No of Pages : 3 **Course Code : 09XT83**

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

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, AUGUST / SEPTEMBER - 2015

MSc – THEORETICAL COMPUTER SCIENCE Semester: 8

09XT83 DATA MINING

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

INSTRUCTIONS:

1. Answer **ALL** questions from PART - A and Answer any **4** questions from PART - B. Question under PART - C is Compulsory

PART - A **Marks : 10 x 3 = 30**

1. Define nominal, ordinal and ratio scaled variable.
2. We want to select two features from the given three using information gain .Which features would we choose? Do you feel they are the best features by looking at the data? Price is the class.

B	C	S	Price
high	low	low	up
high	high	low	up
high	low	high	down
low	low	low	down
low	high	high	up

3. When can we say the association rules are interesting? Can we say that a rule {a}→{b} is symmetric?
4. Name two statistical methods used in Data Mining and indicate an application area in each case.
5. What is entropy and what is its significance in attribute selection in classification?
6. What is “bagging” and how is it different from “boosting?” When would you use either of these techniques?
7. What is cross-validation and why is it important?
8. The k means algorithm relies on iterating between two steps. List these two steps
9. Name one way that the clusters found by the agglomerative clustering algorithm differ to those found by the k-means clustering algorithm.
10. How do you define a noise object in DBSCAN? What is the advantage that OPTICS have over DBSCAN? Define the terms core object and reachability distance with respect to OPTICS.

PART - B **Marks : 4 x 12.5 = 50**

11. 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.

    a) List all candidate 4-itemsets obtained by the candidate generation procedure (joining step) in Apriori.

b) List all candidate 4-itemsets that survive the candidate pruning step (pruning step) of the Apriori algorithm.

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

(1) {b} -> {c}

(2) {a} -> {d}

(3) {b} -> {d}

(4) {e} -> {c}

(5) {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)

12. For the data set given below which decides whether a given type of food is appealing or not, answer the following questions

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	H	Sour	Small
No	H	Salty	Large
Yes	H	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	H	Salty	Large

i) What is the initial entropy of Appealing?

ii) Assume that Taste is chosen for the root of the decision tree. What is the information gain associated with this attribute?

iii) Draw the full decision tree learned for this data (without any pruning).

13. What is Ensemble learning? Explain any two ensemble learning methods? How do these notions relate to the concept of generating a good ensemble? What are the advantages and disadvantages of each method?

No of Pages : 3 **Course Code : 09XT83**

14. Give one advantage of hierarchical clustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering. Usesingle-link and complete-link agglomerative clustering to cluster the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4). Use Manhattan distance. Show the dendrograms.

15. Write short notes on the following giving real time examples

    a) Outlier analysis

    b) Text Mining

    c) Web Mining

PART - C **Marks : 1 x 20 = 20**

16. 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) Assume Min support=40%

TID	items_bought
T100	{ I6, I1, I3}
T200	{ I1, I2, I4, I5, I3}
T300	{ I3, I2, I5}
T400	{ I6, I7}
T500	{ I1, I3, I2, I4, I5}
T600	{ I1, I3, I6}
T700	{ I1, I2, I5, I7}
T800	{ I2, I8, I5, I1}
T900	{ I4, I6}
T1000	{ I1, I2, I5 }

/END/

FD/RL