UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2018/2019 : SEMESTER I

WQD 7003 : Data Analytics

Jan 2019 Time: 2 hours

INSTRUCTIONS TO CANDIDATES:

Answer **ALL** questions (40 marks).

(This question paper consists of 5 questions on 3 printed pages)

1. What is the difference between agglomerative and divisive hierarchical clustering?

(4 marks)

2. Describe THREE (3) types of data smoothing techniques to handle Noisy (Outlier) data.

(9 marks)

3. What is the use of regression in data analytics? What are TWO (2) types of regression and what is the difference between them?

(7 marks)

4. In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Based on your understanding of Bayes, answer the following question.

   a) Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual test positive, what is the probability he or she is a user?

   (6 marks)

   b) What are the advantages of using a Naive Bayes for classification?

   (4 marks)

5. The Weather dataset (Table 1) is a small dataset and is entirely fictitious. It supposedly concerns the conditions that is suitable for playing some unspecified game. The condition attributes of the dataset are in {Outlook, Temperature, Humidity, Wind} and the decision attribute is PlayTennis to denote whether to play tennis. In its simplest form, all four attributes have categorical values. Values of the attributes Outlooks, Temperature, Humidity, and Wind are in {sunny, overcast, rainy}, {hot, mild, cool}, {high, normal}, and {weak, strong} respectively.

Table 1: Weather Datasets

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

a) Using ID3, calculate the gain in the Gini index when splitting on attributes, Outlook Temperature, Humidity, and Wind, respectively. Show your calculation details. According to the gain, which one will you choose as the first attribute to split in the decision tree induction?

(6 marks)

b) Build a decision tree from the given Weather dataset. You should build a tree to predict PlayTennis, based on the other attributes.

(4 marks)

END