**Question 1**

I used K-means because of its simplicity. The trade offs are sensitivity to outliers and initial seeds. The other options are hierarchical clustering or ANN (Competitive Learning) which are more complex.

To identify the number of clusters I’ve plotted the K-Means distortion vs. number of clusters below. As the graph shows after 3rd cluster the distortion level does not decease much as we increase the number of clusters, therefore in first sight 3 clusters look reasonable to start with.



This graph shows the 3 clusters using K-Means and the plot is 2-D using A1 & A2



Doing PCA reveals that just using 2 of the features explains about 96% of variance.



Next I used linear regression to draw the decision regions using PC1 and PC2 with actual labels and cluster assignments.



**Question 2**

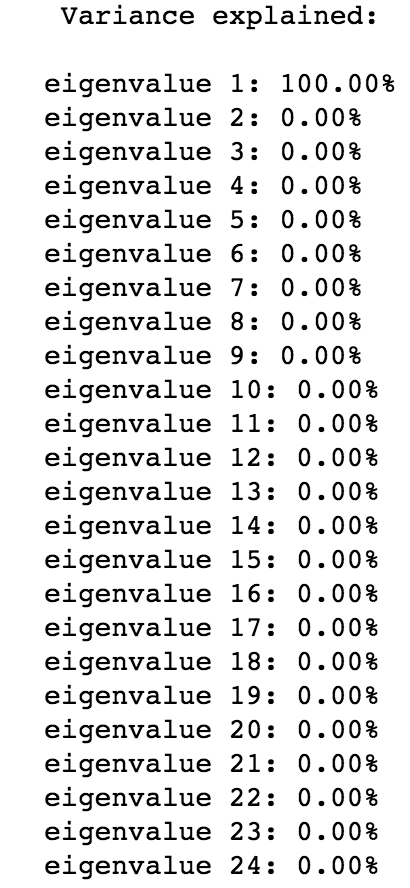
I checked for any missing feature values, but there is no missing data.

Using PCA all variables have comparable variances so PCA is not much useful.



Next I used LDA to look at all variables not only based on the variance but also how much they contribute to class-separation. The LDA results showed that 1 Eigen pair is by far the most informative one so we won’t loose much information if we use a 1-D feature space.

Variances explained using LDA:



Graph of LDA variable shows the “0” classes at one end of the 1-D line and “1” classes at the other end.



I also draw the ROC curves for 3 other different methods and with a 10-fold cross validation using all features. The result show that logistic regression performs better that the other methods with an AUC of 0.81 which is not appealing much.

