CONFIDENTIAL QIT3033

UNIVERSITI UTARA MALAYSIA

FINAL EXAMINATION
FIRST SEMESTER SESSION 2008/2009

CODE/SUBJECT NAME : QIT3033 / KNOWLEDGE ACQUISITIONS IN DECISION MAKING

DATE : 15 NOVEMBER 2008
TIME : 2.30 ~ 5.00 PM (2 ½ HOURS)
VENUE : DP 2/4

INSTRUCTIONS:

1. This book script contains SEVEN (7) questions in TWELVES (12) printed pages, excluding the cover page and appendix.
2. Answer ALL questions in the space provided.

MATRIC NO : ________________________________
(in words ) ( in digits)

I/C NO :

NAME OF LECTURER : ________________________

GROUP : □ TABLE NO: □□□

DO NOT OPEN THIS SCRIPT UNTIL YOU ARE TOLD TO DO SO

CONFIDENTIAL
QUESTION 1 (6 Marks)

Data Mining task falls in two major categories which are Predictive Modeling (Supervised Learning) and Descriptive Modeling (Unsupervised Clustering).

State THREE (3) main differences between the two tasks.

<table>
<thead>
<tr>
<th>Predictive Modeling</th>
<th>Descriptive Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
QUESTION 2 (15 Marks)

You are given THREE (3) problem scenarios. For each of the following problem scenarios;

a. Decide if a solution would best be addressed with Predictive Modeling (Supervised Learning), Descriptive Modeling (Unsupervised Clustering) or Database query.
b. If you decide that Supervised Learning or Unsupervised Clustering is the best answer,
   i. state any initial goal or objective that you would like to test,
   ii. list several attributes you believe to be relevant for solving the problem.

Scenario 1:
You are the Marketing Executive of a retail outlet selling cosmetic products. You need a list of all your female customers below the age of 18 who has bought product A in the last three months, so that you can send leaflet to those customers about the new cosmetic range for teenagers under the same brand.

Answer:
Scenario 2:
You are the Chief Executive Officer (CEO) of a telecommunication company. You need to develop a general profile of your customer so that you can set a strategic plan to market your product according to the needs and characteristics of your customers.

Answer:
Scenario 3:
You are a lecturer in one of the public university. You have been teaching a course with high failure rate for quite a while. This semester you receive a number of new students enrolling in the same course. You would like to predict how well these students will perform in this course, in order to identify those who are weak so that you can concentrate more on them.

Answer:
QUESTION 3 (14 Marks)

Computer are good at learning ‘concepts’. Concepts are set of objects, symbols, or events grouped together because they share certain characteristics. Concepts are the output of data mining session. There are three (3) common concept view; Classical view, Probabilistic view and Exemplar View.

a. What is the main different between Classical View and the other two views? (2 Marks)

b. You are to develop a concept definition for a good employee.

i. List at least THREE attributes that you would use in your definition? (3 Marks)

ii. Give definition of a good employee from a classical point of view. (4 Marks)

iii. State the definition from an exemplar point of view. Give TWO exemplars. (5 Marks)
QUESTION 4 (28 Marks)

Consider the training data shown in the Table 1 for a binary classification problem (C0 and C1).

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Gender</th>
<th>Marital Status</th>
<th>Type of House</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>Married</td>
<td>Terrace</td>
<td>C0</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>Single</td>
<td>Semi-D</td>
<td>C0</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>Single</td>
<td>Semi-D</td>
<td>C0</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>Single</td>
<td>Apartment</td>
<td>C0</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>Single</td>
<td>Bungalow</td>
<td>C0</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>Single</td>
<td>Bungalow</td>
<td>C0</td>
</tr>
<tr>
<td>7</td>
<td>F</td>
<td>Single</td>
<td>Terrace</td>
<td>C0</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>Single</td>
<td>Terrace</td>
<td>C0</td>
</tr>
<tr>
<td>9</td>
<td>F</td>
<td>Single</td>
<td>Semi-D</td>
<td>C0</td>
</tr>
<tr>
<td>10</td>
<td>F</td>
<td>Divorce</td>
<td>Apartment</td>
<td>C0</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>Married</td>
<td>Apartment</td>
<td>C1</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>Married</td>
<td>Bungalow</td>
<td>C1</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>Married</td>
<td>Semi-D</td>
<td>C1</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>Divorce</td>
<td>Bungalow</td>
<td>C1</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>Divorce</td>
<td>Terrace</td>
<td>C1</td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>Divorce</td>
<td>Terrace</td>
<td>C1</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>Divorce</td>
<td>Semi-D</td>
<td>C1</td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>Divorce</td>
<td>Semi-D</td>
<td>C1</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>Divorce</td>
<td>Semi-D</td>
<td>C1</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>Divorce</td>
<td>Bungalow</td>
<td>C1</td>
</tr>
</tbody>
</table>

Table 1

a. Based on the table, complete the following partial decision trees and compute the Gini index (Weighted Average Impurity) for each split.

i. Gender attribute:

```
Parent Node       Count
------------------|-----
Class = C0        10
Class = C1        10
```

```
Node N1
-------
Class = C0
Class = C1
```

```
Node N2
-------
Class = C0
Class = C1
```

Gini Index = 

(6 Marks)
ii. Marital Status attribute:

<table>
<thead>
<tr>
<th>Parent Node</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td>10</td>
</tr>
<tr>
<td>Class = C1</td>
<td>10</td>
</tr>
</tbody>
</table>

Single  
Married  
Divorce

<table>
<thead>
<tr>
<th>Node N1</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node N2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node N3</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

Gini Index =

(8 Marks)

iv. Type of House attribute:

<table>
<thead>
<tr>
<th>Parent Node</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class ≈ C0</td>
<td>10</td>
</tr>
<tr>
<td>Class ≈ C1</td>
<td>10</td>
</tr>
</tbody>
</table>

Terrace  
Semi-D  
Apartment  
Bungalow

<table>
<thead>
<tr>
<th>Node N1</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node N2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node N3</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node N4</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = C0</td>
<td></td>
</tr>
<tr>
<td>Class = C1</td>
<td></td>
</tr>
</tbody>
</table>

Gini Index =

(10 Marks)
b. Based on your answer in question (a), which attribute is better?  

(1 Mark)

c. List THREE characteristics of Decision Tree induction.  

(3 Marks)

QUESTION 5 (9 MARKS)

a. Why do we need to transform a Linear Regression to a Logistic Regression?  

(2 Marks)
b. You have been given a model that could be used to predict the probability of new customers will buy the house offering. Assumes that buying a house as the dependent variable and the linear regression equation is:

\[ ax + c = 0.1003I + 2.4001M - 8.3230G - 0.3211A + 4.5043. \]

Determine which one of these customers will buy a house by transforming the equation to the Logistic Regression.

<table>
<thead>
<tr>
<th>Customer</th>
<th>Customer 1</th>
<th>Customer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income (I)</td>
<td>10k</td>
<td>48k</td>
</tr>
<tr>
<td>Marital Status (M)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gender (G)</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Age (A)</td>
<td>25</td>
<td>27</td>
</tr>
</tbody>
</table>

(7 Marks)
QUESTION 6 (12 MARKS)

Figure 1 shows a fully connected feed-forward neural network. Assume that the training data ‘Stock Price’ range is from RM20 to RM95 and a neural network modeling is applied to predict the ‘Stock Price’. Answer the following questions based on the diagram.

![Diagram of a neural network with nodes and weights](image)

**Figure 1**

a. The input to individual nodes in Neural Network must be numeric and fall within the closed interval \([0,1]\). Convert the attribute ‘Stock Prices’ of RM50, RM65 and RM82 so that it can be used as an input for the Neural Network model.

(6 Marks)
b. Compute the input and output value for Node X, using the weight given in Table 2.

<table>
<thead>
<tr>
<th>Initial weight values for the Neural Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{1x}$</td>
</tr>
<tr>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 2

(4 Marks)

c. By using the output value of Node X, what is the predicted value for the future 'Stock Price'?

(2 Marks)
QUESTION 7 (16 MARKS)

Cluster the following six points (with \((x, y)\) representing locations) \(A1(2, 10)\), \(A2(2, 5)\), \(A3(8, 4)\), \(A4(5, 8)\), \(A5(7, 5)\) and \(A6(1, 2)\) into three clusters. Initial cluster centers are: \(A1(2, 10)\), \(A4(5, 8)\) and \(A7(1, 2)\). Use K-means algorithm to find the three cluster centers after the first iteration.

a. Fill all the values for the first iteration process in the table given by using the Euclidean Distance formula.

(12 Marks)

<table>
<thead>
<tr>
<th>Point</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>((2, 10))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>((2, 5))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>((8, 4))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>((5, 8))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>((7, 5))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>((1, 2))</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Compute the respective new cluster centers for Cluster 1, Cluster 2 and Cluster 3.

(4 Marks)

END OF QUESTIONS