No of Pages : 4 Course Code : 15XW81

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

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, APRIL 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.
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) 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?.

b) i) Assume the following data is given: { 21,21,4,9,8,15,24,25,34,29,26,28}. Apply data discretization by binning the data into 3 bins using equal-width and equi-depth binning, respectively. Also Smooth the data by bin means and bin boundaries (3)

ii) A year ago, Anitha began working at a computer store. Her supervisor asked her to keep a record of the number of sales she made each month.

The following data set is a list of her sales for the last 12 months:

34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37.

Like Anitha, Krishna works at a computer store. He also recorded the number of sales he made each month. In the past 12 months, he sold the following numbers of computers:

51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13.

- Give a five-number summary of Krishna's and Anitha's sales.
- Make two box and whisker plots, one for Anitha's sales and one for Krishna's.
- Briefly describe the comparisons between their sales. (4)

c) What is data discretization? For the following data set discretize using the entropy based method.

Hours studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

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

b) i) Assume the APRIORI algorithm identified the following 7 4-item sets that satisfy a user given support threshold: abcd,abce,abcf, acde,adef, bcde, and bcef; what are initial candidate 5-itemsets created by the APRIORI algorithm? Which of those survive subset pruning? (3)

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) based on the table of transactions below. (4)

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.

Trans ID	**Items**
T100	ABCE
T200	ACDE
T300	BCE
T400	ACDE
T500	CDE
T600	ADE

(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 of 50% and the minimum confidence of 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) Ensembles use multiple classifiers to make decisions. What properties should a set of base classifiers have to form a good ensemble? What is the basic idea behind bagging?

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 10 estimates 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. (3)

ii) Consider a star schema with four dimensions A, B, C, and D. Suppose a query involves one row of A and B each. How many rows of the fact table will be in the result set, assuming that each dimension has 500 rows and the fact table records allowable events? (4)

c) Suppose that a data warehouse consists of the three dimensions time, doctor, patient and the two measures count and charge , where charge is the fee that a doctor charges a patient for a visit.

(i) Enumerate three classes of schemas that are popularly used for modeling data warehouses.

(ii) Draw a schema diagram for the above data warehouse using star schema.

(iii) Starting with the base cuboid [day,doctor,patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

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

b i) 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 is to which cluster or cluster(s) is b assigned or is it treated as an outlier? (3)

ii) How do you calculate the similarity for nominal and ordinal attributes? For the given data set of 4 people buying a product, the attributes are gender and satisfaction(ordinal) with range -2(very dissatisfied) to 2 (very satisfied), calculate which two persons are most similar by aggregating the similarity scores. (4)

	Gender	Satisfaction
Person 1	Male	2
Person 2	Female	1
Person 3	Male	2
Person 4	Male	-1

c) Give two advantages of Hierarchical clusterng over kmeans clutering. For the following data, cluster using complete link clustering techniques. Use Manhattan distance for distance computation.

A1(2, 10) A2(2, 6) 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) Write short notes on

i) Spatial data mining

ii) Outlier analysis

c) What factors make Birch advantageous over other algorithms in dealing with large databases? Explain the construction of CF tree with an example.

/END/

FD/RL