Data Science Lab 1

ENGSCI255 Semester 1, 2018

This lab is due by 5pm on Friday 4 May. For each question hand in the commands you used and the output generated.

1. Perform an exploratory analysis of the data in the file `titanic.csv` in R, on Canvas (using, for example, box and whisker plots, scatter plots, tables)

2. Set the seed of the random number generator to 99

   (a) Use k-means clustering to cluster the data into two groups using only the `age` and the `fare` attributes, performing 20 repetitions.

   (b) Generate a scatter plot of the clusters, and comment on how the two clusters are separated.

   (c) Suppose a passenger (not in the data set) is 30 years old and their fare was $20, which of the two clusters would they be in. Based on this clustering, estimate the probability that they survived.

   (d) Cluster the data again using some other attributes and comment on how well the clustering groups survivors together. You should submit and discuss two additional clusterings.

3. Set the seed of the random number generator to 50, and then generate a training data set of 250 people (the remaining data will be the test set).

   (a) Using only `sex` as the independent attribute create a classification tree based on the training data, and visualise the tree.

   (b) Predict the `Survived` property of the people in the test set, creating a table showing the performance of the classification tree.

   (c) Create four other classification trees based on the training data using different attributes, and different stopping criteria.

   (d) Evaluate the performance of each on the test set. Comment of the similarities and differences between the trees, and discuss why some trees may perform better than others.