Midterm 2324 S2

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 AY2023/2024)

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, label your answer file by your student number (such as A0123456B.R), and UPLOAD IT TO CANVAS/DSA1101/Assignments/Midterm Submission.

1. (*60 points*) The data file `abalone2.csv` is extracted from an original data set which includes information of 4177 abalones. The purpose of the study is to predict the age of the abalone. The given file contains some variables listed below.

Variable	Description
age	age of abalone (year)
sex	M = male, F = female, I = infant
length	the longest shell measurement (mm)
diameter	perpendicular to length (mm)
weight	whole abalone's weight (gram)

- use `set.seed(803)`.
- **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 smaller than one (e.g. 0.0123) and to three decimal places if it's larger than one (e.g. 2.345).

Part I: Data Preparation

1. (2 points) Load the file `abalone2.csv` into R and name it as **data**.
2. (2 points) Report the name of all the columns in **data**, in order as they appear in **data**.
3. (2 points) Write code to create a new categorical variable, **Year**, which equals to "young" if the age of the abalone is less than or equal to 10.5, and equals to "old" if the abalone's age is larger than 10.5.

Part II: Data Exploration

4. (2 points) Write code to create a frequency table for variable **Year** created above. Report the number of abalones that are in "old" group.
5. (4 points) Create a QQ plot for the sample of abalone's age. Give your comments.
6. (4 points) Create a box plot for the sample of abalone's age. Does it show any outliers? If yes, how many outliers are there?
7. (6 points) Find the mean weight of all the abalones that are the large outliers in the box plot of age above. Compared to the mean weight of all abalones in the data set, what's is the difference? Give your comments.
8. (2 points) Create a scatter plot of **age** against **weight**. Give your comments.

talk about variability.

Part III: Linear Model

9. (4 points) Fit a linear regression model for the response variable **age**, named as **M1**, using all input features, **sex**, **length**, **diameter** and **weight**. Report p-values of the regressors that are NOT significant in the model, at significance level 0.1.

10. (4 points) Give your comments on the suitability to fit a linear model for **age**.

11. (2 points) An abalone has information listed below. Write code to predict the age of this abalone. Report the prediction value.

    `sex` = M, `length` = 120 mm, `diameter` = 90 mm, and `weight` = 240 grams.

Part IV: KNN

Consider **Year** as the response variable. Training and testing data are the same, using all the observations in the data set. We would want to form KNN classifiers using numeric input features which helps to predict if an abalone belongs to "old" or "young" group.

12. (6 points) Use all the observations given with standardized numeric input features to form the KNN classifiers where $11 \leq \boldsymbol{k} \leq 50$, and Type 2 Error rate for each classifier is kept in a vector named **type.2**.

13. (4 points) Write code to produce a scatter plot of Type 2 Error, **type.2** against values of $\boldsymbol{k}$ in the question above.

14. (4 points) Using Type 2 Error rate as the criterion, report the best $\boldsymbol{k}$ found and the Type 2 Error rate of the KNN classifier with that $\boldsymbol{k}$.

    Note that, when checking the goodness of fit of the classifiers, "old" is treated as positive and "young" is treated as negative.

Part IV: Decision Trees

Consider **Year** as the response variable. Training and testing data are the same, using all the observations in the data set. We would want to form classifiers using Decision Tree, based on all input features which helps to predict if an abalone belongs to "old" or "young" group.

15. (4 points) Write code to form a decision tree, named as **fit**, with complex parameter $cp = 0.003$ and Information Gain for splitting the branches.

16. (4 points) Plot the tree **fit**. Based on this plot, predict if the abalone in Question 11 is "young" or "old".

17. (4 points) Find and report the accuracy of **fit**.

END OF QUESTIONS