**CSC334/424: Assignment #3 Total points: 65 for graduate students; 35 + 30 extra credit for undergraduate students. Due: Thursday, March 3rd, 2016 by 10pm. Note: Undergraduate students must attempt all problems and will be graded the same way as graduate students. However, a score of 35 of the 65 points will be considered full credit.**

Problem 1 (20 points): Data were collected on two species of flea beetles (a) Halticus oleracea and (b) Halticus carduorum. Measures of thorax length (THORAX), elytra length (ELYTRA), length of the second antennal joint (AJ2), and length of the third antennal joint (AJ3) in microns. These data are stored under beetle.txt. Perform a linear discriminant analysis between the two beetle species assuming equal population proportions.

1. Test for the equality of the variance-covariance matrices between the two species. What are your conclusions? **After running the LDA, the results below from the Box’s Test reveal that the significance value is high (0.586) and we can fail to reject the null hypothesis of equal variance-covariance (we want to fail to reject a value above 0.001). The log determinants (19.428, 19.567, and 19.768) are also fairly similar. We can proceed with the rest of the LDA analysis.**



1. Give the linear discriminant classification function for each of the beetle species. Under what condition would an unidentified specimen be classified as Halticus oleracea?



**Then, for each case, the function scores would be calculated using the following equation:**

**Score = -4.464 -0.093\*Thorax + 0.039\*Elytra + 0.024\*AJ2 + 0.037\*AJ3**

**An unidentified specimen be classified as (a or class 1.00) Halticus oleracea if the score value is <= -0.0975****as specified in the ‘Functions at Group Centriods’ table, which is used to establish the cutting point for classifying cases.**

**Cut Score = (-1.948 + 1.753) / 2 = -0.0975**

1. Suppose that an unidentified specimen with the following measurement is obtained:

|  |  |
| --- | --- |
| **Variable** | **Measurement** |
| **Thorax** | 184 |
| **Elytra** | 275 |
| **AJ2** | 143 |
| **AJ3** | 192 |

Which species would you classify this specimen into? **The insect is classified into the species that has the highest linear discriminant function corresponding to the cut-off values from the ‘Functions at Group Centriods’ table. According to the equation the unidentified species would be classified as Halticus oleracea (-0.315 <=-0.0975 == TRUE)**.

**-0.315 = -4.464 -0.093\*184 + 0.039\*275 + 0.024\*143 + 0.037\*192**

1. Give the apparent confusion matrix for the data. Estimate the percentage of beetles of each species that will be misclassified under the linear discriminant rule. **The apparent confusion matrix is below, with the calculated misclassification percentage of 2.6% as also indicated in the note a where the percentage of correctly classified groups is 97.4%**

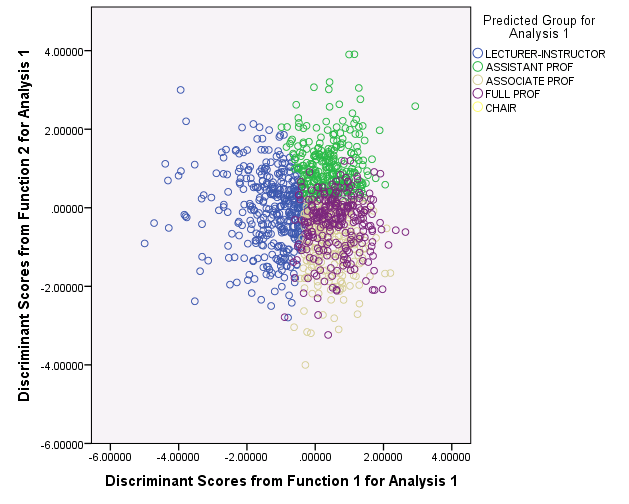


1. Give the cross-validation confusion matrix for the data. Estimate the percentage of beetles of each species that will be misclassified under the linear discriminant rule. **The cross-validation confusion matrix is below, with the calculated misclassification percentage of 7.9% as also indicated in the note a where the percentage of correctly classified groups is 92.1%**

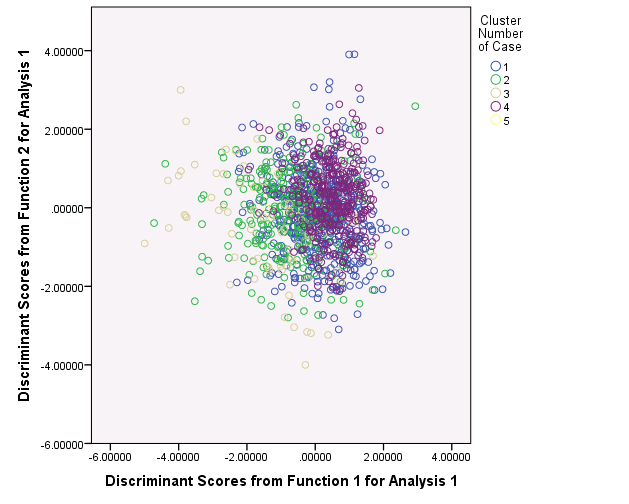


Problem 2 (15 points):

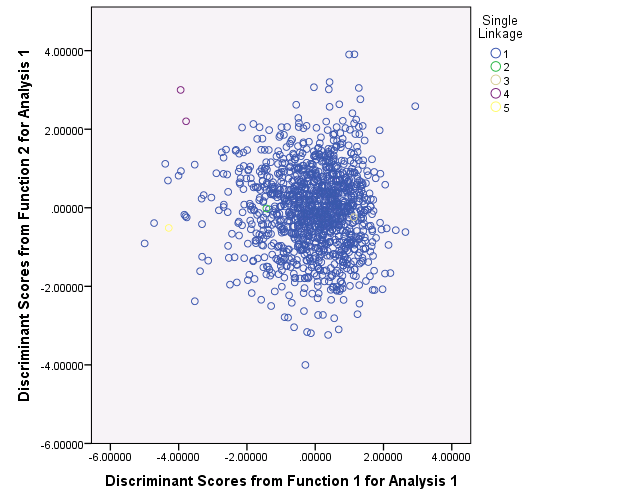
1. Both LDA and multiple linear regression can learn a function mapping multiple variables to predictions of values for another variable. What is the main difference in when they are used? **LDA assumes a linear relationship among predictors within each group, where MLR makes no up front assumptions about within group correlation. LDA also is used when the dependent variable is categorical/discrete where MLR the independent variable could be continuous or categorical/discrete.**
2. Both LDA and PCA can be thought of as finding an optimal vector onto which to project data. What is the difference between the techniques in terms of what criteria the vector is chosen to optimize? **LDA seeks to find a low-dimensional space such that when x is projected, classes are well-separated and maximizes the separation of means. PCA knows nothing of classes or natural groupings in the data and seeks maximize/optimize variance with the highest eigenvector.**
3. Briefly describe what is being optimized in Fisher’s Linear Discriminant (the optimization criterion function). **Fisher’s Linear Discriminant seeks to maximize the difference between the means, normalized by a measure of the within-class scatter and the between-class scatter. Then, we are looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible.**
4. Hierarchical clustering can allow you to determine the number of clusters after the clustering has already been completed – how? **Hierarchical clustering can be divided in agglomerative and divisive. Agglomerative (bottom up, clumping): starts with n singleton cluster and form the sequence by merging clusters. Divisive (top down, splitting): starts with all of the samples in one cluster and form the sequence by successively splitting clusters. From those methodologies you can determine afterwards with a dendrogram tree.**
5. Describe one advantage of DBSCAN over k-means and one advantage of k-means of DBSCAN. **DBSCAN supports outlier detection and is not very sensitive to noise, where K-means is unable to handle noisy data and outliers. K-means is more efficient than DBSCAN clustering because K-means takes less amount of time for the particular change of data in the database whereas DBSCAN takes much larger amount of time.**

Problem 3 (30 points): We will revisit the data on faculty from the second lab, now called faculty.sav. Recall that the PCA projection was not very helpful for visualizing the data. We will now consider a projection using LDA. Recall that you can get more than one discriminant vector. **Run LDA** with the same data as before (variables item13 through item24) and use faculty rank as the dependent variable. **Keep two discriminant vectors and save the scores of all data points on these discriminants. Plot the LDA projection by plotting those new score variables.**

1. Run k-means clustering and use the cluster assignment to color the points.



1. Run hierarchical clustering and use the cluster assignment to color the points.



1. Compare the k-means clustering to the correct labels. **Below is the first 6 samples of comparing the k-means clustering assignment versus the correct labels for faculty rank. An excel if statement was developed to calculate a mismatch between the cluster assignment and the total correct and mismatch were summed. The k-means cluster technique did not proficiently classify the data set (as indicated by the 1149 mis-matched labels).**



Figure : Source contains full table

1. Compare the k-means clustering to the hierarchical clustering. What part of the process of hierarchical clustering accounts for the sparse outlying cluster?  **Below is the hierarchical cluster assignment compared to the correct labels for the faculty rank. The hierarchical performed much better at correctly classifying the labels of faculty rank as seen by the correct amount of 316 versus 279 for k-means. The hierarchical does not perform well with outliers, so it has seemed to classified some of the furthest points as completely different classes. The analysis also only sweeps through the data once, so mistakes made early on cannot be corrected. It is also with merit to say that the independent variables chosen to identify the faculty rank are only very loosely correlated with Faculty Rank (the LDA assumes a linear relationship between the variables and the significant value was not a “Fail to reject” as we would have liked).**

****

Include the plots in your answer.