Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

MCA - SEMESTER–IV • EXAMINATION – WINTER • 2014

Subject Code: 2640005 Date: 06-12-2014

Subject Name: Data Warehousing and Data Mining (DWDM)

Time: 10:30 am - 01:00 pm Total Marks: 70

Instructions:

1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**

Q.1 **(a)** Brief the following terms: **07**

1. What is Jack Knife?
2. What is Fact Table?
3. What is Linear Regression?
4. What is decision Tree Pruning?
5. What do you mean by Data Mart?
6. What do you mean by Snowflake Schema?
7. What is Clustering?

(b) State whether the statement is true or false and justify your answer **07**

1. "Data mining can be viewed as a result of the natural evolution of information technology."
2. Backpropagation is a neural network learning algorithm.
3. The confusion matrix is a useful tool for analyzing how well a classifier can recognize tuples of different classes.
4. Cross-validation and bootstrap are two techniques that increase the accuracy of decision tree.
5. A temporal database typically stores relational data that include time-related attributes.
6. A time-series database stores sequences of ordered events, with or without a concrete notion of time.
7. A heterogeneous database consists of a set of interconnected, autonomous component databases.

Q.2 **(a)** What is KDD process & data pre-processing? How is it different from data mining process? **07**

(b) In the process of data cleaning, how can we fill up the missing values? Write down its methods. **07**

OR

(b) What are the interestingness measures of association rule mining? Explain three interestingness measures giving appropriate examples **07**

Q.3 **(a)** Explain with figure: 3-Tire Data Warehouse Architecture. **07**

(b) List and explain various types of OLAP Servers **07**

OR

Q.3 **(a)** What is classification? Also explain Supervised and Unsupervised learning in detail **07**

(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 patient for a visit. **07**

1. Enumerate three classes of schemas that are popularly used for modeling data warehouses.
2. Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).
3. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be per- formed in order to list the total fee collected by each doctor in 2004?
4. To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge)

Q.4 **(a)** Discuss two Ensemble methods for increasing the accuracy of classifier **07**

(b) Explain parametric methods and non-parametric methods of reduction? **07**

OR

Q.4 **(a)** List and explain describe major issues in data mining. **07**

(b) List and explain several OLAP operations with example. **07**

Q.5 **(a)** Discuss the following as attribute selection measure with example **07**

1. Information gain
2. Gain ratio

(b) What is Apriori property? How the Apriori property is used in finding frequent itemset. **07**

OR

Q.5 **(a)** Discuss the application of Data Mining in Retail Industry. **07**

(b) Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points $a=(x1, y1)$ and $b=(x2, y2)$ is defined as: **07**

$\rho(a, b) = |x2 – x1| + |y2 – y1|$.

Use k-means algorithm to find the three cluster centers after the second iteration

****