NATIONAL UNIVERSITY OF SINGAPORE

DSA1101 INTRODUCTION TO DATA SCIENCE

(Semester 1 : AY 2024/2025)

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 **EIGHT (8)** printed pages (including the cover page).
3. The data file used for this exam is given in the folder "For Finals" on Canvas/DSA1101/Files.
4. This is an OPEN BOOK BLOCK INTERNET 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(888)** 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,
   - Copy and paste your R code to the Examplify text box. Indentation and alignment may not be retained when pasting, but that is acceptable.
   - Save an exact copy of your R file on your laptop, and upload it to Canvas/DSA1101/Assignments/Final Exam Submission immediately after the exam.
10. Do not modify your code in your submission file. Any difference found (except for indentation or alignment) between Examplify and Canvas submissions will be penalized.

1. (*10 points*) No R code is required. Type your answers as comments in the R code file.

In 1990, 1200 randomly selected post-menopausal women were recruited for a study to investigate the effect of using Post-Menopausal Hormone (PMH) on the incidence of breast cancer. 200 of them were randomly chosen to be the users and the rest 1000 were the non-users.
The table below summarizes the results after 5 years.

	Breast Cancer Status	
PMH Use	Present	Absent
Yes	79	121
No	318	682

(a) Calculate and report the conditional probabilities that help to investigate if using PMH increases the chance of getting breast cancer.

(b) Calculate and report the odds ratio for the table above. Interpret it in the context of this study.

2. (*10 points*) **Multilple Choice Questions**.

A dataset collected contains information of patients and their cancer status. Let Column Y be the response variable (cancer status) with two categories, 0 and 1, where 1 = diseased and 0 = no-diseased. Let p denote $P(Y = \text{diseased})$.

(a) In order to get the prediction of p of the test points using a fitted 10-NN classifier, the R code is of the form

```
pred.prob = knn(train, test, cl, k = 10,___)
```

where in the blank space, we add

A. `type = "prob"`

B. `type = "class"`

C. `prob = TRUE`

D. `type = "response"`

E. `type = "raw"`

(b) Using a fitted decision tree, named `fit`, the prediction of p of the test points is obtained by

```
pred.prob = predict(fit, new.data = test, ___)
```

where in the blank space, we add

A. `type = "prob"`

B. `type = "class"`

C. `prob = TRUE`

D. `type = "response"`

E. `type = "raw"`

(c) Using a fitted Naive Bayes classifier, named `nb`, the prediction of p of the test points is obtained by

```
pred.prob = predict(nb, new.data = test, ___)
```

where in the blank space, we add

A. `type = "prob"`

B. `type = "class"`

C. `prob = TRUE`

D. `type = "response"`

E. `type = "raw"`

(d) Using a fitted logistic regression, named `lr`, the prediction of p of the test points is obtained by

```
pred.prob = predict(lr, new.data = test, ___)
```

where in the blank space, we add

A. `type = "prob"`

B. `type = "class"`

C. `prob = TRUE`

D. `type = "response"`

E. `type = "raw"`

3. (*10 points*) Consider a dataset about female horseshoe crabs during their mating season. Researchers want to investigate the male crabs that group around the female and potentially fertilize her eggs (called satellites). The table below gives the description for each variable in this study.
The first ten rows of the dataset is given in Figure 1.

Variable	Description
`satell`	1 = female crab has satellite(s), 0 = not having any satellites
`color`	female crab's color
`weight`	female crab weight (kg)

```
> head(data, 10)
   color satell weight
1  light      1   3.05
2   dark      0   1.55
3  light      1   2.30
4   dark      0   2.10
5   dark      1   2.60
6  light      0   2.10
7  light      0   2.35
8   dark      0   1.90
9  light      0   1.95
10  dark      0   2.15
```

Figure 1: The first 10 rows of the dataset

A model was fitted for the dataset given, called model M. The summary output for model M is given in Figure 2.

```
> M = glm(satell~ color +  weight, data = data, family = binomial)
> summary(M)

Call:
glm(formula = satell ~ color + weight, family = binomial, data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8107     0.8935  -4.265 2.00e-05 ***
colorlight    0.6772     0.3561   1.902   0.0572 .
weight        1.6950     0.3820   4.437 9.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 192.13  on 170  degrees of freedom
AIC: 198.13

Number of Fisher Scoring iterations: 4
```

Figure 2: Summary output of model M

Note: For the questions that follow, report numerical answers to three significant figures if they are smaller than one and to three decimal places if they are larger than one.

(a) Type the equation of model M into the R code file as comments. Explain in detail if any notation is used in the equation.

(b) Report the coefficient of variable `color` in model M and interpret it.

(c) Report the coefficient of variable `weight` in model M and interpret it. Is it significant in model M? Explain.

(d) Calculate and report the odds ratio of having satellites between two crabs below:

Crab A: weight = 3 kg, color = light;

Crab B: weight = 2 kg, color = dark.

4. (*70 points*) Consider a dataset about car evaluation where the quality of a car is evaluated based on some information from the record. Dataset is given in the file `car-eval-dsa1101.csv` The table below gives the description for each variable in this study.

Variable	Description
`buying`	buying price of the car: low, med = medium, high, and vhigh = very high
`maint`	price of the maintenance: 1= low, 2 = medium, 3 = high, 4 = very high
`doors`	number of doors of the car: 2, 3, 4, 5 = more than 4 doors
`persons`	car's capacity in terms of persons to carry: 2, 4, 6 = more than 4 persons
`lug.boot`	the size of luggage boot: 1 = small, 2 = medium, 3 = big
`safety`	estimated safety of the car: 1 = low, 2 = medium, 3 = high
`quality`	low, high

For the questions below,

- load the dataset into R and name it as **df**.
- report numerical answers to three significant figures if they are smaller than one and to three decimal places if they are larger than one.
- For all the models/classifiers in the following questions, there is no need to split the dataset given into train set and test set.
- For all the models/classifiers in the following questions, use the six columns `buying` to `safety` as the input features.

Part I: Decision Trees (20 points)

1. Write code to create a new column for `df` named `status`, where `status` equal to 1 if `quality` of car evaluation is high, and `status` equal to 0 otherwise.
2. Let `m` denote a vector of possible values for argument `minsplit` which is from 25 up to 50. Write code to fit a decision tree (using Information Gain) that helps to predict `status` of car evaluation using the dataset given for each value of `minsplit` in `m`.

   For each tree fitted, write code to obtain the values of TPR and FNR.
3. Which is the best value of `minsplit` such that the value of TPR of the tree built from that `minsplit` is at least 0.9. Write code to show how to determine the best value of `minsplit`.
4. With the value of `minsplit` chosen in Question 3, write code to fit a decision tree, to be named as `DT`. Report the name of the most important variable in that tree.
5. Write code to plot the ROC curve of the tree `DT`. Derive and report the value of AUC.

Part II: Naive Bayes Classifier (10 points)

6. We now use the naive Bayes classifier for the dataset given. Let `status` be the response variable. Write code to form the classifier, to be named as **NB**.

7. Write code to calculate the precision of `NB` for the given dataset using a threshold $\delta = 0.1$.

8. Write code to plot the ROC curve of the classifier `NB`. Derive and report the value of AUC.

Part III: Logistic Regression Model (20 points)

Since all the input features are ordered, we can fit a logistic regression for `status`, by considering each feature as a quantitative variable.

9. Write code to transform `buying` to numeric with values 1, 2, 3, and 4 for the categories `low, med, high`, and `vhigh`, respectively.

10. Using `status` as the response variable, write code to form a logistic regression model for it (called `LR`), using the six input features given.

11. Under model `LR`, compare the odds of having status being 1 (high quality) between a car that has big luggage boot and that of a car having small luggage boot, given that they both have the same information for other features.

12. Write code to calculate the precision of model `LR` for the given dataset using a threshold $\delta = 0.1$.

13. Write code to plot the ROC curve of model `LR`. Derive and report the value of AUC.

Part IV: KNN classifier (20 points)

Since all the input features are ordered, we can fit a KNN classifier for `status`, by considering each feature as a quantitative variable.

14. Write code to standardize the six input features, and store them in a data frame named `data.x`.

15. Consider KNN classifiers with k being the odd numbers from 3 up to 43. Write code to form those classifiers and store the values of TPR and FNR for each value of k.

16. Report the value of k that its TPR is at least 0.9 and its FNR is non-zero.

17. With the value of k determined in Question 16, write code to form a KNN classifier and find its precision using a threshold $\delta = 0.1$.

18. Write code to plot the ROC curve of the KNN classifier in Question 17. Derive and report the value of AUC.

19. Write code to create a plot that has all the four ROC curves created in Questions 5, 8, 13, 18 where each curve should have a different color and a legend box is included in the plot.

20. Among all the four ROC curves, which curve has highest AUC value?

–END OF PAPER–