**Discriminant and Classification Analysis – Iris Data**

We chosen to perform Linear Discriminant Analysis on the Iris dataset. The iris dataset contains fifty observations from three species of Iris, Setosa, Versicolor and Virginica. Each is a multidimensional observation consisting of four variables, length and width of the iris’ sepals, and length and width of the iris’ petals.

Linear/Quadratic Discriminant Analysis aims to classify observations into known groups. In this case we have the 150 iris observations. This technique attempts to answer the question, supposing we have a new observation belonging to an unknown group, how do we determine this unknown group? That is, given that we have prior information about the correct classification of many previous observations, how should we use this knowledge to inform our new classification?

For linear discriminant analysis, we will do this by first supposing each group in our data come from a multivariate normal population with common covariance matrix. We will then estimate the population mean vector for each group as that group’s sample mean vector, and the common covariance matrix as the pooled covariance matrix across all groups. For a given observation vector, we finally classify it as belonging to the group which gives the smallest Mahalanobis distance to that group's mean vector. If instead the covariance matrices are not all equal, we perform Quadratic Discriminant Analysis, in which the Mahalanobis distance to the ith group now uses the ith group's covariance matrix instead of the pooled covariance matrix.

A common method for evaluating the performance of this classification technique is to perform leave one out cross validation. This is a method in which we train the model on all but one observation, and then test it on that left out observation, calculating a squared error. We then leave out a different observation and repeat this process for all observations. We can then average the squared errors to get an estimate of the test mean squared error. However, it would be even better to split the data into a training and test set, then train the model on the former and evaluate it on the latter. Since the test data are unseen to the model, this provides a more objective measure of the test error, and so we employ this method in addition to resubstitution and cross validation.

A test set was created from the Iris using stratified random sampling to extract 5 observations randomly from each of the three groups of flowers, Setosa, Versicolor and Virginica. The remaining observations constitute the training set and are used to construct the classification model. Since there are an equal number of observations per group, we will assume equal priors (that is, equal probability of belonging to a group, without any other information). Bartlett's test for homogenous covariance matrices provides strong evidence that the within group covariance matrices are unequal (Figure 3-1). Therefore, we should proceed with Quadratic Discriminant Analysis. From the "Generalized Squared Distance to species" table (Figure 3-2), we see that the distances between means for groups 2 and 3 are very close at about 5.55, compared to the distances between groups 1&2 and 1&3 which are 106.21 and 149.89 respectively. If we are judging these flowers purely by their sepal and petal dimensions, these generalized squared distances imply that the Setosa species is most easily distinguished from the Versicolor and Virginica species, since its mean vector has the farthest distances from these two. In contrast, the Versicolor and Virginica species are the two least distinguishable species, as their mean vectors are quite close to each other. This foreshadows that any misclassifications will most likely be between Versicolors and Virginicas. Looking at the classification output, this is exactly what each of the resubstitution, cross validation, and test tables show (Figures 3-3, 3-4, 3-5). The resubstitution table reports only one error, a Virginica misclassified as a Versicolor, yielding an overall error rate of only .74%. But we will not heed the resubstitution table as it is overly optimistic. The cross validation confusion table shows two errors: one Versicolor are misclassified a Virginica, and vice versa. This yields an overall cross validation error rate of 1.48%, and therefore an accuracy of 98.52%, which is very good. However, when we apply the model to the 15 unseen test data points, we get two Versicolors misclassified as Virginicas, yielding a test error rate of 13.33%.

Let us instead take a nonparametric approach, relaxing the assumption that the data from each group come from a multivariate normal distribution with its own group mean vector. That is, let us take the k nearest neighbors approach, in which we must specify some neighborhood around each data to judge by majority vote its group membership. Doing so with k=5, k=10, or k=15, we see that each of these choices of flexibility yield the same cross validation error rate of .74% and test error rates of 13.33% (Figures 3-6, 3-7, 3-8). Therefore, in any of these three cases, the accuracy is at least as great as the parametric approach. In addition, for each of these test errors, it is always 2 Versicolors that are misclassified as Virginicas, as was the case for the parametric approach. This suggests that these two misclassifications may be the same two pesky observations throughout all tests.

As a further analysis we can examine the discriminant functions that have the best separation amongst each other. For multivariate data, the coefficients turn out to be the eigenvectors of the matrix as it was defined for MANOVA. We will obtain such eigenvectors. The relative importance of these eigenvectors is determined by their corresponding eigenvalue proportions, that is . From Figure 3-9, we see that the first eigenvalue takes up 98.89% of the total proportion, so a single discriminant function sufficiently describes species separation. The choice is now to select between the standardized or raw eigenvectors; we could pick the former if the variables are commensurate, that is measured on the same scale and with comparable variances. The variables are indeed measured on the same scale, as they have the same units. However, the variances are not commensurate – for Sepal Length, Sepal Width, Petal Length, and Petal Width they are respectively 0.6856935, .1899794, 3.1162779, and 0.5810063. We observe that Petal Length has a dominating variance, for example being approximately 16 times the variance of Sepal Width. Therefore, we will use the standardized coefficients. SAS outputs these as the Pooled Within-Class Standardized Canonical Coefficients (See Figure 3-10), which simplifies to in accordance with Eq (8.17) of Rencher. We see that the variable Petal Length has the highest coefficient magnitude for the first discriminant function (Figure 3-11). Therefore, it contributes most to the separation of the groups.