School of Information Technology, IIT Kharagpur

IT60107 Data Warehousing and Data Mining **End Semester Exam**

Date: November 27th, 2004. **Total Time: 3 Hours** **Max. Marks: 100**

Answer any **five** questions. Clearly state any reasonable assumptions you make.

1. A data warehouse is to be built for a weather monitoring station which will record temperature, pressure, humidity and wind velocity for different latitude, longitude, altitude at each second and maintain data for hundreds of years. (a) Suggest an efficient star schema design for this data warehouse clearly identifying the fact table(s) and dimension tables(s), their primary keys and foreign keys. (b) For your schema, write an SQL statement to retrieve average hourly temperature for each week of the year 2002. An example is shown below to explain the expected output.

Week	Hour	Av. Temp
1	1	XXX
1	2	XXX
…….	…….	…….
1	24	XXX
2	1	XXX
2	2	XXX
…….	…….	…….
2	24	XXX
……..		
……..		
52	1	XXX
52	2	XXX
…….		
52	24	XXX

[15+5=20]

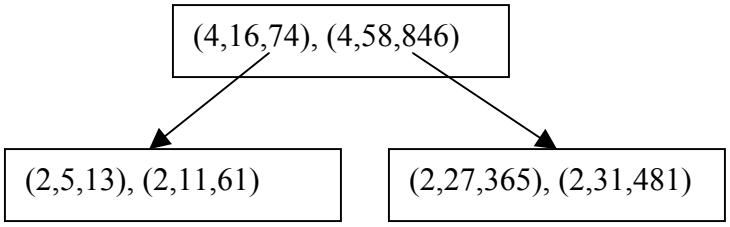
2. Build a Decision Tree for classification using the training data in the table given below. Divide the Height attribute into ranges as follows:
(0,1.6], (1.6,1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, 5.0] **[20]**

Gender	Height	Class
F	1.56 m	Short
F	2.10 m	Tall
F	1.90 m	Tall
F	1.88 m	Tall
F	1.65 m	Short
M	1.85 m	Medium
F	1.59 m	Short
M	1.69 m	Short
M	2.20 m	Tall
F	2.10 m	Tall
F	1.80 m	Tall
M	1.95 m	Medium
F	1.75 m	Medium
M	1.82 m	Medium
F	1.78 m	Medium

3. We want to cluster five documents into two clusters having inter-document distances shown below. Assume that at a certain step, documents C and E are selected as medoids. Which two documents will be selected as medoids in the next iteration using the PAM algorithm? You must explain the steps followed in arriving at your answer. **[20]**

Document	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

4. Consider the BIRCH clustering technique using CF-tree. Assume that there can be at most **2** entries in each leaf node as well as in each non-leaf node and the diameter threshold is **1**. At one stage, following is the structure of the CF-tree.



The next two data items to be added to the tree are 3 and 7. Show (a) the structure of the tree after you add 3 and (b) the structure of the tree after you add 7 (after adding 3). You must explain the steps followed in arriving at your answer. **[15+5=20]**

5. Consider the 5 transactions given below. If minimum support is 60% and minimum confidence is 80%, determine using the *FP-tree* algorithm (a) the frequent itemsets and (b) the association rules. **[15+5=20]**

Transaction	Items
T1	Bread, Jelly, Butter
T2	Bread, Butter, Cheese
T3	Bread, Milk, Butter, Coke, Cheese
T4	Coke, Bread, Jelly
T5	Coke, Milk, Cheese

6. A chain of departmental stores called IndiaMart having head office at Mumbai, has implemented a data warehouse distributed in three of its locations – Kolkata, Delhi and Chennai. The data at the three locations are shown below. 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'. (a) Suggest a method by which the top management at Mumbai can determine the association rules valid for the whole of IndiaMart without actually bringing all the detailed transaction-level data to Mumbai. (b) Determine the association rules using the suggested method. Consider minimum support to be 40% and minimum confidence to be 80%.

[5+15=20]

Kolkata

A1	A2	A3	A4
1	0	0	0
0	1	0	1
0	0	0	1
0	0	1	0
0	1	0	1

Delhi

A1	A2	A3	A4
0	1	1	1
1	1	0	0
0	0	1	0
0	1	1	1
0	0	1	0

Chennai

A1	A2	A3	A4
0	1	1	0
0	1	0	0
0	1	0	1
1	0	1	0
0	1	1	0