

School of Computing, Engineering and Mathematics

Lab-Based Final Exam – Introduction to Data Science

Session 02, 2019

|  |  |
| --- | --- |
| Complete your details in this section when instructed by the Exam Supervisor at the start of the exam.  You should also complete your details on any extra answer papers provided. | |
| STUDENT ID: |  |
| STUDENT FIRST NAME: |  |
| STUDENT SURNAME: |  |

|  |  |  |  |
| --- | --- | --- | --- |
| UNIT NUMBER: | 301033 | | |
| UNIT NAME: | Introduction to Data Science | | |
| QUESTIONS FORMAT: | Word processed document in PDF format; logically presenting answers to each question incorporating R outputs including graphs and charts. | | |
| WEIGHT: | Total exam marks: **60** - **50%** of total assessment. | | |
| UNIT CO-ORDINATOR: | Dr. Liwan Liyanage | | |
| LECTURER: | Ms. Prathayne Nanthakumaran, Mrs. Lakmini Wijesekara | | |
| TIME ALLOWED: | 2 Hours | TOTAL PAGES: | 3 |

|  |
| --- |
| **Final Exam INSTRUCTIONS**  Please note that you are expected to answer the questions clearly. Give the R commands, outputs, comments and discussion clearly and logically. Once completed submit the answer scripts as a PDF via TurnItin link within vUWS site. You also need to show the file you uploaded before leaving the examination room and show the document to the Supervisor.  **Resources**:  Open book. Students are allowed to use any material related to the subject Lecture notes and practical notes available on vUWS. Summaries and handy hints given in class or done by yourself; useful links and readings listed and uploaded in vUWS. |

# Question 1 (3 + 3 + 3 + 3 + 3 = 15)

This Question uses the data set “*glass.csv*”. The data represents the weight percent in corresponding oxides and refractive index together with the type of glass (window glasses and non-window glasses).

The *glass.csv* variables as follows:

RI = Refractive Index

Na = Sodium

Mg = Magnesium

Al = Aluminium

Si = Silicon

K = Potassium

Ca = Calcium

Ba = Barium

Fe = Iron

Type = Type of class

1 – Window glasses

0 – Non-window glasses

1. Calculate the mean and the variance for each appropriate variable and discuss if scaling is necessary and justify your findings.
2. Apply scaling and derive the principal components. R code and output should be clearly stated.
3. Explain the percentage variation captured by each principal component. Use and give relevant R output and support your findings.
4. Write the first two principal components in terms of the original variables in the given dataset.
5. Using Biplot, explain the association between original variables in the dataset.

# Question 2 (3 + 2 + 3 + 3 + 2 + 2 + 3 + 2 = 20)

This Question uses the data set “glass.csv” used in Question 1

* 1. Use K Means Clustering method and identify two clusters with K=2.
  2. In order to visually display the two clusters obtained in part a, plot the first two principal components obtained in question 1 and colour according to the k-means classes.
  3. Construct the misclassification table and calculate the misclassification rate. Discuss the accuracy of classifying window and non-window glasses when using k-means clustering.
  4. Alternatively perform Hierarchical Clustering on glass dataset. [Consider Euclidean distance as the dissimilarity measure and the closest distance between two clusters as the maximum distance between them]
  5. Display the dendrogram and cut it at a height that results in two distinct clusters.
  6. In order to visually display the two clusters obtained in part e, plot the first two principal components obtained in question 1 and colour according to the cluster profiles in part e.
  7. Construct the misclassification table and calculate the misclassification rate. Discuss the accuracy of classifying window and non-window glasses when using hierarchical clustering
  8. Compare results obtained in parts “b and c” with parts “f and g” and justify most appropriate clustering method to classify the dataset as window glasses and non-window glasses.

# Question 3 (3 + 4 + 2 + 2 + 3 + 2 + 4 + 4 + 1 = 25)

This Question uses the data set “*Wine\_Quality.csv*”.

1. Identify and state the Quantitative and Qualitative variables in the given dataset.
2. Construct the box plot for Volatile Acidity, Alcohol, Residual Sugar, Sulphates and pH on the target variable “*WineQuality*” and interpret your findings.
3. Build a logistic regression model to classify the “*WineQuality*” in terms of Volatile acidity, Alcohol and Residual Sugar. (No need to prove the significance of the model)
4. Build a logistic regression model to classify the “*WineQuality*” in terms of Volatile acidity, Alcohol, Residual Sugar, Sulphates and pH.
5. Considering the significance of the parameter estimates, select and state the best model out of the models obtained in part c and d to classify the quality of wine. Clearly, explain your answer with proper justification.
6. Write down the equation to calculate the probability of getting low wine quality for a given set of predictors using part e.
7. Classify the observations with probability > 0.6 as “low” and “high” otherwise. Hence, calculate the misclassification matrix and misclassification rate and comment on your results.
8. Calculate four other measures such as True Positive Rate, False Positive Rate, True Negative Rate and False Negative Rate that can be used to assess the accuracy of the model.
9. Mention another supervised learning method that can be used to solve the given problem.

- END -