves the discriminant function for future classification of new data.

CROSSVALIDATE - classifies each training data point as if it were new. Provides estimates of classification accuracy and error.

CROSSLISTERR - prints any mis-classified observations from the CROSSVALIDATE option.

POOL = - determines whether to assume equal variances across classifications. POOL = TEST will test for the equivalency and use the results for the subsequent analysis.

If POOL=YES option is used, SAS assumes the variance (variance-covariance matrix of the responses) is the same across all classes. This results in a **Linear Discriminant Function**. When the POOL=NO option is chosen, each class has a unique variance structure and a **Quadratic Discriminant Function** is produced. Often the best choice is POOL=TEST which will give a statistical test for equal variance structure.

The **CLASS** statement lists the variable that represents the known classes or groups.

The **VAR** statement lists all relevant response variables and the **PRIORS** statement is used to specify the anticipated, e.g. prior, probabilities that observation belong to each group. Options for PRIORS statements are EQUAL and PROPORTIONAL. EQUAL probabilities provides no preference for any class and PROPORTIONAL uses probabilities equal to the proportion of observations in each class of the training data. If the user has no information on how observations should be classified, the EQUAL option is most appropriate.

Example

In this example, the complete set of flour data containing both cultivars is used. The discriminant procedure will be used to classify observations as either Ida Rose or Russet Burbank based on the six RVA measures. The SAS code to carry this out is listed below.

```
PROC DISCRIM CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=YES;
      CLASS CULTIVAR;
      VAR PEAK_VISC TROUGH_VISC FINAL_VISC BREAKDOWN TOTAL_SETBACK
            TIMEPEAK_VISC;
      PRIORS EQUAL;
```

In the example, the variances are pooled (linear discriminant function), cross validation will be done and the results placed in the dataset DIS_FUNC. I have also specified that the prior probabilities be equal for each class, i.e. probability = 0.5. The printed output is given below.

```
                     The DISCRIM Procedure

       Observations     450         DF Total               449
       Variables          6         DF Within Classes      448
       Classes            2         DF Between Classes       1


                    Class Level Information

           Variable                                          Prior
Cultivar   Name       Frequency       Weight   Proportion   Probability

IR         IR               225     225.0000     0.500000      0.500000
RB         RB               225     225.0000     0.500000      0.500000
```

The first section presents summary information on the number of observations, variables, and classes. The identification, proportion and prior probability of each class are also given.

A section listing the formulation for distance calculations is also provided, but omitted from this printout.

The third section gives the coefficients for the discriminant function of each class. These are the linear combinations of the responses that “define” each cultivar. A constant or intercept term is also included.

```
        Linear Discriminant Function for Cultivar

Variable         Label                  IR            RB

Constant                         -17.80777     -10.84606
Peak_Visc        Peak_Visc        -1.54603     -10.51420
Trough_Visc      Trough_Visc       0.90295       0.57117
Final_Visc       Final_Visc        0.66918       9.96298
Breakdown        Breakdown         1.53603      10.52007
Total_Setback    Total_Setback    -0.60204      -9.93633
TimePeak_Visc    TimePeak_Visc     3.00518       2.26983
```

The next section displays an error matrix giving the number of observations correctly classified and mis-classified. These are computed by simply running the training data back through the discriminant function to see how they get classified. Here we see that 190 Ida Rose observations, or 84%, were correctly classified leaving a 15% rate of error. Russet Burbank does better with a 6.6% error rate. The overall error rate is 11.11%

```
Number of Observations and Percent Classified into Cultivar

      From
      Cultivar           IR           RB        Total

      IR                190           35          225
                      84.44        15.56       100.00

      RB                 15          210          225
                       6.67        93.33       100.00

      Total             205          245          450
                      45.56        54.44       100.00

      Priors            0.5          0.5


            Error Count Estimates for Cultivar

                            IR           RB       Total

      Rate              0.1556       0.0667      0.1111
      Priors            0.5000       0.5000
```

The CROSSVALIDATION option provides a better assessment of classification accuarcy. This classification is also done for each observation, however, the discriminant function used in each case is constructed by taking that observation out of the data set. Thus, every data point is reclassified as if it were a new unknown observation. This provides a more conservative accuracy assessment. For these data, Ida Rose now shows an error rate of 16.6% while Russet Burbank is 7.1%. Overall, 11.7% of the observations were mis-classified. It would now be possible to add the CROSSLISTERR option to the SAS code to list the mis-classified observations. These could be examined to determine why they did not classify as expected.

```
Number of Observations and Percent Classified into Cultivar

      From
      Cultivar           IR           RB        Total

      IR                188           37          225
                      83.56        16.44       100.00

      RB                 16          209          225
                       7.11        92.89       100.00

      Total             204          246          450
                      45.33        54.67       100.00

      Priors            0.5          0.5


             Error Count Estimates for Cultivar

                            IR           RB       Total

      Rate              0.1644       0.0711      0.1178
      Priors            0.5000       0.5000
```

Discriminant analysis is most useful for classifying new, unknown data. If a new set of potato data were available as dataset NEW, the output from the previous run could be used to classify the NEW data using the code:

```
PROC DISCRIM DATA=DIS_FUNC TESTDATA=NEW TESTLIST;
      CLASS CULTIVAR;
```

The DATA=DIS_FUNC option utilizes the previously specified discriminant function and the TESTDATA=NEW option specifies the new data set to be classified. The TESTLIST option will print out each new observation and its classified value.