SMM634 Coursework 2 2018/2019

Deadline 22 March 2019

NAME:

A sample of 12 riding-lawnmower owners and 12 non-owners is sampled from a city and the income in thousands of dollars and lot size in thousands of square feet are recorded. A riding-mower manufacturer wants to see if these two variables adequately separate owners from non-owners.The dataset, `lawnmower.dta`, is available on Moodle.

1. Produce a scatterplot of income (`income`) and lot size `lotsize` by the owner variable (`owner`). What do you notice? [Marks: 4]

2. Use LDA on these data and obtain a confusion matrix. What conclusions can you draw? [Marks: 4]

3. Employ now LDA using only a subset (training sample) of the observations (you could choose randomly 20 observations out of the sample size) and then examine how well it predicts the held out data (test sample). Comment on your findings. [*Hint*: in order to randomly choose 20 observations out of 24 you can use function `sample()` in R.] [Marks: 6]

4. Now fit a QDA model to the `lawnmower` data and compare your results with those in (c). [Marks: 3]

5. As a general question, assume that a large international air carrier has collected data on employees in three different job classifications: 1) customer service personnel, 2) mechanics and 3) dispatchers. The director of Human Resources wants to know if these three job classifications appeal to different personality types. Each employee is administered a battery of psychological tests which include measures of interest in outdoor activity, sociability and conservativeness. Write down the R syntax to fit a LDA on these data. [Marks: 3]