STAT40150 Assignment 1

Assoc. Prof. C. Gormley

2019/2020

• **There is a total of 80 marks for this assignment and it is worth 15% of your final module mark.**

• **Answer all questions and carry out all analyses in R.**

• **Due date: end (Sunday 11pm) of week 5. Please submit your assignment by uploading one pdf file to BrightSpace containing your answers to the questions, with the (fully commented) R code (if required) embedded within each question's solution. It should be possible for the grader of your assignment to copy and paste this code into R, thereby reproducing your work.**

• **Late assignments will be graded but will score 0.**

• **Please include your student name, student number and module code at the top of your assignment.**

• **Your assignment should include solutions to the questions posed below; these solutions may include written/typed responses, output and plots from R. All code should be fully commented within your answers to clearly illustrate how you arrived at your solution for each question. NB: where relevant if R code is not provided, marks will not be awarded.**

• **Any plots included should be clearly labelled.**

• **While discussion of the problems is encouraged plagiarism is not permitted. Anyone found to have been involved in plagiarism will score 0.**

Before beginning this assignment open a new R script and use the `set.seed` function to set the seed to your student number. Prior to working with the Glass and `fgl` data sets used below, randomly generate a number between 1 and n (where n is the number of rows in the data set), and delete that observation/row from the data set. Ensure that you include the code used in these steps in the R code associated with each question.

Question 1

The Glass.csv dataset (posted on Brightspace) examines the chemical compositions of 15th-17th century archaeological glass vessels excavated in Antwerp. The data record the concentrations of 13 elements, determined by x-ray methods, in 180 glass vessels. The data consist of measurements on four different types of glass (detailed in the `Group` variable in the data set).

In all questions, provide details of your reasoning and fully commented R code.

1 (a) (i) Compute the correlation matrix R of the 13 chemical elements using the fact that $R = D^{-1/2}SD^{-1/2}$ where S is the sample covariance matrix and $D^{-1/2}$ is a diagonal matrix with the inverse of each variable's standard deviation on the diagonal. Ensure your calculation is correct by showing that your result and that computed by the `cor` function in R are equal. [5 marks]

1 (a) (ii) Using R, determine the first two eigenvalues and eigenvectors of the covariance matrix S. Using R, show that they are indeed eigenvalues and eigenvectors of S. [4 marks]

1 (a) (iii) Using R, verify that the first two eigenvectors are orthonormal. [2 marks]

1 (a) (iv) Compute the variance of each element and produce a suitable summarising plot. Would you advise standardizing the glass data prior to analysis? Explain your reasoning. [4 marks]

1 (b) Suppose $\mathbb{E}[X_1] = 5$ and $Var[X_1] = 6$. Suppose $\mathbb{E}[X_2] = 8$ and $Var[X_2] = 7$. Suppose also that the covariance between X_1 and X_2 is 2.5.

1 (b) (i) In R, calculate the expected value and variance of $X_2 - X_1$. [2 marks]

1 (b) (ii) In R, calculate the correlation of $U = X_2 - X_1$ and $V = X_2 - 2X_1$. [3 marks]

TOTAL: 20 MARKS

Question 2

Return to the same dataset explored in Question 1 and consider the thirteen elements.

2 (a) In R, cluster the glass vessels using k-means clustering. How many clusters would you suggest are present in this data set? Detail any decisions you make when running this procedure. Provide details of your reasoning and fully commented R code. [5 marks]

2 (b) Cluster the glass vessels using hierarchical clustering. Detail any decisions you make when conducting the hierarchical clustering. From the dendrogram, how many clusters would you suggest are present in this data set? Cut the dendrogram at the desired number of clusters. [5 marks]

2 (c) Examine the cross tabulation of the cluster solutions obtained in (a) and (b) and then quantitatively compare them using an appropriate measure. Comment on the agreement between the two solutions. Provide details of your reasoning and fully commented R code. [3 marks]

2 (d) Create a pairs plot of the data, highlighting vessels from different glass types using colour and/or plotting characters. Comment on the relative size of the first glass type group and on the distribution of the `PbO` variable. Explore the impact of removing these data from your analysis. Why would detecing such issues be challenging in a truly unsupervised and high-dimensional setting? [7 marks]

TOTAL: 20 MARKS

Question 3

3 Load the MASS library in R, and its fgl dataset. Use the help file to understand what the data set contains. Decide whether or not you need to scale the data. **Write your own function** to calculate the misclassification error for linear discriminant analysis applied to the first 9 variables in this dataset, using the final variable as the known class. Split the data such that floor(n*(2/3)) observations are in the training set and the remainder in the test set, compute the misclassification rate, and repeat this 100 times. Create a suitable plot to illustrate the misclassification rates for each class and the overall misclassification rates. What is the average overall misclassification rate? [30 marks]

TOTAL: 30 MARKS

Question 4

Assume it is known that each of a set of observations belong to one of two equally probable classes. The number of variables p recorded on each observation x is 1 and it is assumed that the density of the data $f_k(x)$ in class k is Gaussian with mean μ_k and variance σ^2. Show that the Bayes' decision boundary corresponds to the point where

$$x = \frac{\mu_1 + \mu_2}{2}.$$

TOTAL: 10 MARKS