UTHM
Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2016/2017

COURSE NAME	:	INTRODUCTION TO DATA MINING
COURSE CODE	:	BWB 43303
PROGRAMME CODE	:	BWQ
EXAMINATION DATE	:	JUNE 2017
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **FOUR (4)** PAGES

Q1 (a) Classify the following attributes as **binary**, **discrete** or **continuous**. Also classify them as **qualitative (nominal or ordinal)** or **quantitative (interval or ratio)**.

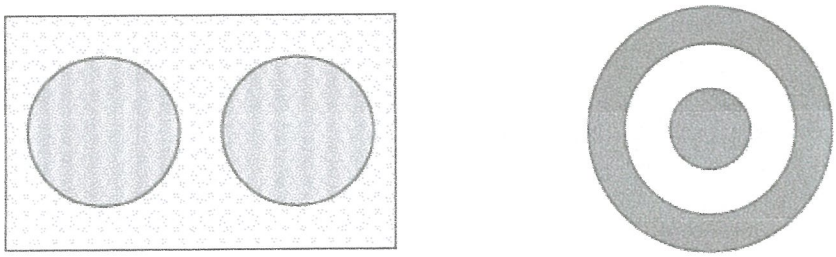
(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
(ii) Military rank.
(iii) Density of a substance in grams per cubic centimeter.
(iv) Angles as measured in degrees between 0° and 360°.
(v) Brightness as measured by people's judgements.

(10 marks)

(b) Name and describe clearly the main features of Genetic Algorithms (GA).

(10 marks)

Q2 (a) Identify the clusters for case (a) and (b) in **Figure Q2(a)** below using the **center-**, **contiguity-**, and **density** based definitions. Also indicate the number of cluster for each case (a) and (b) and give an explanation of your reasons.



(a) (b)

Figure Q2(a): Cases of clusters

(12 marks)

(b) Use the similarity matrix in **Table Q2(b)** below to perform single and complete link hierarchical clustering. Show your results by constructing a clear dendrograms.

Table Q2(b): Similarity matrix

	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

(10 marks)

Q3 (a) Construct an association rule from the market basket items that satisfy the following conditions. Describe clearly whether the rules are subjectively interesting or uninteresting.

Market Basket Items: Milk, Bread, Tuna, Cooking Oil, Laundry detergent, Soft-drink, Caviar.

Conditions:

(i) A rule that has high support and high confidence.
(ii) A rule that has reasonably high support but low confidence.
(iii) A rule that has low support and low confidence.
(iv) A rule that has low support and high confidence.

(8 marks)

(b) Perform a contingency table for each of the following rules using the transactions shown in **Table Q3(b)** below:

Table Q3(b): Market basket transactions

Transaction ID	Items Bought
1	$\{a,b,d,e\}$
2	$\{b,c,d\}$
3	$\{a,b,d,e\}$
4	$\{a,c,d,e\}$
5	$\{b,c,d,e\}$
6	$\{b,d,e\}$
7	$\{c,d\}$
8	$\{a,b,c\}$
9	$\{a,d,e\}$
10	$\{b,d\}$

Rules:

(i) $\{b\} \rightarrow \{c\}$
(ii) $\{a\} \rightarrow \{d\}$
(iii) $\{b\} \rightarrow \{d\}$
(iv) $\{e\} \rightarrow \{c\}$
(v) $\{c\} \rightarrow \{a\}$

(5 marks)

(c) Compute and rank the rules in decreasing order based on your answers in **Q2(b)**, according to:
(i) support.
(ii) confidence.

(20 marks)

Q4 **Table Q4** summarizes a data set with three attributes X, Y, Z and two class labels C_1, C_2.

Table Q4: Summarization of three attributes with two class labels

Instances	X	Y	Z	Target Class
1	T	T	1.0	C_1
2	T	T	6.0	C_1
3	T	F	5.0	C_2
4	F	F	4.0	C_1
5	F	T	7.0	C_2
6	F	T	3.0	C_2
7	F	F	3.0	C_2
8	T	F	7.0	C_1
9	F	T	5.0	C_2

(a) Compute the entropy in **Table Q4** with respect to the C_1 class.

(3 marks)

(b) Calculate the information gains of X and Y relatively to these training examples.

(8 marks)

(c) Obtain the information gain for every possible split in attribute Z.

(12 marks)

(d) Based on the classification error, identify the best split between X and Y.

(2 marks)

– **END OF QUESTIONS** –