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) Classification is supervised learning method and Clustering is unsupervised learning method. Differentiate between Classification and clustering.

(b) Handling a missing value is a part of Data Cleaning process. What are the different methods to handle missing value ?

(c) Differentiate between Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP).

(d) Real-world databases produces Time Series Databases (TSDB). What is time series database ?

(e) Why is tree pruning useful in decision tree induction ? What is a drawback of using a separate set of tuples to evaluate pruning ?

(f) Concept hierarchies' define a sequence of mappings in general concepts. Describe why concept hierarchies are useful in data mining.

(g) What is Multidimensional Association Rule ? Explain in brief. **(7x4)**

2. (a) What is classification ? Compare the advantages and disadvantages of eager classification versus lazy classification. Discuss K- Nearest-neighbor classifier.

(b) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.

(i) Draw a Star Schema for the above data warehouse.

(ii) Starting with the basic cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004 ? **(9+9)**

3. (a) Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. Explain Apriori Algorithm with example.

(b) Briefly describe the clustering methods with examples in each case. **(9+9)**

4. (a) What is noise ? Describe the possible reasons for noisy data.

(b) What are Bayesian classifiers ? Explain briefly Bayes' theorem.

(c) Briefly compare : Enterprise warehouse, Data mart and Virtual warehouse. **(6+6+6)**

5. (a) List out the Online Analytical Processing (OLAP) operations in multidimensional data model.

(b) Web usage mining mines weblog records to discover user access patterns of Web pages. Write a short note on web usage mining.

(c) A group of students are linked to each other in a social network via advisors, courses, research groups, and friendship relationships. Present a clustering method that may partition students into different groups according to their research interests. **(6+6+6)**

6. (a) Use the two methods given below to normalize the following group of data: 200; 300; 400; 600; 1000

1. min-max normalization by setting min = 0 and max = 1
2. z-score normalization

(b) What are the issues related to data integration of pre-processing step ? **(9+9)**

7. (a) Describe the following methods which evaluate the accuracy of a classifier.

1. Holdout Method
2. Random subsampling
3. K-fold cross validation

(b) Write short notes on any two of the followings :

1. Multilayer Feed-Forward Neural Network
2. Genetic algorithm
3. Sequential Pattern Mining **(9+9)**

- o O o -