UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2012/ 2013

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 3363 / BIT 33603
PROGRAMME	:	3 BIT
EXAMINATION DATE	:	DECEMBER 2012 / JANUARY 2013
DURATION	:	2 HOURS 30 MINUTES
INSTRUCTION	:	ANSWER **ALL** QUESTIONS.

THIS QUESTION PAPER CONSISTS OF FOUR (4) PAGES

Instructions: Answer **ALL** questions.

Q1 Cross Industry Standard Process (CRISP-DM) is the industry standard methodology for data mining and predictive analytics. This methodology defined and validated a data mining process that is applicable in diverse industry sectors, and makes large data mining projects faster, cheaper, more reliable and more manageable.

(a) Explain **SIX (6)** steps of CRISP-DM.

(12 marks)

(b) Discuss **THREE (3)** methods for data transformation.

(8 marks)

Q2 (a) Define the term k-means clustering.

(4 marks)

(b) Describe the algorithm to develop these clusters.

(6 marks)

(c) Suppose that we have 5 objects and each object has location in x and y-axis as shown in **Table 1**. Group these objects into K=2 clusters based on their location.

Table 1: Location of the objects

	x-axis	y-axis
Object 1	1	1
Object 2	2	1
Object 3	4	3
Object 4	5	4
Object 5	3	6

(10 marks)

Q3 In data mining application, the use of neural network in classification task is common and gives lower classification error rate compare to other learning community. Data analysis methods vary on the way how they detect patterns.

(a) Define the term classification.

(2 marks)

(b) Outline **TWO (2)** steps using neural network classification.

(4 marks)

(c) Given 3-input neural network with input patterns are $x_0 = 1$, $x_1 = 1$, $x_2 = -1$ has the set of weights as $w_0 = 0.3$, $w_1 = -2.0$, $w_2 = 1.5$, and suppose that the desired output is 1.

(i) what is the actual output ?

(ii) what is the value of error δ?

(iii) Assuming that the weights are updated after each pattern and the value of η is **0.33**, what are the new values for the weights?

(d) Using these new values of weights, what would the output be for the same input pattern?

(14 marks)

Q4 (a) Describe the term Association Rule.

(8 marks)

(b) In the following, **Table 2** has eight transactions on items {A,B,C,D,E}, use the Apriori Algorithm to compare the frequent item sets with their minimum support as 3.

Table 2: Transaction on items {A,B,C,D,E}

Tid	Items
1	{A,B}
2	{A,B,C}
3	{B,C,D}
4	{B,C}
5	{A,B,C,D}
6	{B,D}
7	{B,E}
8	{B,D,E}

Clearly indicate the steps of the algorithm to give all generators of closed frequent item sets and their closure.

(12 marks)

Q5 (a) Define decision tree structure.

(4 marks)

(b) Define the term entropy.

(6 marks)

(c) Based on the following scenario:

A study on disease diagnosis is conducted and a dataset is given in **Table 3**. The study shows whether the patient will cause different disease. The decision variable is identified in the last column.

Table 3: Hypothetical Training Data for Disease Diagnosis

Patient ID	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	no	no	no	yes	yes	allergy
2	no	no	no	yes	no	allergy
3	yes	no	no	yes	yes	allergy
4	no	yes	no	yes	no	cold
5	no	yes	no	yes	yes	cold
6	yes	yes	no	yes	no	cold
7	yes	yes	no	yes	yes	cold
8	no	no	yes	no	no	Strep throat
9	yes	yes	yes	yes	yes	Strep throat
10	yes	no	yes	no	no	Strep throat

Calculate the entropy of each attribute and generate decision tree.

(10 marks)

- END OF QUESTION -