NATIONAL UNIVERSITY OF SINGAPORE

DSA1101 INTRODUCTION TO DATA SCIENCE

(Semester 1 : AY 2023/2024)

Time: 2 Hours

PRACTICE PAPER
(no solution is provided)

The data file `data_finals.csv` is extracted from an original dataset which was obtained from the WHO and United Nations website. The given file contains information on 89 countries in the year 2014.
We consider the the following variables:

Variable	Description
Status	status of a country (Developed, Developing)
Life_expectancy	life expectancy (year)
Adult_mortality	adult mortality rate
infant_deaths	number of infant deaths for that country
Alcohol	alcohol consumption (per capita (15+) consumption, in liters of pure alcohol)

For the questions below,

- **Status** is considered as the response variable.
- all four features are used to form models/classifiers.
- use `set.seed(1101)`.
- please report numerical answers to three significant figures if it's smaller than one and to three decimal places if it's larger than one.

Data Preparation (10 points)

1. Load the dataset into R and name it as **data**. Write code to remove the first three columns of the dataset which we will not use for this question.
2. Report the number of developed countries in the data given.
3. Write code to transform the variable **Status** to numerical format where `Developing` = 0 and `Developed` = 1.

Part I: Linear Regression Model (10 points)

4. Write code to form a linear model (called M1) for **Status**. Report the R^2 of model M1.
5. How many fitted values of model M1 are less than 0?
6. What are the possible limitations when fitting a linear model to the response **Status**?
7. Write code to derive the AUC value of the ROC curve of model M1. Report the value.

Part II: Logistic Regression Model (35 points)

8. Write code to form a logistic regression model (called M2) for **Status**. Write down the fitted model and explain in detail any notations used.

9. Report the coefficient of **Alcohol** in model M2. Interpret it.

10. A country has information listed below. Write code to predict the probability that it is a developed country.

    `Life_expectancy` = 83, `Adult_mortality` = 57, `infant_deaths` = 2, and

    `Alcohol` = 3.

11. Write code to derive the AUC value of the ROC curve of model M2. Report the value.

12. Let δ denotes the threshold used for a classifier based on the probability derived from model M2. Write code to plot a figure to show how the TPR and the FPR of the classifier change when the threshold δ changes.

13. Is $\delta = 0.5$ a good choice? Explain.

14. Report the values of TPR and FPR when $\delta \in (0.19, 0.2)$. Give your comments.

Part III: Naive Bayes Classifier (15 points)

15. We now use the naive Bayes classifier for the dataset given. Write code to form the classifier, named as M3.

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

17. Using the classifier M3, predict the probability that the country with information listed in question 10 will be classified as ‘developed’. Report the probability.

18. Write code to derive the AUC value of the ROC curve of the classifier M3. Report the value.

Part IV: k-Nearest Neighbours (10 points)

19. Use all the observations and standardized features to form the best k-NN classifier where $2 \leq k \leq 10$, based on accuracy. Report the best k found and the accuracy of the k-NN classifier with that k (called M4).

20. Write code to derive the AUC value of the ROC curve of the classifier M4. Report the value.

Part V: Decision Trees (20 points)

21. Write code to form a decision tree to predict the status of a country with complexity parameter of 0.01, where variable selection and split points are based on Gini index.

22. Among the four features used to form the tree, which one(s) is/are more important in predicting the status of a country?

23. Write code to plot both ROC curves of the classifier M3 (naive Bayes, in red) and of the decision tree (in blue) in one figure. Add a legend for the figure which indicates the name of the classifier of each curve.

24. Among all the models and classifiers formed above, which one has the highest AUC value?

–END OF PAPER–