UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2011/ 2012

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 3363 / BIT 33603
PROGRAMME	:	BACHELOR OF INFORMATION TECHNOLOGY
EXAMINATION DATE	:	JANUARY 2012
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **FIVE (5)** FROM **SIX (6)** QUESTIONS.

THIS QUESTION PAPER CONSISTS OF SIX (6) PAGES

Instructions: Answer **FIVE (5)** from **SIX (6)** questions.

Q1 Based on the following scenario:

```
A credit card company wants to promote their new credit card. Due to
mailing costs, the company decided to send the promotional material
to a limited number of credit card holders. Due to its popularity,
the company decided to use data mining techniques to find out what
attributes which can group the customers into two or more clusters.
The company can then check historical records to determine which
group is more likely to have higher acceptance of the offer. Based on
that, the company may decide to mail the promotional material to a
particular group of people only.
```

(a) Why data mining technique is suitable for this scenario?

(5 marks)

(b) Explain which training (supervised or un-supervised) is applicable for clustering task in this scenario.

(5 marks)

(c) Outline **FIVE (5)** challenges of data mining.

(10 marks)

Q2 Today, 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. This methodology makes large data mining projects faster, cheaper, more reliable and more manageable. Even small scale data mining investigations benefit from using CRISP-DM.

(a) Differentiate **TWO (2)** reasons why there should be a standard process in data mining

(2 marks)

(b) Discuss **THREE (3)** benefits of using CRISP-DM.

(6 marks)

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

(12 marks)

Q3 Based on the following scenario:

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.

(a) Describe **FOUR (4)** kinds of data quality problems.

(4 marks)

(b) Outline **ONE (1)** solution for each data quality problems in **Q3(a)**.

(4 marks)

(c) A data string is given in **Figure Q3**. Analyze the pre-process the data into a new data set with `[0.1, 0.9]` range by considering data smoothing and data normalization formula (Equation 1).

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

99
101
56
23
115
84

Figure Q3

(12 marks)

Q4 Based on the following scenario:

A study on customer expenses is conducted and a dataset is given in **Table 1**. The study shows either customer will buy a computer or not. The decision or the dependent variable is identified in the last column. Summary of all the entropy calculation are tabulated in **Table 2, 3** and **4**.

Table 1: Customer dataset

ID	Age	Income	Student	Credit rating	Buy Computer
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	0.0935
Income	0.6110
Student	0.3885
Credit rating	0.5922

Table 3: Entropy information for the branch attribute (<=30)

Attribute	Average Entropy
Income	0.4000
Student	0
Credit rating	0.9510

Table 4: Entropy information for the branch attribute (>40)

Attribute	Average Entropy
Income	0.951
Student	0.951
Credit rating	0

(a) Outline a decision tree using the entropy measure given.

(8 marks)

(b) Convert the decision tree in **Q4(a)** to a production rules.

(8 marks)

(c) Analyze the result of a customer with age greater than 40 and with fair credit rating

(4 marks)

Q5 Based on **Figure Q5**:

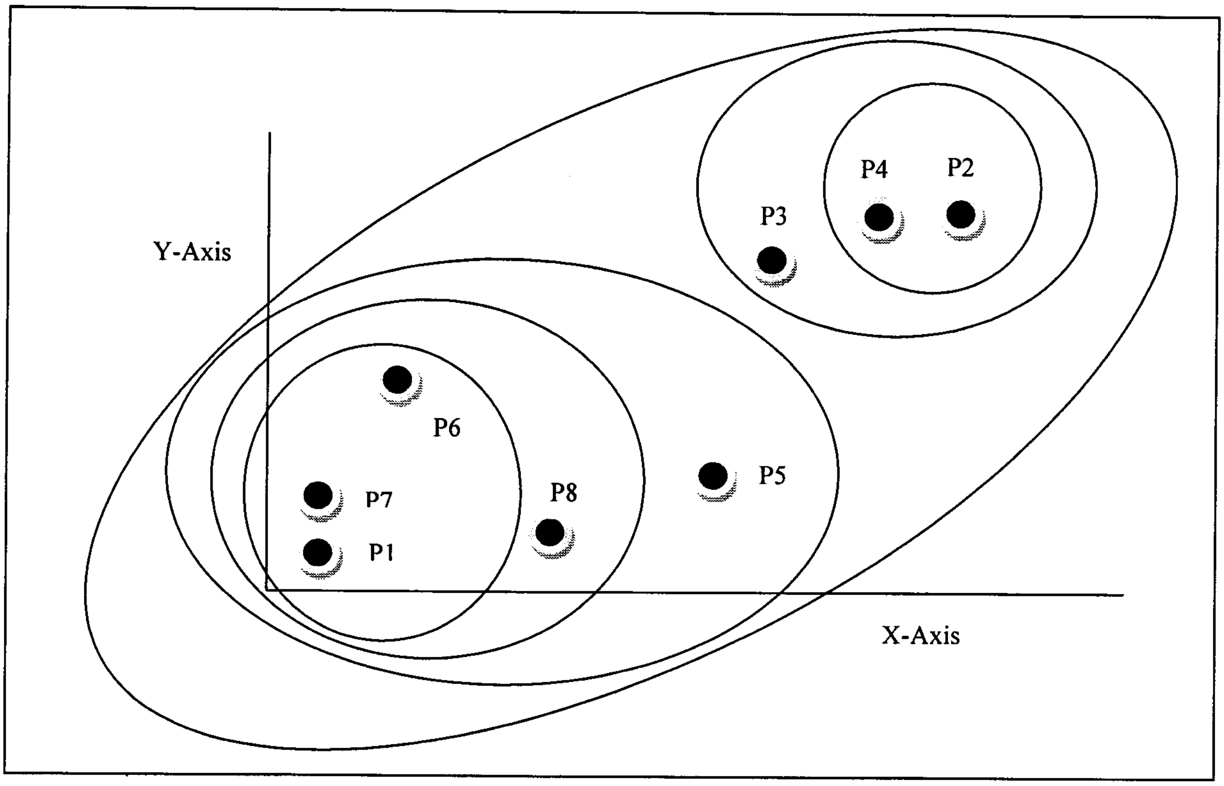

Figure Q5

(a) Discuss the term cluster analysis.

(4 marks)

(b) Outline the ***dendogram*** for hierarchical clustering as illustrated in **Figure Q5**.

(10 marks)

(c) Characteristics of the input data are important in cluster analysis. Differentiate **THREE (3)** of those characteristics.

(6 marks)

Q6 Based on the following scenario:

Malaysian oil palm industry is having a major problem in detecting symptoms of disease cause by fungus that destroys thousands hectares of Malaysia oil palm 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

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

(a) Outline **TWO (2)** steps process needed before using neural network classification task.

(4 marks)

(b) Draw and label a schematic diagram of neural network architecture by considering an optimal weights for neural network is 35.

(8 marks)

(c) Categorize **FOUR (4)** issues regarding evaluating classification methods.

(8 marks)