UTHM
Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
(ONLINE)
SEMESTER I
SESSION 2020/2021

COURSE NAME	:	TECHNIQUES IN DATA MINING
COURSE CODE	:	BWB 44603
PROGRAMME CODE	:	BWQ
EXAMINATION DATE	:	JANUARY / FEBRUARY 2021
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER ALL QUESTIONS. **OPEN BOOK EXAMINATION**

THIS QUESTION PAPER CONSISTS OF **SIX (6)** PAGES

Q1 (a) One method in data mining is to select the similarity measures for the data analysis. However, results can vary depending on the similarity measures used. Consider the data in **Table Q1(a)**, answer the following questions.

Table Q1(a)

	A1	A2
x1	1.5	1.7
x2	2	1.9
x3	1.6	1.8
x4	1.2	1.5
x5	1.5	1

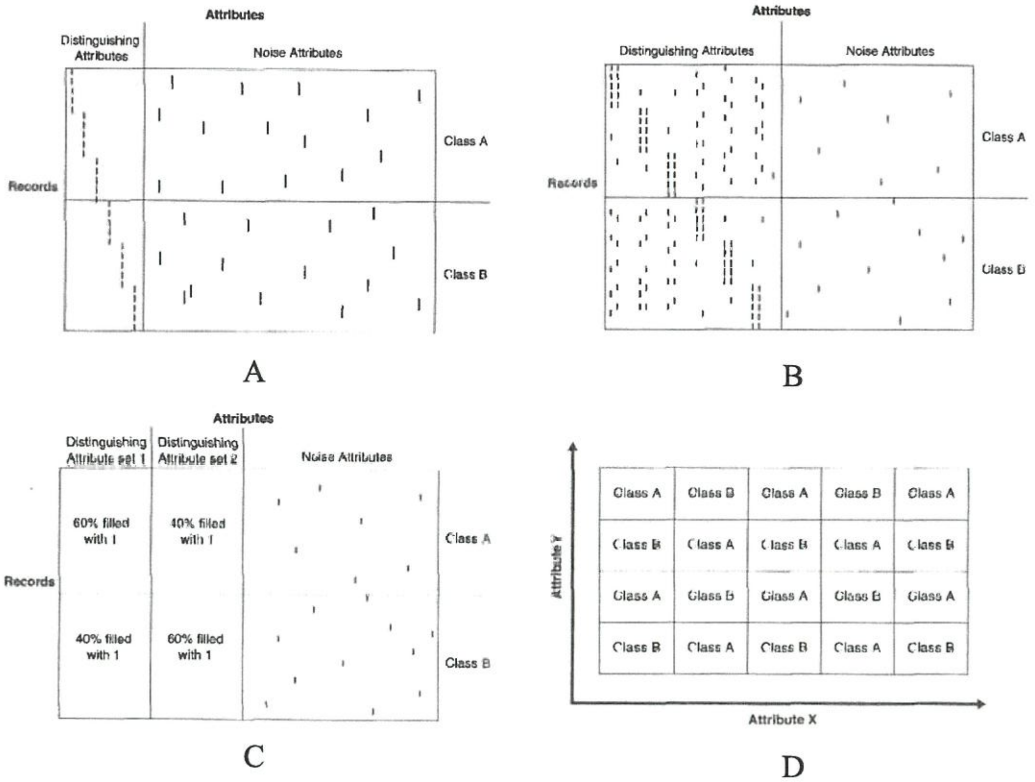
(i) Given a new data point, x = (1.4, 1.6), rank the database points based on the similarity by using Euclidean distance, Manhattan distance and cosine distance.

(6 marks)

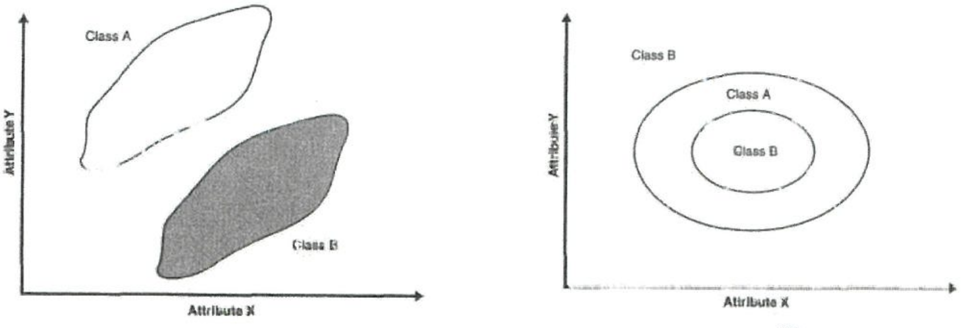
(ii) Normalize the data set so the norm of each data point is equal to 1 by using Euclidean distance.

(6 marks)

(b) Comments on how the **decision tree, naive Bayes** and **k-nearest neighbor classifier** will be performed on the data set shows in **Figure Q1(b)**.



Class A	Class B	Class A	Class B	Class A
Class B	Class A	Class B	Class A	Class B
Class A	Class B	Class A	Class B	Class A
Class B	Class A	Class B	Class A	Class B



E F

Figure Q1(b)

(18 marks)

Q2 For a data mining assignment, Syuhada has collected a traffic accidents data that happen in Panchor, Muar as in **Table Q2.**

Table Q2

Weather condition	**Driver's Condition**	**Traffic Violation**	**Seat Belt**	**Crash Severity**
Sunny	Alcohol-impaired	Exceed speed limit	No	Major
Raining	Healthy	None	Yes	Minor
Sunny	Healthy	Disobey stop sign	Yes	Minor
Sunny	Healthy	Exceed speed limit	Yes	Major
Raining	Healthy	Disobey red traffic	No	Major
Sunny	Alcohol-impaired	Disobey stop sign	Yes	Minor
Raining	Alcohol-impaired	None	Yes	Minor
Sunny	Healthy	Disobey red traffic	Yes	Major
Sunny	Alcohol-impaired	None	No	Major
Raining	Healthy	Disobey red traffic	No	Major
Sunny	Alcohol-impaired	Exceed speed limit	Yes	Major
Raining	Healthy	Disobey stop sign	Yes	Minor

(a) Construct a binarized version of the data set.

(12 marks)

(b) Identify the maximum width of each transaction in the binarized data.

(1 mark)

(c) Assume that the support threshold is 30%, determine the number of candidate and the frequency of the itemset that will be generated.

(2 marks)

(d) From **Table Q2**, recreate a data set that contains only the following asymmetric binary attributes:

- *Weather: Raining*
- *Driver's condition: Alcohol-impaired*
- *Traffic violation: Yes*
  Seat Belt: No
- *Crash severity: Major*

For Traffic violation, only None has a value of 0. The rest of the attribute values are assigned to 1.

Assume that support threshold is 30%, determine the number of candidate and the frequent itemset will be generated.

(7 marks)

(e) Compare your answer in **Q2(c)** and **Q2(d)**.

(1 mark)

Q3 **Table Q3** is a similarity matrix for a set of data.

Table Q3

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

(a) Construct single and complete link hierarchical clustering (manually).

(14 marks)

(b) Display a dendrogram and clearly show the order in which the points are merged.

(4 marks)

Q4 Given the Bayesian network shown in **Figure Q4**.

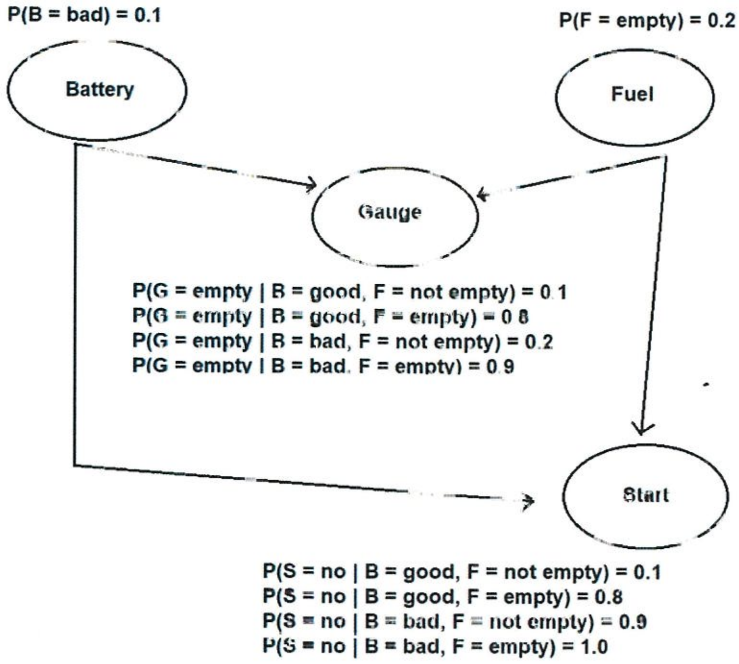

P(B = bad) = 0.1
P(F = empty) = 0.2
Battery
Fuel
Gauge
P(G = empty | B = good, F = not empty) = 0.1
P(G = empty | B = good, F = empty) = 0.8
P(G = empty | B = bad, F = not empty) = 0.2
P(G = empty | B = bad, F = empty) = 0.9
Start
P(S = no | B = good, F = not empty) = 0.1
P(S = no | B = good, F = empty) = 0.8
P(S = no | B = bad, F = not empty) = 0.9
P(S = no | B = bad, F = empty) = 1.0

Figure Q4

Compute the following probabilities.

(a) $P(B = \text{good}, F = \text{empty}, G = \text{empty}, S = \text{yes})$

(3 marks)

(b) $P(B = \text{bad}, F = \text{empty}, G = \text{not empty}, S = \text{no})$

(3 marks)

(c) Given that the battery is bad, compute the probability that the car will start.

(3 marks)

Q5 **Table Q5** is a data set that contains two attributes, Raining and Sunny, and two class labels, Yes and No. Each attribute can take three different values: 0, 1, or 2. The concept is following the rules:
Yes class is Sunny =1
No class is Raining $= 0 \vee$ Raining $= 2$.

Table Q5

Raining	Sunny	Number of Instances	
		Yes	No
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

(a) Construct a decision tree on the data set by using classification error rate index. (15 marks)

(b) Identify which attributes will be chosen as the first split. (1 mark)

(c) Build a confusion matrix on the training data and then evaluate accuracy, precision and recall. (4 marks)

– END OF QUESTIONS –