Total No. of Printed Pages:3

B.E (Computer) Semester-VII (Revised Course 2016-17)
EXAMINATION AUGUST-2020
Data Mining

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

Instructions : 1) Attempt **THREE FULL** Question at least **ONE** question from **EACH Part.**
2) Assume suitable data if necessary.

PART- A

Q.1
a) Draw a neat labeled diagram and explain the process of knowledge discovery in databases (KDD) **5**
b) Given a list of coordinates 'cords' and a point T. Find the item in the list that is closest to the point T. **4**
Cords=[(455,12), (188,90),(74,366), (10,10)]
T=(18,448)
c) Define binning. Perform data smoothing by bin means, bin boundaries and bin median for the following sorted data of Age attribute. **7**
4,8,9,15,21,21,24,25,26,28,29,34 with bins of depth =4
d) Find mode, median, mean and variance for the following data series: **4**
5,10,13,35,50,50,99

Q.2
a) What is attribute oriented induction (AOI)? Explain with a suitable example. **5**
b) Define Data Cube. Explain with neat diagram different types of OLAP Schemas. "The snowflake schema saves storage space compared to the star schema. "Justify. **9**
c) Compute Cosine Similarity and Extended Jaccard coefficient for following two document vectors:A=(3,6,0,3,6) and B= (1,2,0,1,2) **4**
d) State the general characteristics of Data Sets that have significant impact on data mining techniques. **2**

Q.3
a) Explain principal component analysis and state the significance of the same in data preprocessing **7**
b) Suppose the given data is : 300,440,700,990,1100 **8**
i) Use Z-score normalization to transform values 440,700 and 990
ii) Use Min- Max normalization to transform all given values into range of [0.0,1.0]
c) Discuss whether or not each of the following activity is Data Mining task **2**
i) Dividing the customers of a company according to their gender
ii) Monitoring seismic waves for earthquake activities
d) Using following set of stock prices: **3**
40,50,70,80,90,100,120,120,140,150,180,200
Find 20^{th} Percentile and 50^{th} Percentile.

PART- B

Q.4 a) Write K- Nearest Neighbor Classifier Algorithm **6**

b) Draw Decision Tree for following Data set. Explain Steps. **12**

TID	Home Owner	Marital Status	Annual Income	Class: Loan Defaulter
1.	Y	S	125	N
2.	N	M	100	N
3.	N	S	70	N
4.	Y	M	120	N
5.	N	D	95	Y
6.	N	M	60	N
7.	Y	D	220	N
8.	N	S	85	Y
9.	N	M	75	N
10.	N	S	90	Y

c) Define Rule based Classifier **2**

Q.5 a) Consider the data set shown below: **8**

Transaction	**Items Bought**
T1	Pasta, Lemon, Bread, Orange
T2	Pasta, Lemon
T3	Pasta, Orange, Cake
T4	Pasta, Lemon, Orange , Cake

Generate Association Rules using Apriori Algorithm.
Consider values of support= 60 % and Confidence =80%

b) Explain with neat figures k- means and different types of clusters **6**

c) Generate FP Tree for the following transaction dataset (Consider Support count =3) **6**

TID	ITEM SET
1	Fan, Axe, Cake, Doll, Gun, Mat, Pan
2	Axe, Bat , Cake, Fan, Lock, Mat ,Oven
3	Bat, Fan, Hat, Oven
4	Bat, Key, Cake , Pan
5	Axe, Fan, Cake , Lock , Pan, Mat, Net

Q.6 a) Differentiate between supervised and Unsupervised Learning **5**

b) Using Agglomerative Hierarchical Clustering Algorithm, Generate dendrogram for the following Proximity Matrix given below. **10**

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

c) Explain in brief Rule Ordering Schemes. **5**

PART- C

Q.7 a) Explain Challenges that motivated development of Data Mining. **5**

b) Suppose that a data warehouse consist of the 4- Dimensions i.e. date, spectator, location and game & the measures were count and charge (where charge is the fare that spectator pays when he/ she is watching a game on given date). Spectators may be students, adults or seniors with each category having its own charge rate. Draw a STAR Schema diagram for the data warehouse. **8**

c) Explain Feature Subset Selection process with a flowchart. **5**

d) What is the difference between Predictive and Descriptive Data Mining Tasks? **2**

Q.8 a) Apply KNN(K-Nearest Neighbor) Classification Algorithm on following data –set & Predict the class for testing data: X(I1=3,I2=7). (Consider k=3) **6**

Sr. No	I1	I2	Class
1	7	7	False
2	7	4	False
3	3	4	True
4	1	4	True

b) Define Outlier. Explain the Issues faced by Statistical Approach to Outlier detection. **4**

c) Consider the following data of eight objects for clustering **10**

Objects	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

Use the k-means Algorithm to cluster the above data into three clusters