

School of Computing, Engineering and Mathematics

Lab-Based Final Exam – Introduction to Data Science

SESSION 1, 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: | Mrs. Lakmini Wijesekara, Ms. Prathayne Nanthakumaran | | |
| 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, analysis, comments and discussion clearly and logically. Once completed submit the answer scripts via TurnItin link within vUWS site. You also need to show the file you uploaded before leaving the examination room.  **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 + 2 + 3 + 2 + 2 + 3 = 15)

This Question uses the data set “Envdata”. The data represents the pollution conditions and maximum wind speed together with the prevalence of Asthma (present or absent of Asthma) associate with several patients in Victoria State.

The Envdata are given for each patient under investigation as follows:

X1=co

X2=so2

X3=o3

X4=ppm10

X5=no2

X6= maxwindspeed

Y = asthma

* 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 and colour according to the k-means classes.
  3. Construct the misclassification table and misclassification rate and discuss the accuracy of predicting presence of Asthma.
  4. Alternatively use K Means Clustering method to identify three clusters using K=3.
  5. In order to visually display the three clusters obtained in part d, plot the first two principal components and colour according to the k-means classes.
  6. Compare results obtained in parts “a and b” with parts “d and e” and justify most suitable number of clusters for this data set using total within cluster variation and total between cluster variation.

# Question 2 (2 + 4 + 3 + 2 + 4 + 2 + 3= 20)

This Question uses the data set “Envdata” used in Question 1

1. Calculate the mean and the variance for each variable and discuss if scaling is necessary and justify your findings.
2. Apply scaling and derive the principal components. (R code and output)
3. Give the Scree Plot and give the percentage variation captured by each principal component.
4. Select the number of principal components most suitable to represent the dataset and justify your answer.
5. Derive and give the principal component loading vectors for the given dataset and explain the results/output.
6. Give the first two principal components using loading parameters obtained in part e.
7. Construct the Biplot and interpret it in terms of original variable contributions.

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

This Question uses the data set “Admission”.

This dataset contains the following variables.

*Serial. No*: observation number

*GRE* : Graduate Record Examinations score (out of 340)

*TOEFL* : Test of English as a Foreign Language Scores (out of 120)

*Uni\_R* : University rating (out of 5)

*SOP* : Statement of Purpose score (out of 5)

*CGPA* : Undergraduate GPA (out of 10)

*Chance* : Chance of admission

1. Construct the matrix plot and correlation matrix and comment.
2. Derive a multiple linear regression model to describe the “Chance of admission” in terms of other numeric variables and give the estimated model. (No need to prove the significance of the model)
3. Describe the model accuracy (Not the model assumptions)
4. Select one suitable explanatory variable and derive the best polynomial regression model and give the estimated model to describe “Chance of admission”. (No need to prove the significance of the model)
5. Describe the model accuracy of the polynomial model. (Not the model assumptions)
6. Improve the model by combining the two models derived from parts b and d and obtain the best final model. (No need to prove the significance of the model)
7. Test the significance of the best model obtained in part f (Show all steps in the significance test when testing for one parameter and describe logically the significance of other parameters briefly).
8. Test the model assumptions of the final model obtained in part f. (Using 4 default diagnostic plots)
9. Describe the model accuracy of the Final model using all the above findings.

- END -