Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2015/2016

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 33603
PROGRAMME	:	3 BIT
EXAMINATION DATE	:	DECEMBER 2015 / JANUARY 2016
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **FOUR (4)** QUESTIONS ONLY.

THIS QUESTION PAPER CONSISTS OF **EIGHT (8)** PAGES

Q1 (a) Answer **Q1(a)(i)** and **Q1(a)(ii)** based on Figure **Q1(a)**.

```
Data are raw facts, number or text that can be processed by a
computer. Data can exist in any kind or format. Due to the
ease of data collection, organizations are accumulating vast
and growing amounts of data in different formats and in
different data repositories. One of the key issues in data
mining is the data quality since 80% of mining efforts often
spend their time on data quality improvement. That is why pre-
processing data is very critical and needed since this process
will increase the accuracy of data mining techniques.
```

FIGURE Q1(a)

(i) Explain **TWO (2)** kinds of data quality problems.

(4 marks)

(ii) Propose **ONE (1)** solution for each data quality problems based on Figure **Q1(a)**.

(2 marks)

(b) A set of data is given as {3, 4, 7, 2, 14, 0, 21, 9, 1, 5}. Pre-process the data into a new data set with [0, 0.9] range by considering data smoothing and data normalization formula (Equation 1).

$$D'(i) = \frac{D(i) - \min(D)}{\max(D) - \min(D)} * (upper - lower) + lower \qquad \text{(Equation 1)}$$

(8 marks)

(c) Differentiate between:

(i) Noise and outliers.

(2 marks)

(ii) Data integration and data reduction.

(2 marks)

(iii) Training set and testing set.

(2 marks)

Q2 (a) Explain the function of each operator in *RapidMiner* as follow:

(i) Set Role

(2 marks)

(ii) Performance (Classification)

(2 marks)

(iii) Split Validation

(2 marks)

(iv) X-validation

(2 marks)

(v) Normalize

(2 marks)

(b) Sketch a detailed process and subprocess in *RapidMiner* for training and testing a data mining model which applies a `split_validation`. The result from the prosess should able to visualize the model and the performance. Assume that the data is cleaned. However no label is yet assigned to one of the attribute.

(10 marks)

Q3 Answer **Q3(a)-Q3(c)** based on Figure **Q3**, Table **1** and Table **2**.

A study on customer expenses is conducted and a dataset is given in Table 1. The study shows either customer will buy a house or not. The decision or the dependent variable is identified in the last column. Summary of the entropy calculation for root node is tabulated in Table 2.

FIGURE Q3

Table 1: Customer Dataset

ID	Age	Income	Government employee	Credit rating	Buy House
1	<=30	High	No	Fair	No
2	<=30	High	No	Good	No
3	31…40	High	No	fair	Yes
4	>40	Medium	No	fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	good	No
7	31…40	Low	Yes	good	Yes
8	<=30	Medium	No	fair	No
9	<=30	Low	Yes	fair	Yes
10	>40	Medium	Yes	fair	Yes
11	<=30	Medium	Yes	good	Yes
12	31…40	Medium	No	good	Yes
13	31…40	High	Yes	fair	Yes
14	>40	Medium	No	good	no

Table 2: Entropy Information for Root Node

Attribute	Average Entropy
Age	0.6935
Income	0.9110
Government Employee	0.7885
Credit Rating	0.8922

(a) Construct a decision tree by calculating all the entropy given in Table **1**.
(10 marks)

(b) Convert the decision tree in **Q3(a)** to production rules.
(8 marks)

(c) What is the result of an old government employee with fair credit rating?
(2 marks)

Q4 (a) Draw and label a schematic diagram of a neural network architecture by considering the optimal number of weights for the network model is 48 based on Figure **Q4(a)**.

(8 marks)

Malaysian rubber industry is having a major problem in detecting symptoms of disease cause by fungus that destroys thousands hectares of Malaysia rubber plantings every year. Only 10,000 raw data had been collected manually in order to diagnose the symptoms whether infected, non-infected or neutral. Some major attributes that express the disease symptoms had also been identified as follows:

(1) The aggressiveness of fungus
(2) Size of fungus
(3) Type of fertilizer
(4) Humidity
(5) location

You have been chosen by Malaysia palm oil industry as data mining expert to solve their problem by using neural network classification task.

FIGURE Q4(a)

(b) Calculate the net output (before activation function) and neuron output (after activation function) for nodes 4, 5, and 6 based on Figure **Q4(b)**. Use initial inputs and weights in Table 3, and initial bias in Table 4.

(12 marks)

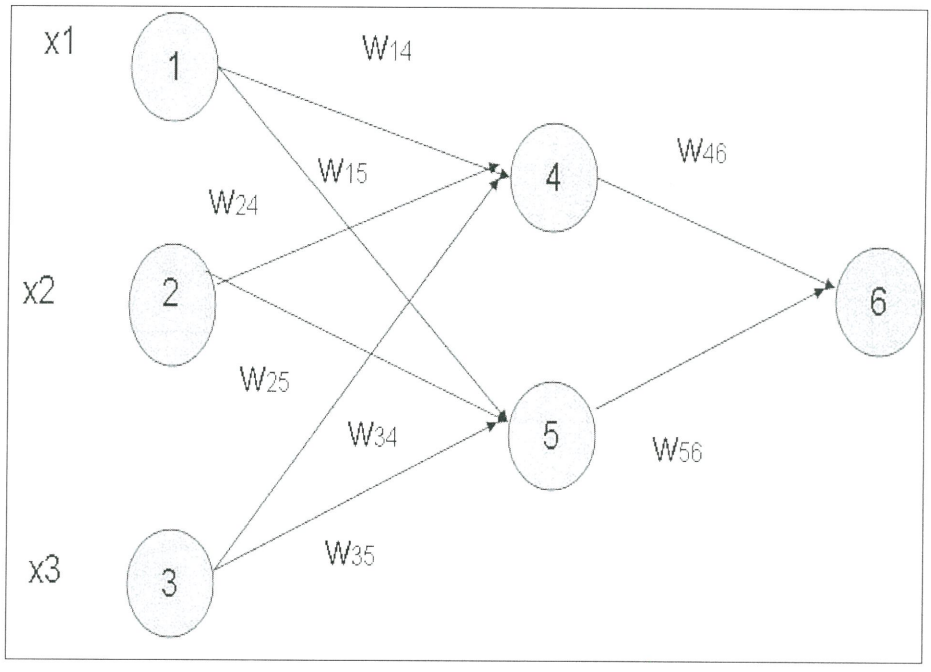

FIGURE Q4(b)

Table 3: Initial Inputs and Weights

x1	x2	x3	w14	w15	w24	w25	w34	w35	w46	w56
1	0	1	0.2	-0.3	0.4	0.1	0.5	0.2	-0.3	-0.2

Table 4: Initial Bias

04	05	06
-0.4	0.2	0.1

Q5 (a) Calculate the support(s) and confidence(c) for each rules below which generated from market basket transactions in Table **5**.

Table 5: Items and their Transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

(i) {Milk,Diaper} → {Beer}

(2 marks)

(ii) {Diaper,Beer} → {Milk}

(2 marks)

(iii) {Diaper} → {Milk,Beer}

(2 marks)

(iv) {Milk} → {Diaper,Beer}

(2 marks)

(b) Answer **Q5(b)(i)** and **Q5(b)(ii)** based on market basket transactions shown in Table **6**.

Table 6: Customer Dataset

Transaction ID	Item
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

(i) Calculate the number of possible candidate itemsets for Table **6**.

(2 marks)

(ii) Using Apriori algorithm, construct step by step procedure to mine frequent itemset by reducing the number of candidates, given the Minimum Support = 2.

(10 marks)

Q6 Answer **Q6(a)-Q6(c)** based on Figure **Q6**.

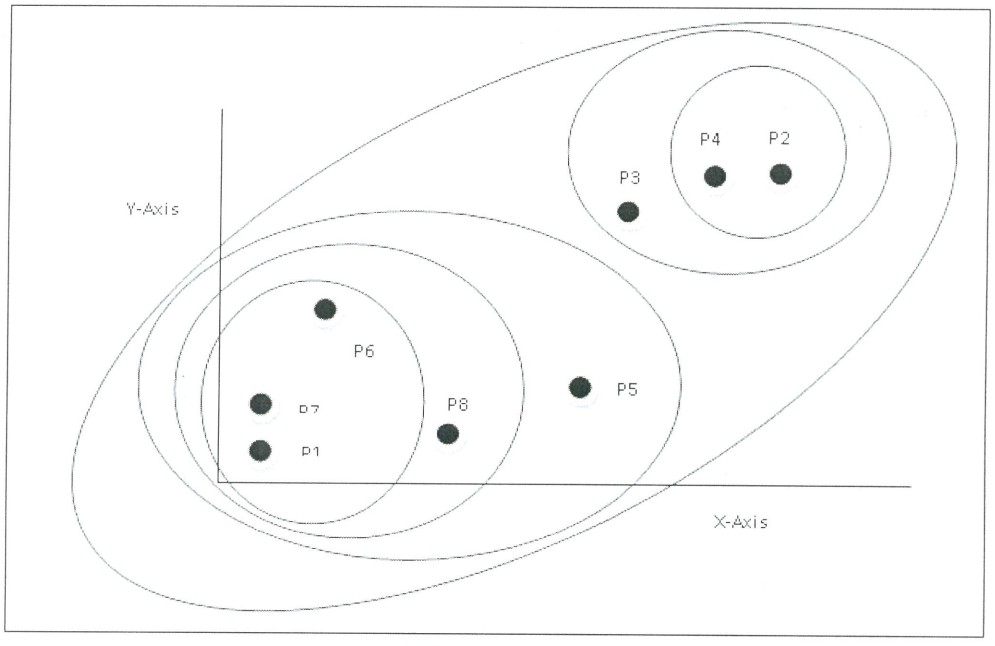

FIGURE Q6

(a) Define the term cluster analysis.

(4 marks)

(b) Draw a *dendogram* for hierarchical clustering as illustrated in Figure **Q6**.

(10 marks)

(c) Discuss **THREE (3)** characteristics of the input data in cluster analysis.

(6 marks)

- END OF QUESTION –