KOLEJ UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

ACADEMIC YEAR 2020/2021

APRIL/MAY FINAL ONLINE ASSESSMENT

AACS2383 INTRODUCTION TO DATA MINING

FRIDAY, 7 MAY 2021 TIME: 9.00 AM – 12.00 NOON (3 HOURS)

DIPLOMA IN COMPUTER SCIENCE (DATA SCIENCE)

Instructions to Candidates:

Answer **ALL** questions in the requested format or template provided.

- This is an open book final online assessment. You MUST answer the assessment questions on your own without any assistance from other persons.
- You must submit your answers within the following time frame allowed for this online assessment:
  - The deadline for the submission of your answers is **half an hour** from the end time of this online assessment.
- Penalty as below **WILL BE IMPOSED** on students who submit their answers late as follows:
  - The final marks of this online assessment will be reduced by 10 marks for answer scripts that are submitted within 30 minutes after the deadline for the submission of answers for this online assessment.
  - The final marks of this online assessment will be downgraded to zero (0) mark for any answer scripts that are submitted after one hour from the end time of this online assessment.
- Extenuation Mitigating Circumstance (EMC) encountered, if any, must be submitted to the Faculty/Branch/Centre within 48 hours after the date of this online assessment. All EMC applications must be supported with valid reasons and evidence. The UC EMC Guidelines apply.

FOCS Additional Instructions to Candidates:

- Include your **FULL NAME**, **STUDENT ID** and **PROGRAMME OF STUDY** in your submission of answer.
- Read all the questions carefully and understand what you are being asked to answer.
- Marks are awarded for your own (original) analysis. Therefore, use the time and information to build well-constructed answers.

AACS2383 INTRODUCTION TO DATA MINING

Declaration by candidates:

By submitting this e-assessment, I declare that this submitted work is free from all forms of plagiarism and for all intents and purposes is my own properly derived work. I understand that I have to bear the consequences if I fail to do so.

Final Online Assessment Submission
Course Code:
Course Title:
Signature:
Name of Student:
Student ID:
Date:

AACS2383 INTRODUCTION TO DATA MINING

Question 1

a) There are few issues involving data quality assessment such as accuracy, completeness, and consistency.
For each of the above-mentioned **THREE (3)** issues, discuss:

(i) how data quality assessment can depend on the intended use of the data. Support your answers with relevant examples.

(6 marks)

(ii) illustrate **TWO (2)** other dimensions of data quality.

(4 marks)

b) Missing values for some attributes are a regular issue that arises in real-world datasets. As a data analyst, describe **THREE (3)** methods on how you are handling this issue to proceed with the data for further analysis.

(9 marks)

c) Assume the following data is given: {22, 12, 61, 57, 30, 1, 32, 37, 37, 68, 42, 11, 25, 7, 8, 16}.

(i) Apply data discretization by binning the data into 4 bins using equal-depth and equi-width binning, respectively.

(5 marks)

(ii) Distinguish the two binning methods. Give an appropriate example application for each of the binning methods.

(5 marks)

(iii) If you know that the data represent ages of persons, what kind of binning method would you then use? (You may propose a third binning method.)

(3 marks)

d) Assume that a data warehouse with four dimensions, date, viewer, location, and game, and the two measures, count, and charge, where the charge is the cost that a viewer pays when watching a game on a given date. Viewers may be students, adults, or seniors, with each category having its charge rate.
Draw a star schema diagram for the data warehouse.

(10 marks)

Question 1 (continued)

e) In general, a star schema or a snowflake schema can be used to create a data warehouse. Briefly examine the similarities and the differences between the two models, and then analyse their benefits and drawbacks regarding one another. Give your opinion of which might be more empirically beneficial and give the reasons behind your answers.

(8 marks)

[Total: 50 marks]

Question 2

a) It is given a training set of positive (+) and negative (-) instances. The instances represented using two real-valued features (X and Y) (Figure 2 (a) and Figure 2 (b)). Suppose we aim to randomly divide this data into a training set (90%) and, a test set (10%) and to train and assess a model.

(i) Among the classifier, KNN (with K=1) or Naïve Bayes which one do you think would have a higher chance of doing well in terms of accuracy for Figure 2(a)? Explain your answers.

(7 marks)

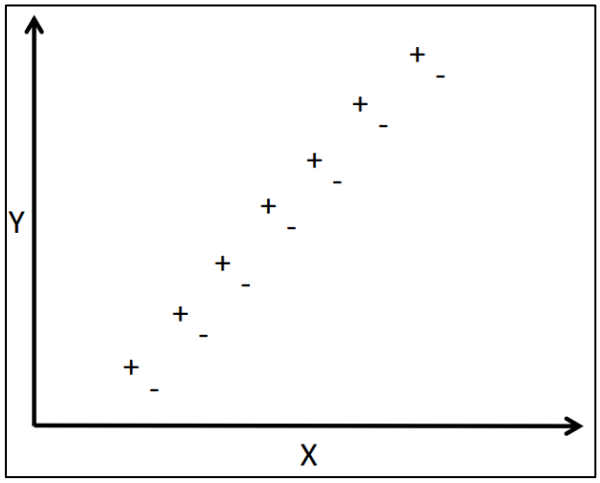

Figure 2 (a)

(ii) Among the classifier, KNN (with K=1) or Naïve Bayes which one do you think would have a higher chance of doing well in terms of accuracy for Figure 2(b)? Give reasons to support your answer.

(8 marks)

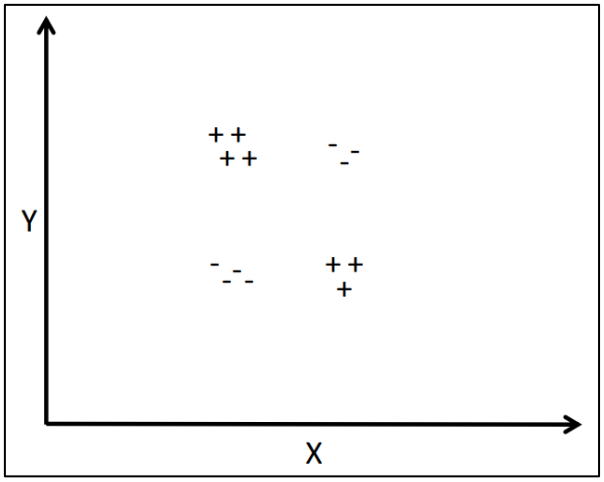

Figure 2 (b)

b) Assume that you are a machine in a wood company and should be learned to distinguish the wood types based on their characteristics. There are mainly two types of woods are available from two types of tree called Oaktree and Pine tree. As a machine you are required to apply a Decision Tree classifier for the following records in Table 1:

Table 1

Sample	Hardness	Density	WoodGrain	Class
1	Strong	High	Tiny	Oak
2	Strong	High	Big	Oak
3	Strong	High	Tiny	Oak
4	Weak	Low	Big	Oak
5	Strong	Low	Big	Pine
6	Weak	High	Tiny	Pine
7	Weak	High	Big	Pine
8	Weak	High	Tiny	Pine

(i) Identify the attribute that will be selected as the root of the tree. Give a reason to support your selection.

(4 marks)

(ii) Construct the decision tree whereby information gain is applied recursively to select roots of sub-trees as presented in Table 1.

(7 marks)

Question 2 b) (Continued)

(iii) Classify these new samples by using the constructed decision tree in 2 b) (ii).

1. What is the class for [Density=Low, WoodGrain=Tiny, Hardness=Strong]?
2. What is the class for [Density=Low, WoodGrain=Tiny, Hardness=Weak]?

(4 marks)

c) Imagine that you run Density-based spatial clustering of applications with noise (DBSCAN) with *MinPoints* = 6 and *epsilon* = 0.1 for a dataset. Thus, you obtained four clusters and 5% of the objects in the dataset are classified as outliers.
Assume you run DBSCAN with *MinPoints* = 8 and *epsilon* = 0.1. Analyse how do you expect the clustering results to change?

(6 marks)

d) In general, clustering evaluation assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method. Compare the two key cluster quality evaluation methods: extrinsic and intrinsic with suitable examples.

(6 marks)

e) Assume that you are using a standard classifier on the training set that contains 10 points. However, each example has one real-valued feature, *x*, and a binary class label, *y*, with value 0 or 1 (Figure 2 (c)). This standard classifier is defined to predict the class label that is in the training dataset, regardless of the input value. Assume in the case of ties, predict class 1.

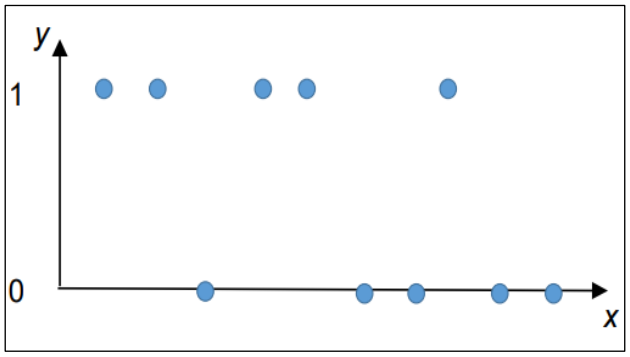

Figure 2 (c)

(i) Calculate the training dataset accuracy.

(2 marks)

(ii) Calculate the Leave-1-Out Cross-Validation accuracy.

(3 marks)

Question 2 e) (Continued)

(iii) What is the 2-fold Cross-Validation accuracy? Assume that the leftmost 5 points (i.e., the 5 points with smallest x values) are in one-fold and the rightmost 5 points are in the second fold.

(3 marks)

[Total: 50 marks]