No of Pages : 4 **Course Code : 15XT93**

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

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, NOVEMBER 2019

MSc – THEORETICAL COMPUTER SCIENCE Semester : 9

15XT93 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.
3. Course Outcome : Table

Qn.1	CO.1	Qn.2	CO 2	Qn.3	CO 3	Qn.4	CO 4	Qn.5	CO 5

1 a) Differentiate between filter, wrapper and embedded methods for feature selection.

b) i) Assume a normalized attribute of an object has a z-score of -2; what does this say about the object's attribute value in relationship to the values of other objects?

ii) Suppose that you are given the following data set (in increasing order) for data analysis:

{1, 2, 4, 7, 11, 18, 24, 25, 32, 36, 37, 38, 41, 60}, n = 14

- Give the five-number summary of the data.
- Show a box-plot of the data. Examine the data for possible outliers

c) Suppose that a data warehouse for Big-University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg_grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

i) Draw a snowflake schema diagram for the data warehouse.

ii) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big-University student.

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

b) 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.

i) 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 original association rule mining formulations uses the support and confidence measures to prune uninteresting rules. Here, we focus on the following rules:

- {b} -> {c}
- {a} -> {d}
- {b} -> {d}
- {e} -> {c}
- {c} -> {a}

Compute and rank the rules in decreasing order according to the following measures: Support, Confidence and Lift (Interest).

Transaction ID	Items Bought
1	(a,b,d,e)
2	(b,c,d)
3	(a,b,d,e)
4	(a,c,d,e)
5	(b,c,d,e)
6	(b,d,e)
7	(c,d)
8	(a,b,c)
9	(a,d,e)
10	(b,e)

c) i) 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). Let the min support be 50% and the minimum confidence be 60%.

List all frequent itemsets along with their support. Which of them are closed and which are maximal? For all itemsets that are maximal, list the strong association rules.

TransID	**Items**
T100	A, B, C, D
T200	A, B, C, E
T300	A, B, E, F, H
T400	A, C, H

(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? For all itemsets that are maximal, list the strong association rules.

3. a) What is the basic idea behind bagging? When applying bagging to trees, what extra feature is used (over basic bagging) in the random forest implementation? What is its purpose?

b) i) Suppose we produce 10 bootstrapped samples from a data set with two classes $Y \in \{0,1\}$ .We grow a classification tree to each bootstrap sample and for a specific value X, produce 10estimates of P(Y = 0|X):

(0:1; 0:15; 0:2; 0; 2; 0:55; 0:6; 0:6; 0:65; 0:7; 0:75

There are two common methods for generating classifications given this data:

Take the average of these probabilities and round that to 0 or 1

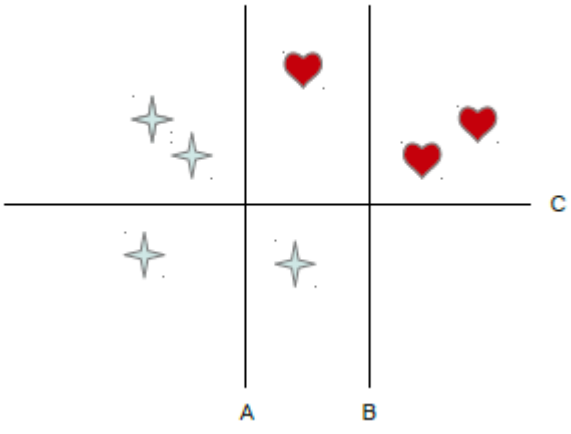
Take a majority vote for the rounded probabilities of each tree.

What would the classification based on each of these methods? Justify.

ii) What are the differences between the three commonly used ensemble learning techniques, stacking, boosting, and bagging?

c) Write the Adaboost algorithm for ensemble learning and explain.

The diagram shows training data for a binary concept where positive examples are denoted by a heart. Also shown are three decision stumps (A, B and C) each of which consists of a linear decision boundary. Suppose that AdaBoost chooses A as the first stump in an ensemble and it has to decide between B and C as the next stump. Which will it choose? Explain. What will be the ε and α values for the first iteration?



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

b) i) How do you calculate the similarity for nominal and ordinal attributes? For the given data set where Gender and Eye colour are nominal attributes, calculate which two persons are most similar

	Gender	Eye colour
Person 1	Male	Blue
Person 2	Male	Black
Person 3	Female	Green
Person 4	Male	Blue

No of Pages : 4 **Course Code : 15XT93**

ii) Discuss conditions under which density-based clustering is more suitable than partitioning-based clustering and hierarchical clustering. Use two examples to illustrate and support your argument. 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 clustering over kmeans clustering. For the following data, cluster using single link and complete link clustering techniques. Use Manhattan distance for distance computation.

A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5)

5. a) Compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. What does the first three moments of a stream signify?

b) i) Give two real world examples where time series mining can be applied. What is dynamic time warping? Explain with an example.

ii) There are three categories of anomalies: point, contextual, and collective. Give an example in each category. What are the differences between distance-based outlier detection and density-based outlier detection?

c) What factors make Birch advantageous over other algorithms in dealing with large databases. Explain BIRCH algorithm with example.

/END/

FD/RL