School of Information Technology
IIT Kharagpur

Course Id: IT60107 Data Warehousing and Data Mining

Date: November 23, 2005 **Total Time: 3 Hours**
Max. Marks: 100

Instructions: Answer any five questions. You may answer the questions in any order. However, all parts of the same question must be answered together. Clearly state any reasonable assumption you make.

1. A hospital cum medical research institute is carrying out a study on the nature of different types of fevers. In order to track every patient as he/she keeps coming back to the hospital, a unique id is maintained. For each patient, they keep track of the body temperature at every hour of the day as long as the patient is admitted in the hospital. They also maintain data about the different types of medicine being given to the patient. Patients may be given more than one medicine in a day. Every medicine is administered as many times in a day as the doctor has prescribed. Since there is history of different types of fevers occurring in various districts, states and regions in the country, the hospital research team wants to maintain such residence details of each patient. One of the goals of the research is to determine if there is any relation between the age and gender of the patients with their body temperature when various medicines are administered. Another goal is to determine if there is a relation between the % of population who are farmers, office goers or teachers in the patient's state with the body temperature of the patients when various medicines are administered.
    a. Design a suitable schema for the hospital cum medical research institute, clearly identifying the Fact table(s), Dimension Tables(s), the Facts, the Dimensions, Primary Keys and Foreign Keys of all the tables. Your schema should at least be able to satisfy the above mentioned research requirements. You may consider other suitable attributes for the dimension table(s).
    b. Classify the fact(s) in your fact table(s) as additive, non-additive and semi-additive.
    c. Write an SQL query that runs on your schema and returns today's average, maximum and minimum body temperature for each married male patient.
    d. Draw a cuboid to represent the result of your query.

    [10+3+5+2=20]
2. Consider the 5 transactions given below. If minimum support is 20% and minimum confidence is 80%, determine the frequent itemsets and association rules using the *FP-Tree* algorithm. **[15+5=20]**

Transaction	Items
T1	Bread (500 grams), Jam (1 bottle), Butter (100 grams), Ketchup (1 bottle)
T2	Bread (500 grams), Butter (100 grams), Egg (1 dozen)
T3	Bread (500 grams), Milk (2 litres), Butter (200 grams)
T4	Egg (2 dozens), Bread (500 grams), Ketchup (1 bottle), Milk (3 litres)
T5	Egg (1 dozen), Milk (1 litre), Jam (1 bottle)

3. Consider the following table of transactions. Each row represents a transaction and each column represents an item. If an item is present in a transaction, it is marked as '1', else it is marked as '0'. Determine the Frequent Itemsets using the Dynamic Itemset Counting algorithm. Use intervals of 5 transactions and min_support = 30%. What is the percentage of savings in terms of the number of database accesses you are achieving in this case over running a priori algorithm? **[18+2=20]**

A1	A2	A3	A4	A5	A6	A7	A8	A9
1	1	1	1	1	1	0	1	1
0	1	0	1	0	0	0	1	0
0	0	0	1	1	0	1	0	0
0	1	1	0	1	0	1	0	0
0	0	0	0	1	1	1	0	0
1	1	1	1	0	1	0	0	1
0	1	0	0	0	1	1	0	1
0	0	0	0	1	0	0	0	0
0	1	0	0	1	0	0	1	0
0	0	1	0	1	0	1	0	1
0	0	1	0	1	0	1	0	0
0	0	0	0	1	1	0	1	1
0	1	0	1	0	1	1	0	0
1	0	1	0	1	0	1	0	0
0	1	1	0	0	0	1	1	1

4. Consider the following set of transactions for a number of customers. Determine the maximal sequences that have at least 40% support. **[20]**

Customer Id	Transaction Date	Items Bought
1	10/11/2005	1, 2
1	11/11/2005	1, 2, 3
1	12/11/2005	2, 3
2	10/11/2005	1, 2
2	11/11/2005	2, 4
3	10/11/2005	1, 2
3	11/11/2005	2, 4
3	12/11/2005	1, 2, 3
3	13/11/2005	1, 2
4	10/11/2005	1, 3
4	11/11/2005	2, 3
4	12/11/2005	1, 2, 3, 4
5	10/11/2005	2, 4
5	11/11/2005	3, 4

5. Build a Decision Tree using the training data in the table given below. Divide the Height attribute into ranges as follows: (0,1.7], (1.7,1.9], (1.9, 2.5] **[20]**

Gender	Height	Class
F	1.60 m	Short
M	1.95 m	Tall
F	1.89 m	Medium
F	1.88 m	Medium
F	1.68 m	Short
M	1.85 m	Medium
F	1.60 m	Short
M	1.69 m	Short
M	2.20 m	Tall
M	2.10 m	Tall
F	1.80 m	Medium
M	1.95 m	Medium
F	1.89 m	Medium
F	1.80 m	Medium
F	1.75 m	Medium

6. Consider a multilayer perceptron (MLP) having 2 units in the input layer, 2 units in the hidden layer and one unit in the output layer. We want to train this MLP with the truth table of an XOR gate. Let us denote the units in the input layer by the subscript **k**, those in the hidden layer by the subscript **j** and that in the output layer by the subscript **i**. Consider that the input layer to hidden layer weights are initially set as follows: w_{11}=1, w_{12}=2, w_{21}=1 and w_{22}=1. Hidden layer to the output layer wights are initially set as follows w_{11}=1 and w_{21}=2. Consider that the transfer functions for the hidden layer units as well as the output layer units are as follows:

   $$\text{Output} = \frac{1}{1+e^{-Input}}$$

   Assume that the input layer units transfer their inputs without any change.

   a. Determine the new weights after an input pattern (1 0) is given as the training data.
   b. Consider that the transfer function for the hidden layer units is changed as follows, with the transfer function for the output layer remaining unchanged:

      $$\text{Output} = \frac{1-e^{-Input}}{1+e^{-Input}}$$

      Derive an expression for the input layer to hidden layer weight increments Δw_{kj}'s in terms of the MLP inputs, hidden layer outputs, desired outputs and weights.

      [12+8=20]