BE899

Total No. of Printed Pages: 05

B.E.(Computer) Semester-VIII (Revised Course 2007-08)
EXAMINATION OCTOBER-2020
Data Mining

[Duration : Two Hours] **[Total Marks : 60]**

Instructions:
1) Answer THREE FULL QUESTIONS with ONE QUESTION from ANY THREE MODULES.
2) Make suitable assumptions if necessary, state clearly assumptions made.

MODULE-I

Q.1 a) Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects. (8)

b) What is Discretization and why is it needed? Explain giving an example. (4)

c) Consider the following data for the 'price' attribute :{4,8,9,15,21,21,24,25,26,28,29,34}. Partition the same into 3 bins using: i) Equi-depth binning ii) Smoothing by bin means iii) Smoothing by bin boundaries. (6)

d) Write the four factors which test the interestingness of the patterns. (2)

Q.2 a) Draw a neat labeled diagram and describe the steps involved in data mining when used for the process of knowledge discovery. (8)

b) Explain the following: (Any Two) (6+6)
i. Regression
ii. Principal Component Analysis.
iii. Aggregation

MODULE –II

Q.3 a) What is entropy? State the significance of entropy in classification. (4)

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. (8)

c) With the help of an error graph explain the state of over fitting in decision trees. Explain the idea of pre-pruning and post-pruning and the related methods to fix the same. (8)

Q.4 a) Provide exhaustive comparison of OLTP and OLAP. Describe the data cube model in detail. (6)

b) Assume training database given in Table 1. It has attributes Age and Car Type. (7+7)
Age – Ordinal Attribute
Car Type- Categorical Attribute
Class-L: low and H: high (risk)
Table 1

Age	Car Type	Class
>21	Maruti	L
>21	Hyundai	H
<21	Maruti	H
<21	Indica	H
>21	Maruti	L
>21	Hyundai	H

Write ID3 algorithm for classification using decision tree. Generate a decision tree for the training data provided in Table 1.

MODULE-III

Q.5 a) What are rule based classifiers? Explain the working of the same using the following example: (2+8)
Consider the following set of attributes and attribute value for the binary classification problem:
Air conditioner ={Working , Broken}
Engine = {Good, Bad}
Mileage={High,Meadium,Low}
Rust = {Yes, No}
The rule based classifier runs on this example and produces the following rule set:
Mileage =High $\longrightarrow$ Value = Low
Mileage = Low $\longrightarrow$ Value = High
Mileage =High $\longrightarrow$ Value = Low
Air Conditioner = Working, Engine = Good $\longrightarrow$ Value = High
Air Conditioner = Working, Engine = Bad $\longrightarrow$ Value = Low
Air Conditioner = Broken $\longrightarrow$ Value = Low
Is ordering needed for these set of rules and why?

(8+2)

b) Consider the set of transactions given in Table 2:
Table 2

TID	Items Bought
001	B,M,T,Y
002	B,M
003	A,T,S,P
004	A,B,C,D
005	A,B
006	T,Y,E,M

007	A,B,M
008	B,C,D,T,P
009	D,T,S
010	A,B,M

i. Assume that we wish to find the association rules with at least 30% support and 60% confidence. Find the frequent item sets and then the association rules.

ii. Which step of the Apriori algorithm is the most expensive? Explain the reasons for your answer.

Q.6 a) Write the k-nearest neighbor algorithm for classification. For the one dimensional data given in Table 3 below, give the k-nearest neighbor classifier for the points x=1, x=11 and x=100 using k=5. (4+6)

Table 3

X	Y
2	1
4	-1
6	1
8	-1
10	1
15	-1
20	1
25	-1
30	1
35	-1
40	1
45	-1
50	1
55	-1
60	1
65	-1
70	1
75	-1
80	1
85	-1
90	1
95	-1
100	1
200	-1

b) Consider an example with the set of transactions given in Table 4: (8)

Table 4

TID	Items Bought

001	B,M,T,Y
002	B,M
003	A,T,S,P
004	A,B,C,D
005	A,B
006	T,Y,E,M
007	A,B,M
008	B,C,D,T,P
009	D,T,S
010	A,B,M

Build an FP – tree for the transaction given in Table 4.

c) What is the significance of Naïve in the Naïve Bayes classifier? Justify. (2)

MODULE-IV

Q.7 a) State the principle of working and compare the following types of clustering from application point of view for each of the following: (6)

i. K-means clustering
ii. Hierarchical clustering
iii. Density – based clustering

b) Consider the following distance matrix in table 5, giving distance between given points: (8)

	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

Table 5

with matrix from Table 5, using agglomerative hierarchical clustering create a dendrogram.

c) List the characteristics of the following types of outliers and cite an example of each type: (6)

i. Global outlier
ii. Contextual outlier
iii. Collective outlier

Q.8 a) Consider the following items to cluster: (10)

{2,4,10,12,3,20,30,11,25}

i. Using the k-means clustering, cluster the given items considering k=2.
ii. Write the k-means algorithm.

b) Which of the data mining algorithms would be primarily used to generate recommendations in a recommender system? Justify. (4)

c) What is an outlier? Why is outlier mining important? Briefly describe distance based outlier detection. (6)