School of Information Technology
IIT Kharagpur

Course Id: IT60107 Data Warehousing and Data Mining (End Semester Examination)

Date: November 22, 2011
Total Time: 3 Hours **Max. Marks: 100**

Instructions: Answer any 5 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. Consider the 5 transactions given below. If minimum support is 40%, (a) determine the frequent itemsets using either the FP-Tree algorithm OR the Dynamic Itemset Counting algorithm (using three stops). (b) Determine the association rules considering minimum confidence of 60%. While determining the association rules only consider the frequent 3-itemsets and frequent 4-itemsets (if any). **[15+5=20]**

Transaction	Items
T1	Bread, Jam, Milk, Butter
T2	Bread, Milk, Butter, Ketchup
T3	Jam, Milk, Ketchup
T4	Bread, Jam, Milk, Butter
T5	Jam, Milk
T6	Jam, Milk, Butter

2. Consider the following set of transactions for a number of customers. Determine the maximal sequences that have at least 50% support. **[20]**

Transaction Date	Customer Id	Items Bought
1/1/2011	1	A, B, C
1/1/2011	2	A, B
1/1/2011	3	A, C
1/1/2011	4	B, D
2/1/2011	1	D
2/1/2011	2	C, D
2/1/2011	3	D
3/1/2011	1	B, E
3/1/2011	2	D
4/1/2011	3	B, E
4/1/2011	4	E
5/1/2011	1	D, E

3. Cluster the following 5 data points using CLARANS with maxneighbor = 2 and numlocal =2. Assume that the initial set of selected medoids is {2, 5}. Show the steps in detail but you need not explain them. They should be self-explanatory.

   2, 5, 8, 10, 11 **[20]**

4. (a) Construct a CF Tree from the following data points (the data points are to be considered in this order): 20, 100, 80, 10, 15, 20. Consider both branching factor as well as maximum no. of leaf node entries to be 3. Consider the diameter threshold to be 8. (b) Once the complete CF tree is built, determine the centroid of the leftmost entry in the leftmost child node of your CF tree. **[15+5=20]**

5. (a) Build a Decision Tree for classification using the training data in the table given below. Divide the Height attribute into 3 ranges as follows: Less than 1.6, 1.6-1.8, Greater than 1.8

Gender	**Height**		**Class**
F	1.58		Tall
M	1.58		Medium
M	1.7		Medium
F	1.65		Tall
F	1.85		Tall
F	1.4		Short
M	1.4		Short
M	1.7		Medium
F	1.75		Tall
M	1.82		Tall
F	1.6		Tall

OR

(b) We wish to train a multilayer perceptron (MLP) with the truth table of an OR gate. Consider this MLP has 2 units in the input layer (corresponding to two inputs of the OR gate), 2 units in the hidden layer and 2 units in the output layer (one unit represents class 0 and the other unit represents class 1). Consider that the input layer to hidden layer weights are initially set as follows: w_{11}=0.4, w_{12}=0.1, w_{21}=0.2 and w_{22}=0.3. Hidden layer to the output layer weights are initially set as follows: W_{11}=0.2, W_{12}=0.2, W_{21}=0.1 and W_{22}=0.1. Consider that the transfer functions for the hidden layer units as well as the output layer units are as follows: $y = 1/(1+e^{-s})$. Assume that the input layer units transfer their inputs without any change and $\eta = 0.5$.

Determine the new weights after an input pattern (1 0) is given as the training data. The expected output is 1. **[20]**

6. Explain the following concepts in brief with suitable examples : **[5x4=20]**
   a. View Materialization
   b. Star Schema
   c. CF vector and its usefulness in computing inter-cluster and intra-cluster distances
   d. Methods for handling slowly changing and rapidly changing dimensions