DSA1101 Final 2310

Introduction to Data Science (National University of Singapore)

NATIONAL UNIVERSITY OF SINGAPORE

DSA1101 INTRODUCTION TO DATA SCIENCE

(Semester 1 : AY 2023/2024)

Time Allowed: 2 Hours

INSTRUCTIONS TO STUDENTS

1. Please indicate only your student number on your answer file. **Do not indicate your name**.
2. This exam paper contains **FOUR (4)** problems and comprises **SIX (6)** printed pages (including the cover page).
3. The data files used for this exam are given in the folder "For Finals" on Canvas/DSA1101/Files.
4. This is an OPEN BOOK exam. You may refer to your lecture notes, tutorials, textbooks or any notes that you have made, but you are not allowed to perform any online search.
5. Both programmable calculators and non-programmable calculators are allowed.
6. Use **set.seed(2811)** for questions in R.
7. Students are required to answer ALL questions. Total mark is 100.
8. During the exam, you are not allowed to communicate with any person other than the invigilators.
9. At the end of the exam, label your answer file by your **student number** and upload it to Canvas/DSA1101/Assignments/Final Exam Submission.

1. (*10 points*) **True or False**

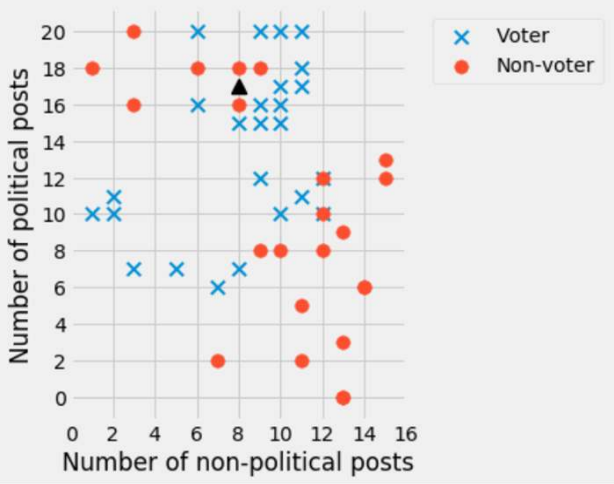
a. (*2 pts*) If we use linear regression to predict response y based on regressor x for n observations, the average of our residuals, $(1/n)\sum_{i=1}^{n} e_i$ will always be zero.

b. (*2 pts*) In order to build a k-nearest neighbors classifier, you do not need to know the class label (response) of any of the training observations.

c. (*2 pts*) A classifier is considered to be overfitting if it performs very well on the training set, but not very well on the test set.

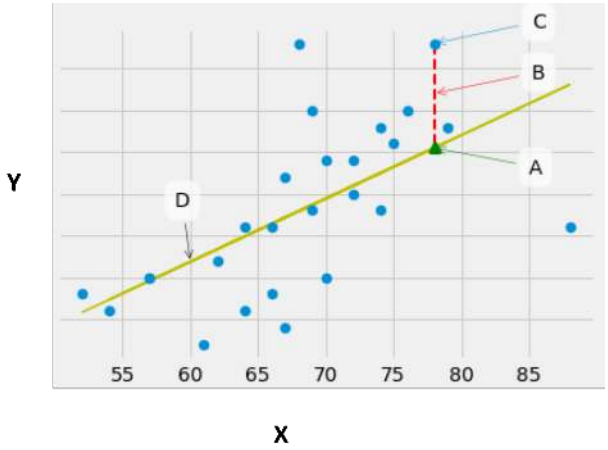
d. (*2 pts*) If I am using House Price and Household Monthly Income (both in Singapore dollars) as two features for my KNN classifier, I do not need to standardize them since they have the same units.

e. (*2 pts*) Candidate A decides to train a classifier to predict whether people will vote in the 2020 U.S. election or not. He gathered data on voting records from the 2018 U.S. election and decides to use two features: the number of political and non-political posts on social media that a person made in the month leading up to the election. A scatter plot of his data is shown below. The candidate is trying to classify the point at (8, 17) shown as a triangle on the graph. **If he uses a 3-nearest neighbor classifier, the classification will be "voter".**



2. (*10 points*) Muilple Choice Questions

A linear regression model was built using a dataset where Y is the response and X is the regressor. The visualization with certain aspects labeled A, B, C, or D is given below. Match each term below to its label on the graph, or "Not Pictured". Some letters may be used multiple times or not at all.



a. (*2 pts*) Predicted value

A *B* *C* *D* Not Pictured

b. (*2 pts*) Residual

A *B* *C* *D* Not Pictured

c. (*2 pts*) Line of Best Fit

A *B* *C* *D* Not Pictured

d. (*2 pts*) Intercept

A *B* *C* *D* Not Pictured

e. (*2 pts*) Observed Value

A *B* *C* *D* Not Pictured

3. (*70 points*) A research study investigated Y = whether a patient having surgery with general anesthesia experienced a sore throat on waking (response variable) as a function of D = the duration of the surgery (in minutes) and T = the type of device used to secure the airway. The dataset is given in a file named `data1-finals.csv`. The table below gives the description for each variable in this study.

Variable	Description
Patient	patient's identity number
Y	0=no; 1=yes
T	0=laryngeal mask airway; 1=tracheal tube
D	duration of the surgery (in minutes)

For the questions below,

- load the dataset into R and name it as **data1**.
- report numerical answers to three significant figures if it's smaller than one and to three decimal places if it's larger than one.

Part I: Logistic Regression Model (35 points)

1. Write code to form a logistic regression model (called M1) for response Y. Write down the fitted model and explain in detail any notations used.
2. Report any regressor that is not significant at significant level 0.1.
3. Report the coefficient of the variable D, duration of surgery, in model M1. Interpret it in the context of this study.
4. Report the coefficient of the variable T, the type of device used to secure the airway, in model M1. Interpret it in the context of this study.
5. Write code to plot the ROC curve of model M1. Derive and report the value of AUC.
6. Let δ denote the threshold used for a classifier based on the probability derived from model M1. Write code to plot a figure to show how the TPR and the FPR of the classifier change when the threshold δ changes.
7. Assume TPR is prioritized over FPR, find the value of δ that gives the best TPR as long as FPR is not larger than 0.5.

8. Two patients will have surgery with the estimated time for duration (D) and the type of device used to secure the airway (T) listed below. Write code to predict the probability that the patient will experience sore throat upon waking up from the surgery.

   Patient A: D = 80, T = laryngeal mask airway;

   Patient B: D = 125, T = tracheal tube.

Part II: Naive Bayes Classifier (15 points)

9. We now use the naive Bayes classifier for the dataset given. Write code to form the classifier, named as M2.

10. Calculate the accuracy of M2 on the given dataset.

11. Using the classifier M2, predict the probability of having sore throat after surgery for each patient listed in Question 8. Report the probabilities.

Part III: Decision Trees (20 points)

12. Write code to form a decision tree (called M3) to predict if a patient has sore throat upon waking up after a surgery, with `minsplit = 4`, where variable selection and split points are based on information gain.

13. Among the features used to form the tree, which one is the most important?

14. Calculate the accuracy of M3 on the given dataset.

15. Using the classifier M3, predict the status of having sore throat after surgery for each patient listed in Question 8. Report the results.

4. (*10 points*) Consider the famous Iris Flower Data set which was first introduced in 1936 by the famous statistician Ronald Fisher. This data set consists of observations from flowers of Iris species. For each observation, four features were measured: the flower's length and width of the sepals and petals (in cm).
The data set is given in a file named `data2-finals.csv`.

1. Use K-means clustering method to cluster all the flowers into k groups where $k = 1, 2, 3, ..., 10$, where for each value of k the value of WSS - the within sum of squares is obtained.

2. Write code to obtain the plot of WSS against k. Which value of k would you choose as the number of clusters for all the observations in the data set? Explain.

3. With the value of k chosen above, report the centroids of all the clusters and the number of the observations in each cluster.

–END OF PAPER–