ROLL NO. _____

Code: CT75 Subject: DATA WAREHOUSING AND DATA MINING

ALCCS

Time: 3 Hours **JUNE 2016** **Max. Marks: 100**

PLEASE WRITE YOUR ROLL NO. AT THE SPACE PROVIDED ON EACH PAGE IMMEDIATELY AFTER RECEIVING THE QUESTION PAPER.

NOTE:

- **Question 1 is compulsory and carries 28 marks. Answer any FOUR questions from the rest. Marks are indicated against each question.**
- **Parts of a question should be answered at the same place.**

Q.1 a. Comment on database processing vs. data mining processing. Support your answer with query examples in each one.

b. Draw the typical architecture of KDD Process.

c. What is a data warehouse? Explain the characteristics of data warehouse.

d. Explain Fuzzy set approach for classification with example.

e. Enlist the KDD Issues.

f. Discuss types of attributes with the help of example.

g. What is the ROCK clustering algorithm? What kind of similarity measure is used in ROCK? **(7×4)**

Q.2 a. State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the *update-driven approach*, rather than the *query-driven approach*? Describe situations where the query-driven approach is preferable over the update-driven approach. **(9)**

b. Briefly compare the following concepts. You may use an example to explain your point.
(i) Data cleaning
(ii) Data transformation
(iii) Refresh **(9)**

Q.3 a. Suppose that a data warehouse for *Big University* consists of the following four dimensions: *student, course, semester*, and *instructor*, and two measures *count* and *avg grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the average grade for the given combination.
Draw a *snowflake* schema diagram for the data warehouse. **(10)**

b. *Data quality* can be assessed in terms of accuracy and completeness. Propose other dimensions of data quality and relate the importance of each one. **(8)**

Q.4 a. Suppose we train a model to predict whether an email is Spam or Not Spam. After training the model, we apply it to a test set of 500 new email messages (also labeled) and the model produces the contingency matrix below. **(9)**
(i) Compute the precision of this model with respect to the Spam class.
(ii) Compute the recall of this model with respect to the Spam class.

		True Class	
		Spam	Not Spam
Predicted Class	Spam	70	10
	Not Spam	40	380

b. Discuss the back propagation learning algorithm for classification. **(9)**

Q.5 a. A patient record table contains the attributes *name, gender, fever, cough, test-1, test-2, test-3,* and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary. **(12)**

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Find the Dissimilarity between objects (patients) based only on the asymmetric variables.
(i) d(Jack, Mary)
(ii) d(Jack, Jim)
(iii) d(Mary, Jim)

b. List and describe the five *primitives* for specifying a data mining task. **(6)**

Q.6 a. Write down the pseudocode of Apriori algorithm. What do you mean by Apriori pruning principle? **(8)**

b. Define maximum marginal hyperplane (MMH) and support vectors in terms of SVM. Discuss SVM—when data is linearly separable. **(10)**

Q.7 Discuss the following. **(3×6)**
(i) Tree pruning in decision tree induction
(ii) What do you mean by supervised and unsupervised data mining techniques? How both kinds of techniques are different from each other?
(iii) OLTP vs Data Warehouse