BACHELOR OF TECHNOLOGY (C.B.C.S.) (2014 COURSE)
B.Tech.Sem - VIII COMPUTER :SUMMER- 2022
SUBJECT : DATA MINING & KNOWLEDGE DISCOVERY

Day : Monday | Time : 02:30 PM-05:30 PM
Date : 20-06-2022 | **S-13682-2022** | Max. Marks : 60

N.B.

1) All questions are **COMPULSORY**.
2) Figures to the **RIGHT** indicate **FULL** marks.
3) Assume suitable data, if necessary.

Q.1 What do you understand by data mining? Describe in detail the tasks of data mining with suitable example. **(10)**

OR

Q.1 Suppose that the data for analysis includes the attributes age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. **(10)**

i) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
ii) How might you determine outliers in the data?
iii) What other methods are there for data smoothing?

Q.2 Explain data cube in detail. Construct a data cube from the below given table. Is this a dense or sparse data cube? If it is sparse, identify the cells that are empty. **(10)**

Product ID	Location ID	Items sold
1	1	10
1	3	6
2	1	5
2	2	22

OR

Q.2 How are concept hierarchies useful in OLAP? Explain in detail with examples the typical OLAP operations on multi-dimensional data. **(10)**

Q.3 Most frequent pattern mining algorithms consider only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transactional data analysis. How can one mine frequent itemsets efficiently considering multiple occurrences of items? Propose modifications to the well-known algorithms, such as Apriori and FP-growth, to adapt to such a situation. **(10)**

OR

Q.3 What is FP-growth? A database D with nine transactions is given below. **(10)**

T ID	**List of items ID**
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Find the frequent itemsets without candidate generation.

P.T.O.

Q.4 Consider the training examples shown in below table for a binary classification problem. **(10)**

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

i) Compute the Gini Index for the overall collection of training example.
ii) Compute the Gini Index for the customer ID attribute.
iii) Compute the Gini Index for the Gender attribute.
iv) Compute the Gini Index for the Car type attribute using multi-way split.
v) Compute the Gini Index for the Shirt size attribute using multi-way split.
vi) Which attribute is better Gender, Car Type or shirt size?

OR

Q.4 What is decision tree induction? Write its algorithm. Explain in detail the design issues and characteristics of decision tree induction. **(10)**

Q.5 What is knowledge discovery? Explain the steps involved in the KDD process with neat labeled diagram. **(10)**

OR

Q.5 In knowledge discovery, explain how data mining system can be integrated with database / data warehouse system? **(10)**

Q.6 Explain the following clustering algorithm using examples. **(10)**
i) K-means
ii) K-mediod.

OR

Q.6 What is hierarchical clustering? Use the similarity matrix in below table to perform single and Complete link hierarchical clustering. Show your results by drawing a dendogram. The dendogram should clearly show the order in which the points are merged. **(10)**

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

*****