UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2014/2015

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 33603
PROGRAMME	:	3 BIT
EXAMINATION DATE	:	DECEMBER 2014/JANUARY 2015
DURATION	:	2 HOURS 30 MINUTES
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **FIVE (5)** PAGES

Q1 Explain each issues in data quality as follows:

(a) Noise/outliers

(2 marks)

(b) Missing values

(2 marks)

(c) Duplicate data

(2 marks)

Q2 Differentiate:

(a) Labelled and unlabelled data

(6 marks)

(b) Training and testing set

(6 marks)

(c) Hierarchical clustering and partitional clustering

(6 marks)

Q3 Calculate the impurity of each node in Figure **Q3** using:-

(a) Gini index

(6 marks)

(b) Entropy

(6 marks)

Node N1	Count
Class = 0	0
Class = 1	6

Node N2	Count
Class = 0	1
Class = 1	5

Node N3	Count
Class = 0	3
Class = 1	3

FIGURE Q3

Q4 Using the training set in **Table 1** and the Euclidean distance measure:

(a) Calculate the distance of each instance in Training set from the Unseen set, which consists of `Attribute 1: 9.1`, and `Attribute 2: 11.0`, respectively.

(10 marks)

(b) Examine and identify the 5-nearest neighbours of the Unseen set.

(5 marks)

Table 1: Training set

Attribute 1	Attribute 2	Class
0.8	6.3	-
1.4	8.1	-
2.1	7.4	-
2.6	14.3	+
6.8	12.6	-
8.8	9.8	+
9.2	11.6	-
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	-
14.0	19.9	-
14.2	18.5	-
15.6	17.4	-
15.8	12.2	-

Q5 Four classifiers are generated for the same training set, which has 100 instances. They have confusion matrices shown in **Table 2**.

(a) Calculate the values of True Positive rate and False Positive rate for each classifier

(8 marks)

(b) Calculate the value for Euclidean distance measure for each one.

(8 marks)

(c) Based on the answer in **Q5 (b)**, identify which classifier would be the best?

(2 marks)

Table 2: Confusion Matrices

		Predicted class	
		+	-
Actual class	+	50	10
	-	10	30

		Predicted class	
		+	-
Actual class	+	55	5
	-	5	35

		Predicted class	
		+	-
Actual class	+	40	20
	-	1	39

		Predicted class	
		+	-
Actual class	+	60	0
	-	20	20

Q6 Figure **Q6** shows the attributes of Abalone data taken from UCI Machine Learning Repository. Graphically illustrate and label the input/output mapping for the prediction of Abalone age using a Multilayer Perceptron.

```
Name / Data Type / Measurement Unit / Description
----------------------------
1-Sex / nominal / -- / M, F, and I (infant)
2-Length / continuous / mm / Longest shell measurement
3-Diameter / continuous / mm / perpendicular to length
4-Height / continuous / mm / with meat in shell
5-Whole weight / continuous / grams / whole abalone
6-Shucked weight / continuous / grams / weight of meat
7-Viscera weight / continuous / grams / gut weight (after bleeding)
8-Shell weight / continuous / grams / after being dried
9-Rings / integer / -- / +1.5 gives the age in years
```

FIGURE Q6

(8 marks)

Q7 Consider the market basket transactions shown in **Table 3**.

Table 3: Customer dataset

Transaction ID	Buying Item
1	Milk, Tea, Chocolate
2	Bread, Butter, Milk
3	Milk, Chocolate, Cookies
4	Bread, Butter, Cookies
5	Tea, Cookies, Chocolate
6	Milk, Chocolate, Bread, Butter
7	Bread, Butter, Chocolate
8	Tea, Chocolate
9	Milk, Chocolate, Bread, Butter
10	Cookies

(a) Calculate the number of possible candidate itemsets for **Table 3**.

(3 marks)

(b) Using Apriori algorithm, break down the step by step procedure to mine frequent itemset by reducing the number of candidates, given the Minimum Support = 3.

(20 marks)

- END OF QUESTION -