No. of Printed Pages : **3**

Sl. No.

C5-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1. Answer question 1 and any FOUR from questions 2 to 7.

2. Parts of the same question should be answered together and in the same sequence.

Time : 3 Hours **Total Marks : 100**

1. (a) Describe three challenges to data mining regarding data mining methodology and user interaction issues.

(b) Explain the role of Starnet Query Model in querying multidimensional databases with an example.

(c) Suppose a group of 12 sales price records has been sorted as follows :

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

(i) equal-frequency partitioning

(ii) equal-width partitioning

(d) Compare and contrast the incremental update operation performed in ROLAP, MOLAP and HOLAP.

(e) Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.

(f) Why is naive Bayesian classification called "naive" ? Briefly outline the major ideas of naive Bayesian classification.

(g) Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as PCA and SVD. **(7x4)**

2. (a) Outline the major research challenges of data mining in bioinformatics application domain.

(b) Use the two methods below to normalize the following group of data : 200, 300, 400, 600, 1000

(i) min-max normalization by setting min = 0 and max = 1

(ii) z-score normalization **(10+8)**

3. (a) 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 the fare that a 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.

(i) Draw a star schema diagram for the data warehouse.

(ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2004 ?

(b) What is Confusion Matrix for classifier ? Write a 4x4 confusion matrix for a classifier with 100% accuracy ? The instances in the four classes are A(20), B(35), C(40), D(10). Also calculate classifier precision in detail. **(10+8)**

4. (a) Use the similarity matrix in Table 1 to perform single and complete link hierarchical clustering. Show your results by drawing dendrogram. The dendrogram should clearly show the order in which the points are merged :

Table : 1

	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

(b) Consider the following data set shown in table where bank officials need to build decision tree to classify bank loan applications by assigning applications to one of the three risk classes that is A, B, C.

Owns Home	Married	Gender	Employed	Credit Rating	Risk Class
YES	YES	MALE	YES	A	B
NO	NO	FEMALE	YES	A	A
YES	YES	FEMALE	YES	B	C
YES	NO	MALE	NO	B	B
NO	YES	FEMALE	YES	B	C
NO	NO	FEMALE	YES	B	A
NO	NO	MALE	NO	B	B
YES	NO	FEMALE	YES	A	A
NO	YES	FEMALE	YES	A	C
YES	YES	FEMALE	YES	A	C

Draw the decision tree for the above table using ID3 algorithm. **(9+9)**

5. (a) The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, $\overline{\text{hotdogs}}$ refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and $\overline{\text{hamburgers}}$ refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{\text{hotdogs}}$	Σ_{row}
hamburgers	2,000	500	2,500
$\overline{\text{hamburgers}}$	1,000	1,500	2,500
Σ_{col}	3,000	2,000	5,000

(i) Suppose that the association rule "hot dogs ⇒ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong ?

(ii) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers ? If not, what kind of correlation relationship exists between the two ?

(b) What are outliers ? Explain their usage in the data analysis. **(12+6)**

6. (a) What is data mart ? How is it different from a Data Warehouse ? Define dependent, independent and hybrid data marts.

(b) What can business analysts gain from having a data warehouse ? **(10+8)**

7. Briefly compare the following concepts. You may use an example to explain your point(s).

(a) Snowflake schema, fact constellation, starnet query model.

(b) Data cleaning, data transformation, refresh.

(c) Enterprise warehouse, data mart, virtual warehouse. **(6+6+6)**

- o O o -