NATIONAL UNIVERSITY OF SINGAPORE

# DSA1101 INTRODUCTION TO DATA SCIENCE

(Semester 2 : 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 **THREE (3)** questions 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. The use of offline Large Language Models (LLMs) is prohibited for this exam.
6. Both programmable calculators and non-programmable calculators are allowed.
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 into 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 the code in your submission file. Any difference found (except for indentation or alignment) between Examplify and Canvas submissions will be penalized.

**1.** (*20 points*)

**Part I: True/False Questions**. No explanation is required.

**(a)** If we use linear regression model with intercept to predict a response $y$ based on a regressor $x$ given 100 observations, the average of the residuals, $(1/100)\sum_{i=1}^{100} e_i$, will always be zero. True or False?

**(b)** If I am using House Price and Household Monthly Income (both in Singapore dollars) as two features for the K-means algorithm to divide Singapore households into groups, I do not need to standardize them since they have the same units. True or False?

**Part II: Multiple Choice Questions**. No explanation is required.

A data set contains information on 100 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 $P(Y = \text{diseased})$.

**(c)** Using a fitted decision tree, named `DT`, the prediction of $p$ for the test points is obtained by

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

where in the blank space, we add

A. `type = "raw"`

B. `prob = TRUE`

C. `type = "class"`

D. `type = "prob"`

E. `type = "response"`

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

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

where in the blank space, we add

A. `type = "raw"`

B. `prob = TRUE`

C. `type = "class"`

D. `type = "prob"`

E. `type = "response"`

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

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

where in the blank space, we add

A. `type = "raw"`

B. `prob = TRUE`

C. `type = "class"`

D. `type = "prob"`

E. `type = "response"`

**(f)** In order to obtain the prediction of $p$ for the test points using a fitted 9-NN classifier, the R code is of the form

```
pred.prob = knn(train, test, cl, k = 9,___)
```

where in the blank space, we add

A. `type = "raw"`

B. `prob = TRUE`

C. `type = "class"`

D. `type = "prob"`

E. `type = "response"`

**Part III: Open-Ended Questions**. Explanation is required.

**(g)** When constructing a classification model that helps to predict if a person has a disease, which is genetically transmitted and has costly treatment with significant side effects, which type of error rate (Type I or Type II) should be prioritized for minimization? Explain your answer.

**(h)** When constructing a classification model that helps to predict if a person has a highly contagious disease, such as Covid-19, which type of error rate (Type I or Type II) should be prioritized for minimization? Explain your answer.

**2.** (*10 points*) A data set on house selling price in the US was randomly collected. A model that helps to predict the selling price of a house using other variables is to be fitted. The description of variables in the data set is given below.

| Variable | Description |
|---|---|
| `price` | house's price in thousand US dollars |
| `size` | house's size in square feet (sqft) |
| `bedrooms` | number of bedrooms |
| `garage` | 1 = with garage; 0 = no garage |

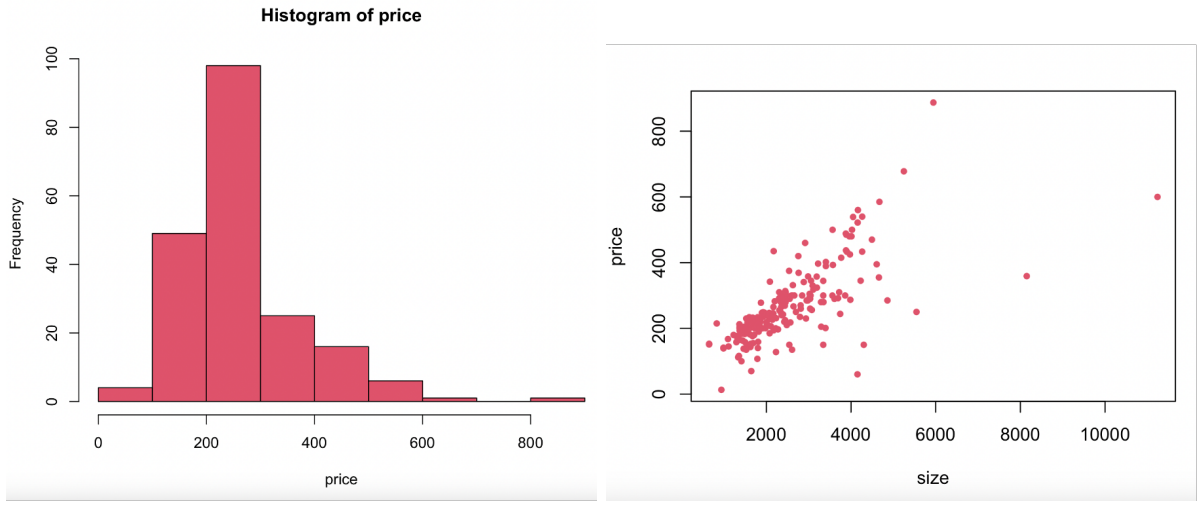


Figure 1: Histogram of price (left) and scatter plot of price against size (right)

**(a)** A histogram of the house's price and a scatter plot of the house's price against house's size are given in Figure 1. Comment if it is suitable to fit a linear regression model for `price`?

**(b)** A linear model, called M, is fitted. The summary output of model M is given in Figure 2 (next page).

Write down the fitted equation of model M.

Report the coefficient of the variable `garage` in model M and interpret it.

**(c)** Comment on the goodness of fit of model M.

**(d)** The information of a new house is given below. Using model M, calculate the predicted price of this house.

`size` = 1500 sqft, `bedrooms` = 3, `garage` = 1

```
> summary(M)

Call:
lm(formula = log(price) ~ size + bedrooms + garage, data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25845 -0.10709  0.05071  0.17037  0.92127

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.626e+00  8.444e-02  54.778  < 2e-16 ***
size        2.108e-04  2.021e-05  10.431  < 2e-16 ***
bedrooms    7.033e-02  2.344e-02   3.000  0.00305 **
garage1     1.561e-01  5.903e-02   2.644  0.00885 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3375 on 196 degrees of freedom
Multiple R-squared:  0.443, Adjusted R-squared:  0.4345
F-statistic: 51.97 on 3 and 196 DF,  p-value: < 2.2e-16
```

Figure 2: Summary output of model M

**3.** (*70 points*) Consider a random sample of 1,000 people. Their health information is recorded for a study about factors that may affect the diabetes status of people. The data set is given in the file `diabetes-dataset-1k.csv`. The table below gives the description for some variables in this study.

| Variable | Description |
|---|---|
| `age` | person's age (in years) |
| `hypertension` | 0 = No, 1 = Yes |
| `heart_disease` | 0 = No, 1 = Yes |
| `bmi` | the Body Mass Index |
| `HbA1c_level` | Hemoglobin A1c level (%) |
| `blood_glucose_level` | blood glucose level (mg/dL) |
| `diabetes` | 0 = No, 1 = Yes |

For the questions below,

- Load the data set into R and name it as **data**.
- Report the final numerical answer to three significant figures if its absolute value is smaller than one (e.g. 0.0123) and to three decimal places if its absolute value is larger than one (e.g. -2.345).

**Part I: Linear Model** (8 points)

For this part, there is no need to split the data set given into a train set and a test set.

1. Write code to fit a linear regression model using all the input features, to be named as **LM**, for predicting the probability of having diabetes. Is there any insignificant variable in model **LM**?
2. Comment on the goodness of fit of model **LM**.
3. State at least two disadvantages of fitting a linear model for a binary response.

**Part II: Decision Trees** (16 points)

All the decision tree models below are using Information Gain for branch split.

4. Run the command **set.seed(2904)**.

   Then, write code to randomly split **data** into two subsets, one set with 800 rows, to be named as **train.set** and other set with the rest of rows, to be named as **test.set**.

5. Let **m** denote a vector of possible values for the argument `minsplit` which ranges from 40 up to 50 inclusively.

   For each value in **m**, write code to

   (i) fit a decision tree using **train.set**;

   (ii) use the fitted tree to predict the diabetes status of people in **test.set**;

   (iii) derive the precision and the Type I error rate for the prediction in (ii).

6. Consider `minsplit = 50`, write code to fit a decision tree, to be named as **DT**. Report the name of the most important variable in that tree.

7. Write code to plot the ROC curve of the tree **DT** in black color based on its prediction for **test.set**. Derive and report the value of AUC.

**Part III: Naive Bayes Classifier** (18 points)

8. We now fit a naive Bayes classifier using **train.set**. Write code to fit the classifier, to be named as **NB**.

9. Write code to plot the ROC curve of the classifier **NB** in blue color based on its prediction for **test.set**. Derive and report the value of AUC.

10. Let $\delta$ denote the threshold used for the ROC curve in Question 9. Write code to plot a figure that shows how the TPR and the FPR change when the threshold $\delta$ changes.

11. Propose a value of the threshold $\delta$ such that the classifier **NB** can attain a TPR of at least 0.9 while the FPR is as low as possible.

12. With the proposed value of $\delta$ in Question 11, write code to calculate the accuracy of the classifier **NB** when it is used to predict the diabetes status for **test.set**.

**Part III: Logistic Regression Model** (16 points)

13. Write code to fit a logistic regression model using **train.set** by including all the given input features. What is the most insignificant feature in that fitted model?

14. Write code to fit a logistic regression model using **train.set**, to be named as **LR**, by including all the given input features except the most insignificant feature found in Question 13.

15. Using the model **LR**, compute and report the odds ratio of having diabetes between a person that has hypertension and a person without hypertension, given that they both have the same values for the other features.

16. With the proposed value of $\delta$ in Question 11, write code to calculate the accuracy of model **LR** when it is used to predict the diabetes status for **test.set**.

17. Write code to plot the ROC curve of the model **LR** in red color based on its prediction for **test.set**. Derive and report the value of AUC.

**Part IV: KNN classifier** (12 points)

18. Standardize all the quantitative input features in **train.set** and **test.set**.

19. Write code to fit a KNN classifier with $k = 5$ using **train.set** to predict the winning probabilities for **test.set**.

20. Write code to find the predicted probability of having diabetes for each person in **test.set**.

21. With the proposed value of $\delta$ in Question 11, write code to calculate the accuracy of the 5-NN classifier above when it is used to predict the diabetes status for **test.set**. Report the accuracy.

–END OF PAPER–