2425S2 midterm

Introduction to Data Science (National University of Singapore)

NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 2 AY2024/2025)

Midterm Test

Time Allowed: 80 minutes

INSTRUCTIONS TO STUDENTS

1. Students are required to complete this test individually.
2. **Your submission should have only one file (.R) which includes the R code and your interpretation/analysis as comments**. Make sure that there is no error when the graders open and run your code file.
3. Be sure to lay out systematically the various parts and steps in your code file.
4. At the end of the test, please upload the answer file to CANVAS/DSA1101/Assignments/Midterm Submission.

1. (*50 points*) The data file `heart_disease_midterm.csv` is a random sample of 300 people who came to a hospital in the US for checking their heart. The sample is about the heart disease status (yes/no) and the variables that might help to predict the heart disease status.
The given file contains columns with their names listed below.

Column's Name	Description
age	age of the person
sex	1 = male; 0 = female
chest.pain	1 = the type of chest pain that is of high risk for health 0 = the type of chest pain that is of low risk for health
bp	resting blood pressure (mm Hg)
chol	serum cholesterol level (mg/dl)
disease	heart disease status (yes/no)

- please run function `setwd()` in a separate line (if you need it) when importing the data set into R.
- **the names given in bold below MUST be used in your R code**.
- please report numerical answers to at least three significant figures if it's absolute is smaller than one (e.g. 0.0123) and to three decimal places if it's absolute is larger than one (e.g. -2.345).

Part I: *Data Exploration (20 points)*

Load the file `heart_disease_midterm.csv` into R and name it as **data**.

1. (2 points) Write code to create a new column for **data**, named **cp**, which equals to "high risk" if the person has **chest.pain** of type 1, and equals to "low risk" if the person has **chest.pain** of type 0.
2. (2 points) Write code to change the labels for **sex** where 1 is replaced by "**male**" and 0 is replaced by "**female**".
3. (2 points) Write code to create a table of proportions for the response column, **disease**. Report the percentage of people that have heart disease.
4. (2 points) Write code to create a QQ plot for the sample of cholesterol level, **chol**. Give your comments.
5. (2 points) Write code to create box plots of people's age by groups of heart disease status. Give your comments.
6. (2 points) Create a contingency table (named **tab1**) for **cp** and **disease**. Report the number of people that have high risk type of chest pain and also have heart disease.

7. (4 points) Using **tab1**, write one command in R to find the two probabilities below.

   (i) the probability of having heart disease in the group of people having high risk type of chest pain.

   (ii) the probability of having heart disease in the group of people having low risk type of chest pain.

   Report the difference of the two probabilities above and interpret the meaning of that difference.

8. (4 points) Find the odds ratio for **tab1** and interpret it.

Part II: KNN (20 points)

For the questions in this Part II, we would want to form KNN classifiers using input features **age**, **chest.pain**, **bp** and **chol** which help to predict if a person has heart disease. We hence will consider **chest.pain** with 0, 1 as a numeric variable.

9. (2 points) Write code to create a new data frame, called **data.KNN** which has five columns: **age**, **chest.pain**, **bp** and **chol** after standardization for all observations; and the 5^{th} column is the column of the response, **disease**.

10. (4 points) Run the command **set.seed(703)**.

    Then, write code to randomly split **data.KNN** into two groups: one group has 240 rows (will be the train set), named as **train.set** and other group has the rest of rows (will be the test set), named as **test.set**.

11. (6 points) Use the train set with four standardized features to form the KNN classifiers where $\boldsymbol{k} = 3, 5, 7, 9, 11, 13, 15$, and TPR value for each classifier is kept in a vector named **tpr**.

12. (2 points) Using True Positive Rate (TPR) as the criterion, write code to find the best $\boldsymbol{k}$ found and report the value of TPR of the KNN classifier with that value of $\boldsymbol{k}$.

13. (6 points) Use the KNN with the best **k** found in Question 12 to predict the disease status of a female at age 60, having high risk type of chest pain, with information listed below.

    age = 60, **sex** = female, **chest.pain** = 1, **bp** = 160, **chol** = 200.

    Hint: you can use the same train set as in Question 11 above to form the classifier. You may standardize the values for the new person using the mean and the standard deviation of all the observations in the full data set.

Part III: DT (10 points)

For the questions in this Part III, we would want to form a decision tree which helps to predict heart disease status, using the full data set given and all 5 input features **age**, **sex**, **bp**, **chol**, and either **chest.pain** or **cp**.

14. (4 points) Write code to form a decision tree, named **DT**, to predict the heart disease status with `maxdepth = 3`, where variable selection and split points are based on Information Gain.

    How many input features are shown in the fitted tree? Among all the features shown in that tree, which one is the most important in predicting the disease status?

15. (2 points) Consider the person with information listed in Question 13. Using the tree **DT** formed above, report the predicted probability that the person will have heart disease.

16. (4 points) Use `predict()` function in R, write code to find the TPR of the tree **DT** for its prediction for all the points in **data**.

    Hint: template of function `predict()` is

    `predict(object, newdata, type = "prob")` where

    `object`: name of the model

    `newdata`: name of the data frame that contains the point(s) for prediction

    `type = "prob"` (default) to get the predicted probabilities; or `type = "class"` to get the predicted class label.

2. (*10 points*) Joshua Lee is working for a company where his monthly salary in 2025 is \$8,000 and the salary will be increased 5% from 01 January every year.
Every month, Joshua saves a fixed amount of 40% monthly salary.
So far, until end of March 2025, he has a saving amount of \$60,000.
He plans to form a start-up company which will requires an initial amount of \$150,000.

Define a function in R, named **F**, which helps to calculate the number of months (counting from 01 Apr 2025) that Joshua Lee could save up enough money to start the company.
Note: Function **F** must have at least one argument, named **salary**, to specify Joshua's monthly salary.

END OF QUESTIONS