No. of Printed Pages : **2**

Sl. No.

C5-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

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

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

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

1. (a) What is data preprocessing ? Explain the steps of data preprocessing.

(b) Explain support and confidence in relation to association rule mining. Make use of appropriate example to explain.

(c) What is starnet query model for querying multidimensional databases ? Give an example to support your answer.

(d) Compare OLAP and OLTP operations.

(e) What are the applications of data mining ?

(f) Explain different types of data on which data mining can be applied.

(g) What is accuracy and error measure in relation to classification ? Explain with example. **(7x4)**

2. (a) What do you mean by data transformation ? Explain various techniques used for data transformation.

(b) What is descriptive data summarization ? Explain the techniques to measure dispersion of data.

(c) What are the major issues in data mining ? Explain them in brief. **(6+6+6)**

3. (a) What is multidimensional data model ? Draw and explain a 3-D cube to store sales data for All Electronics company, according to the dimensions time, item, and location.

(b) Draw and explain star schema for multidimensional databases.

(c) What are the various approaches for the Materialization of Different Kinds of Data cubes for on-line analytical processing of multidimensional data ? Explain in brief. **(6+6+6)**

4. (a) What is time series data ? Is it same as sequence data ? How is the trend analysis of a time series data done ? List the major components or movements for characterizing time-series data.

(b) Explain how a back propagation algorithm performs learning on the basis of a Multilayer Feed-Forward Neural Network.

(c) What is Outlier Analysis ? How Can Outlier Detection Improve Business Analysis ? Explain one technique for Outlier Analysis. **(6+6+6)**

5. (a) What are the problems associated with apriori algorithm and how they can be resolved with FP growth method ? Explain FP-growth algorithm for the following dataset of transactions.

TID	List of item_IDs
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

(b) How can you improve efficiency of apriori algorithm ?

(c) Compare eager classification with lazy classification techniques. **(9+6+3)**

6. (a) Define supervised and unsupervised learning techniques for data mining. Differentiate between these two learning techniques highlighting the key points.

(b) What is the difference between Partitioning Methods : k-Means and k-medoids ? According to you, which method is more flexible and robust ? Also, discuss the time complexity of both.

(c) Suppose you are required to measure the dissimilarity between categorical variables for cluster analysis. How would you perform the calculation ? Explain with reference to following dataset. **(8+4+6)**

Object Identifier	Test-1 (Categorical)
1	Code-A
2	Code-B
3	Code-C
4	Code-A

7. (a) Calculate the Correlation coefficient of given data as below :

x	12	15	18	21	27
y	2	4	6	8	12

(b) What are the assumptions of linear regression regarding residuals ? What is the coefficient of correlation and the coefficient of determination ?

(c) Draw the architecture of a typical data mining system and explain each of the components of it. **(6+6+6)**

- o O o -