Assignment 4 – Linear Discriminant Analysis

Question 1.

Using the multivariate data in the file fld1.xlsx:

1. determine the discriminant line found by Fishers Linear Discriminant.
2. Plot both the data and the discriminant line on a scatter plot
3. Using this line, determine the class of each of the data points in the dataset, assuming that the threshold is 0 (i.e. positive values are in one class and negative values in the other).
4. Determine what percentage of data points are incorrectly classified.

NOTE: The first 2 columns in fld1.xlsx are data columns. The third column is the class to which each data point belongs.

Question 2.

Using the multivariate data in the file spam.xlsx, determine the discriminant line found by Fishers Linear Discriminant. Using this line, determine the class of each of the data points in the dataset, assuming that the threshold is 0.

NOTE: The first 57 columns in spam.xlsx are data columns. The 58th column is the class to which each data point belongs.

Determine what percentage of the data points are incorrectly classified. You will notice that most of the data in the first class is classified correctly while the data in the second class is not. Therefore, it makes sense to adjust the threshold. Try the classification again a few times while adjusting the threshold so that it is a small negative number to see if you can improve the overall classification error rate (percentage of errors in BOTH classes).

FYI, information about the dataset is given in the folder with the spam dataset. This is a real dataset, from the Machine Learning Repository at UCI (University of California Irvine). I simply made it a bit smaller by only using the first 500 data points from the first class and the first 500 data points from the second. The original database has over 4000 data points and it was awkward to work with that in Excel. I have not, however, reduced the number of attributes. For those that are interested in a real-life example, this is one. It is a dataset with attributes used to try to detect spam from email.