No of Pages : 3 **Course Code : 15XW81**

Roll No:
(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, SEPTEMBER / OCTOBER 2019

MSc - SOFTWARE SYSTEMS Semester: 8

15XW81 DATA MINING

Time : 3 Hours **Maximum Marks : 100**

INSTRUCTIONS:
1. Answer **ALL** questions. Each question carries 20 Marks.
2. Subdivision (a) carries 3 marks each, subdivision (b) carries 7 marks each and subdivision (c) carries 10 marks each.

1. a) Define nominal, ordinal and ratio scaled variable with examples.

   b) For the following group of data calculate

   200,400,800,1000,2000

   i. Mean and Variance

   ii. Normalize the data using min max normalization

   iii. In z score m=normalization what value should the first number 200 be transformed to?

   c) The following list gives the scores of the same students in a midterm exam. 40, 36, 40, 40, 28, 36, 34, 34, 36, 34, 34, 38, 34, 24, 12

   Draw the boxplot of the midterm scores. Write your observations. Examine the data for possible outliers

2. a) What do you mean by maximal and closed itemset. Give an example.

   b) i) Consider the following set of frequent 3-itemsets:

   {1,2,3}, {1,2,4}, {1,2,5}, {1,3,4}, {1,3,5}, {2,3,4}, {2,3,5}, {3,4,5}

   Assume that there are only five items in the data set.List all candidate 4-itemsets obtained by the candidate generation procedure.(joining step) in Apriori. List all candidate 4-itemsets that survive the candidate pruning step (pruning step) of the Apriori algorithm.

   ii) The confidence for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean? For the dataset shown below,

   a) Compute the support for itemsets {e}, {b,d} and {b,d,e} by treating each transaction ID as a market basket.

   b) Compute the confidence for the association rules {b,d}→ {e} and {e} → {b,d}. Is confidence a symmetric measure?

No of Pages : 3 **Course Code : 15XW81**

Transaction ID	Items bought
0001	{a,d,e}
0024	{a,b,c,e}
0012	{a,b,d,e}
0031	{a,c,d,e}
0015	{b,c,e}
0022	{b,d,e}
0029	{c,d}
0040	{a,b,c}
0033	{a,d,e}
0038	{a,b,e}

c) i) What is Association rule mining? Explain using terms as support, confidence and frequent itemset. On what factors does FP Growth algorithm perform better than Apriori for Association rule mining? For the given transaction data set construct the FP tree and give a step by step process of generating the frequent item sets. (Show the conditional pattern base and conditional FP tree for every item).Minimum support=2.

Transaction Data Set

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

(OR)

ii) Briefly explain the Apriori principle of frequent itemset mining. Why is it so important for frequent itemset mining? For the above dataset with min support be 50% and the minimum confidence be 60%, find all frequent itemsets. Which of them are closed and which are maximal?

3. a) What is the basic idea behind bagging. Why does it perform better than a single classifier?

b) Explain the methods for evaluating the accuracy of a classifier. Why is 10 fold cross validation the most preferred method for evaluation? Consider the confusion matrix of the classifier M

Data set 1		Predicted class	
		+	-
Actual class	+	45	5
	-	10	40

Calculate the any two metrics you can for comparing classifiers based on the above data.

c) Explain the fundamental difference between the Bagging and Boosting ensemble learning methods? Write the Adaboost algorithm for ensemble learning and explain.

4 a) Define a core object and density reachability. Is density reachability a symmetric function?

b) i) Given 5-dimensional numeric samples A= (1, 0, 2, 5, 3) and B= (2, 1, 0, 3, -1), find

1) The Euclidean distance between points

2) The city block distance

3) The Minkowski distance for p=3

ii) How does DBSCAN form clusters? What is a border point in DBSCAN? Assume we have a border point b which is in the radius of 3 different core points c1, c2, and c3. How does DBSCAN deal with this situation—to which cluster or cluster(s) is b assigned or is it treated as an outlier?

c) Give two advantages of Hierarchical clusterng over kmeans clutering. Use single link and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendograms.

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

5. a) What is the significance of robots.txt?

b) Explain in detail the Web Crawling operation.

c) Write short notes on Web mining and Spatial Data mining

/END/

FD/RL