Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2013/2014

COURSE NAME	:	DATA MINING
COURSE CODE	:	BIT 33603
PROGRAMME	:	3 BIT
EXAMINATION DATE	:	DECEMBER 2013/JANUARY 2014
DURATION	:	2 HOURS 30 MINUTES
INSTRUCTION	:	ANSWER ALL QUESTIONS

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

Q1 In today's competitive world, the introduction of wireless and internet broadband had changed the trend of the telecommunication industry in Malaysia. Customer churn remains to be as one of the most pressing concerns for network providers.

(a) Discuss THREE (3) factors cause the customer churn.

(6 marks)

(b) Give the appropriate technique of churn analysis in data mining to prevent loss in revenue for the company?

(4 marks)

(c) Describe briefly the standard process of data mining methodology to overcome the situation.

(10 marks)

Q2 In retail industry of company ABC, the marketing staff would like to have a better understanding of their different customers. Effective data mining needs expert knowledge about the relevant business process and knowledge about data attributes.

(a) Discuss the need of data mining in retail industry.

(5 marks)

(b) Identify which application of data mining require more such expert knowledge. Clustering or classification task? Give your reason.

(10 marks)

(c) Describe the major challanges in data quality?

(5 marks)

Q3 (a) Explain the ID3 Algorithm for decision trees.

(5 marks)

(b) A study on government employee expenses is conducted and a dataset is given in Table 1 shows either employee will buy a house. The decision of the dependent variable is identified in the last column. Summary of all the entropy calculation are identified in Table 2,3,4.

Table 1: Employee dataset

Age	Income	Employee	Expenses	Buy House
Young	High	No	Fair	No
Young	High	No	Good	No
Middle	High	No	Fair	Yes
Old	Medium	No	Fair	Yes
Old	Low	Yes	Fair	Yes
Old	Low	Yes	Good	No
Middle	Low	Yes	Good	Yes
Young	Medium	No	Fair	No
Young	Low	Yes	Fair	yes
Old	Medium	Yes	Fair	yes
Young	Medium	Yes	Good	yes
Middle	Medium	No	Good	Yes
Middle	High	Yes	Fair	Yes
Old	Medium	No	Good	No

Table 2: Entropy for root node

Attribute	Average Entropy
Age	0.6935
Income	0.9110
Govt Employee	0.7885
Expenses	0.8922

Table 3: Entropy information for the branch node (Young)

Attribute	Average Entropy
Income	0.400
Govt Employee	0
Expenses	0.9510

Table 4: Entropy information for the branch node (Old)

Attribute	Average Entropy
Income	0.9510
Govt Employee	0.9510
Expenses	0

(i) Construct a decision tree using the entropy measure given from the tables. (5 marks)

(ii) Convert the decision tree in 3(b)(i) to a production rules. (5 marks)

(iii) Apply 3(b)(ii) to find the result of an old goverment employee with fair credit expenses?

(5 marks)

Q4 (a) Define the term cluster analysis.

(5 marks)

(b) Differentiate between hierarchical clustering and partitional clustering.

(5 marks)

(c) Describe the algorithm of K-mean clustering.

(4 marks)

(d) Apply the algorithm to determine the following given points into K different clusters using Euclidean distance. (Assume that k=2.)

X	2	5	1	5	2
Y	4	7	2	6	5

(6 marks)

Q5 Consider the market basket transactions shown in Table 5,

Table 5: 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) Use Apriori algorithm, find the maximum number of association rules can be extracted, which satisfy the requirement of 30% support and 70% confidence.

(10 marks)

(b) Find the number of closed frequent itemsets and maximal frequent itemsets from the given Table.

(6 marks)

(c) Explain briefly why FP-Growth Algorithm better than Apriori Algorithm ?

(4 marks)

- END OF QUESTION -