Total No. of Printed Pages:04

B.E. (Computer) Semester- VIII (Revised Course 2007-08)
EXAMINATION MAY/JUNE 2019
Elective-III - (4) Data Mining.

[Duration : Three Hours] **[Max.Marks :100]**

Instructions:-
1. Answer any Five full questions at least one full question from each module.
2. Make suitable assumptions wherever necessary.

Module I

Q.1 a) Discuss whether or not each of the following activities is a data mining task. **05**
i. Monitoring seismic waves for earthquake activities.
ii. Predicting the outcomes of tossing a (fair) pair of dice.

b) Draw a neat labelled diagram and describe the complete data mining process. **05**

c) Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity. **05**
i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
ii) Time in terms of AM or PM.

d) What is dimensionality reduction? Propose different techniques used for dimensionality reduction? Explain one technique in detail. **05**

Q.2 a) The following attributes are measured for members of a herd of Asian elephants: weight, height, tusk length, trunk length, and ear area. Based on these measurements, what sort of similarity measure would you use to compare or group these elephants? Justify your answer and explain any special circumstances. **05**

b) Write the four factors which test the interestingness of the patterns. **02**

c) Explain discretization with examples in the context of data preprocessing. Compare it with binarization. **05**

d) What is the Euclidean and cosine distance between each of the (first, second and third) vectors (1,0,0), (1,4,5), and (10,0,0)? **06**

e) State methods used for noise elimination in data preprocessing. **02**

Module II

Q.3 a) Explain the different visualization techniques used in data exploration. **04**

b) Suppose that a data warehouse consists of the four dimensions date, spectator, location and game, and the two measures count and charge, where charge is a fare that the spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. Draw the star schema diagram for the data warehouse. **08**

c) What are the characteristics of overfitting when learning decision trees? Assume you observe overfitting, explain any two approaches which could be taken in order to learn a "better" decision tree? **08**

Q.4 a) Provide examples and state reasons to demonstrate the difference between star, snowflake and fact constellation schemas. **06**

b) Refer Table 1. Consider X as College_Major and Y as likes "Programming". **04**

Table 1

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Math	Yes
History	No

Calculate entropy of Y as H(Y) and the entropy of X as H(X).

c) Write a decision tree algorithm and explain different measures used for best split in decision trees. **08**

d) State different multivariate summary statistics. **02**

Module III

Q.5 a) Consider the database given in Table 2: **08**

Table 2

Tld	Items bought
100	A,B,C,D,E,F
200	M,B,C,D,E,F
300	A,F,D,E
400	A,U,H,D,F
500	H,B,B,D,E

Assume that the minimum support threshold is 60% and minimum confidence threshold is 80%. Find association rules using Apriori algorithm.

b) Discuss the fp-tree technique by providing its algorithm. How does it differ from apriori based association rule mining method? **08**

c) What is the role of the learn-one-rule function? **02**

d) What are maximal frequent itemsets? **02**

Q.6 a) Write the apriori principle? Draw a fp-tree for the database given in Table 2. **08**

b) Why is the k-nearest neighbor called as a lazy classifier? Explain with the help of an example by comparing it with the decision tree classifier. **06**

c) State different approaches used for improving the efficiency of apriori algorithm. Describe two approaches in detail. **06**

Module IV

Q.7 a) Consider the following distance matrix giving distance between the points given in Table 3: **06**

Table 3

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Show the hierarchy of clustering created by the single-link clustering algorithm. State point of comparison between single linkage with complete linkage.

b) Write the phases of BIRCH clustering. Consider the following points to construct CF. (3,4) (2,6) (4,5) (4,7) (3,8) **06**

c) Explain the DBSCAN method. State its advantages over k-means clustering. **06**

d) State strengths and weakness of proximity based outlier detection. **02**

Q.8 a) Suppose that the data mining task is to cluster the following eight points with (x;y) representing location into three clusters. **09**
A1(2;10); A2(2;5); A3(8;4); B1(5;8); B2(7;5); B3(6;4); C1(1;2); C2(4;9)
Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show:
i) The three cluster centers after two rounds of execution.
ii) List the strengths and weakness of k-means
iii) Write k-means algorithm

b) How do you differentiate noise and outlier? Why is outlier mining important? Describe the following different approaches used for outlier analysis: **09**
i) Statistical-based outlier detection.
ii) Distance-based outlier detection.
iii) Density-based outlier detection.

c) Demonstrate the working principle of bisecting k-means. **02**