CSC334/424: Assignment #3
Due: Thursday, March 3rd, 2016 by 10pm
Total points: 65 for graduate students; 35 + 30 extra credit for undergraduate students

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.

i. Test for the equality of the variance-covariance matrices between the two species. What are your conclusions?
ii. Give the linear discriminant classification function for each of the beetle species. Under what condition would an unidentified specimen be classified as Halticus oleracea?
iii. 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?

iv. 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.
v. 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.

Problem 2 (15 points):

(a) 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?

(b) 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?

(c) Briefly describe what is being optimized in Fisher's Linear Discriminant (the optimization criterion function).

(d) Hierarchical clustering can allow you to determine the number of clusters after the clustering has already been completed – how?

(e) Describe one advantage of DBSCAN over k-means and one advantage of k-means of DBSCAN.

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.

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

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

c. Compare the k-means clustering to the correct labels.

d. Compare the k-means clustering to the hierarchical clustering. What part of the process of hierarchical clustering accounts for the sparse outlying cluster?

Include the plots in your answer.