Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2014/2015

COURSE NAME	:	INTRODUCTION TO DATA MINING
COURSE CODE	:	BWB 43303
PROGRAMME	:	3 BWQ
EXAMINATION DATE	:	JUNE 2015 / JULY 2015
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

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

Q1 (a) There are two major tasks in data mining; Predictive and Descriptive Task. Define the meaning of these tasks.

(2 marks)

(b) Give two examples of clustering application in the real world problem and explains.

(4 marks)

(c) There are seven steps of Data Mining. List all these steps.

(7 marks)

(d) For the following vectors x and y, calculate the similarity by using **cosine**, **correlation** and **Jaccard** distance.

(i) $x = (0,1,0,1)$ $y = (1,0,1,0)$

(ii) $x = (1,1,0,1,0,1)$ $y = (1,1,1,0,0,1)$

(12 marks)

Q2 (a) Construct an FP-tree from the data in **Table Q2(a)**.

Table Q2(a): Market Basket Transactions Data

TID	Items
1	a c
2	a d e
3	a b c
4	a b d
5	b c e

(12 marks)

(b) Consider the market basket transactions shown in **Table Q2(b)**.

Table Q2(b): Market Basket Transactions

Customer ID	Transaction ID	Items Bought
1	0001	Milk, Bread, Eggs
1	0024	Milk, Yogurt, Flour, Eggs
2	0031	Milk, Yogurt, Bread, Eggs
2	0015	Milk, Flour, Bread, Eggs
3	0022	Yogurt, Flour, Eggs
3	0029	Yogurt, Bread, Eggs

(i) Compute the support for itemsets {milk}, {bread, eggs} and {milk, bread, eggs} by treating each **transaction ID** as a market basket.

(3 marks)

(ii) Use the results in **Q2(b)(i)** to compute the confidence for the association rules for {milk, bread} $\rightarrow$ {eggs} and {eggs} $\rightarrow$ {milk, bread}.

(2 marks)

(iii) Transform the market basket data in **Table Q2(b)** into the binary transaction by treating each **customer ID** as a market basket. Each item should be treated as a binary variable (1 if an item appears in **at least one transaction** bought by the customer, and 0 otherwise).

(3 marks)

(iv) Compute the support for itemsets {eggs}, {milk, bread} and {milk, bread, eggs} based on your answers in part **Q2(b)(iii)**.

(3 marks)

(v) From the answer in **Q2(b)(iii)**, compute the confidence for itemsets {milk, bread} $\rightarrow$ {eggs} and {eggs} $\rightarrow$ {milk, bread}.

(2 marks)

Q3 **Table Q3(a)** summarizes a data set with three attributes X, Y, Z and two class labels C_1, C_2 .

Table Q3(a): Summarization of three attributes with the class labels

X	Y	Z	**Number of instances**	
			C_1	C_2
F	T	T	0	2
F	F	T	4	0
T	F	T	0	4
F	F	F	2	0
T	T	F	0	0
F	T	F	0	10
T	F	F	0	0
F	F	F	10	0

(a) According to the Gini index, which attribute would be chosen as the first splitting attribute? For each attribute, show the contigency table and the gains in Gini index.

(13 marks)

(b) Repeart for the next two children of the root node. Build the three level decision tree.

(12 marks)

Q4 You are to cluster eight points: $x_1 = (2,10)$, $x_2 = (2,5)$, $x_3 = (8,4)$, $x_4 = (5,8)$, $x_5 = (7,5)$, $x_6 = (6,4)$, $x_7 = (1,2)$, $x_8 = (4,9)$. Suppose, you assigned x_1, x_4 and x_7 as initial cluster centers for K-means clustering . Using k-means with the Manhattan distance, compute the three clusters for each round of the algorithm until convergence.

(25 marks)

-END OF QUESTION-

FINAL EXAMINATION

SEMESTER/ SESSION: SEM II / 2014/2015 **COURSE: 3 BWQ**

SUBJECT: INTRODUCTION TO DATA MINING **CODE: BWB 43303**

Formula

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N}, c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

$$-\sum_{i=0}^{c-1} p(i \mid t) \log_2 p(i \mid t) \qquad 1 - \sum_{i=0}^{c-1} [p(i \mid i)]^2$$

$$1 - \max[p(i \mid i)]^2 \qquad \Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \qquad \cos(x,y) = \frac{x.y}{\| x \| \| y \|}$$

$$corr(x,y) = \frac{\text{cov}(x,y)}{s_x s_y} \qquad Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

$$\text{cov} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y}) \qquad \text{standard deviation}(x) = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$d(x,y) = | x_1 - x_2 | + | y_1 - y_2 |, \qquad d(x,y) = \left(\sum_{k=1}^{n} | x_k - y_k |^r \right)^{1/r},$$

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}, \qquad d(x,y) = \left(\sum_{k=1}^{n} w_k | x_k - y_k |^r \right)^{1/r}$$