Seat No.: ________ Enrolment No.___________

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER–VII • EXAMINATION – SUMMER • 2015

Subject Code: 171601 **Date: 01/05/2015**

Subject Name: Data warehousing and Data mining

Time:02.30pm-05.00pm **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)** Explain different OLAP operation with example. **07**

(b) i) What are the major challenges of mining a huge amount of data in comparison with mining a small amount of data? **04**

ii) Why strong association rule is not always interesting? Explain with example. **03**

Q.2 **(a)** 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. **07**

1) Draw a star schema diagram for the data warehouse.
2) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

(b) Define sampling. Explain different type of sampling techniques with example. **07**

OR

(b) What is noise? Explain the different techniques to remove the noise from data. **07**

Q.3 **(a)** How to compute the dissimilarity between objects described by the following types of variables: **07**

1) Interval-scaled variables
2) Asymmetric binary variables
3) Categorical variables

(b) How multilevel association rules can be mined efficiently using concept hierarchy? **07**

OR

Q.3 **(a)** Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: **07**

$A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$:

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show

1) The three cluster centers after the first round execution
2) The final three clusters

(b) Explain linear regression? What are the reasons for not using the linear regression model to estimate the output data? **07**

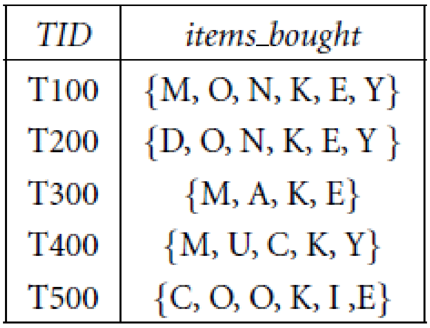
Q.4 **(a)** What is decision tree induction? Write Basic algorithm for inducing a decision tree from training tuples. **07**

(b) i) List strengths and weakness of neural network as classifier. **04**

ii) How can distance be computed for attributes that having missing valves in K-Nearest Neighbor classifier? **03**

OR

Q.4 **(a)** A database has 5 transactions. Let min_sup = 60% and min_conf = 80%. **07**



TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

1) Find all frequent itemsets using Apriori algorithm
2) List all the association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., "A", "B", etc.):
$\forall x \in$ transaction; buys(X, item1) $\wedge$ buys(X, item2) $\rightarrow$ buys(X, item3) [s, c]

(b) What are the methods to evaluate accuracy of classifier/predictor? **07**

Q.5 **(a)** Write a short note on web usage mining. **07**

(b) Discuss basic principle of Attribute Oriented Indication **07**

OR

Q.5 **(a)** a) What is time series database? How to characterize the time series data using trend analysis? **07**

(b) i) What are measures for assessing quality of text retrieval mining system? **04**

ii) What are the terminating conditions to stop training process of neural network classifier? **03**

****