Seat No.: _____ Enrolment No._____

# GUJARAT TECHNOLOGICAL UNIVERSITY

**M.C.A -IV$^{th}$ SEMESTER–EXAMINATION – MAY- 2012**

**Subject code: 640005 Date: 19/05/2012**

**Subject Name: Data Warehousing & 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)** | Define Data Warehouse. Briefly describe the following terms: (i) Subject-oriented; (ii) Integrated; (iii) Time-variant; (iv) Non-volatile; (v) Fact; (vi) Dimension | **07** |
| | **(b)** | What is meant by Concept Hierarchy and what is its purpose? Is it applicable for Fact or for Dimension? | **03** |
| | **(c)** | Data Mining is classified into two categories: (i) Descriptive Data Mining, and (ii) Predictive Data Mining. Write briefly the basic purpose of these two categories of Data Mining. To which of the two categories the following data mining taskd fall into:<br>• Concept Description<br>• Detection of an outlier | **04** |
| | | | |
| **Q.2** | **(a)** | What is meant by Supervised Learning and Unsupervised Learning? A data mining task may use either supervised learning or unsupervised learning or either of the two or none of the two types. Indicate against each data mining task, the learning type, i.e. supervised or unsupervised or either or none:<br>• Classification<br>• Clustering<br>• Characterization<br>• Discrimination<br>• Association Rule Mining | **07** |
| | **(b)** | Suppose that a data warehouse of a hospital consists of three dimensions, namely time, doctor, and patient with the concept hierarchy as follows:<br>Time: day, month, quarter, year<br>Doctor: doctor, specialization (e.g. ophthalmologist, pediatrician, etc.)<br>Patient: patient, category (e.g. outdoor, indoor)<br>There are two measures, namely count and charge, where charge is the fee that a doctor charges a patient for a visit.<br>(i) Draw a star schema for the above data warehouse.<br>(ii) 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 the year 2010? | **07** |
| | | **OR** | |
| | **(b)** | Using the data warehouse given in the above example (Q. 2 (b) main part), let us assume that the hospital's interest is to analyze the volume of patients and the revenue generated (through fee) for a group of doctors under each specialization and for each doctor on monthly and yearly basis. Which of the pre-computed cubes (list out all the relevant data cubes) are required for this task? Justify your | **07** |

| | | | |
|---|---|---|---|
| | | answer. In this case, whether metadata has any role to play? If yes, briefly describe the role of metadata. If only base cube is available, what OLAP operations will be required to do the task? | |
| | | | |
| **Q.3** | **(a)** | The cube definition statement has the following syntax<br>define cube <cube_name> [<dimension_list>] : <measure_list><br>The dimension definition statement has the following syntax<br>define dimension <dimension_name> as (<attribute_or_subdimension_list>)<br><br>Using the above syntax, define the fact constellation schema in DMQL for the two 'facts' and the corresponding 'dimensions' given below.<br>First 'fact': sales with four dimensions, namely time, item, branch, and location with the two measures as (i) value_sold (i.e. sales value), (ii) quantity_sold<br>Second 'fact': shipping with five dimensions, namely time, item, shipper, from_location, to_location with the two measures as (i) amount_shipped, and (ii) quantity_shipped.<br>Concept hierarchy of dimensions is given below:<br>time: day, month, quarter, year; item: item_name, brand, type, supplier_type<br>branch: branch_name, branch_type; location: city, state, country<br>shipper: shipper_name, shipper_location, shipper_type<br>Note-1: any location whether of shipper or from_location or to_location will have same hierarchy as that of location.<br>Note-2: if there is a difficulty in defining the fact constellation, then define the schema for the two facts separately in DMQL. | **07** |
| | **(b)** | What is Data Mart? Differentiate between Data Mart and Data Warehouse. Describe 3-tier architecture of Data Warehouse. | **07** |
| | | **OR** | |
| **Q.3** | **(a)** | Write down the salient points differentiating OLAP against OLTP. What are the basic characteristics of ROLAP, MOLAP, and HOLAP? | **07** |
| | **(b)** | Data pre-processing includes (i) Data Cleaning, (ii) Data Integration, (iii) Data Transformation, (iv) Data Reduction. Write briefly the basic tasks done under each type of pre-processing. | **07** |
| | | | |
| **Q.4** | **(a)** | A database has the following transactions. Let min_sup = 60% and min_conf = 80%<br>TID items_bought (in the form of brand-item_category)<br>T01 {Sunset-Milk, Dairyland-Cheese, Best-Bread}<br>T02 {Goldfarm-Apple, Dairyland-Milk, Best-Cheese, Wonder-Bread, Tasty-Pie}<br>T03 {Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}<br>T04 {Sunset-Milk, Dairyland-Cheese, Wonder-Bread}<br>List the frequent k-itemset for the largest k at the granularity of item_category (e.g. $item_i$ could be "Milk") for the following rule template:<br>For all X $\epsilon$ transaction, buys(X, $item_1$) ^ buys(X, $item_2$) => buys(X, $item_3$) [s, c]<br>Also list all of the strong association rules (with their support s and confidence c) containing the frequent k-itemset for the largest k. | **07** |
| | **(b)** | Suppose that the data mining task is to cluster the following 8 points (with (x, y) representing location) into three clusters.<br>$A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$.<br>The distance function is Manhattan 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 | **07** |

| | | | |
|---|---|---|---|
| | | add two points only, i.e. $A_2$ and $A_3$, in appropriate clusters and compute the new centers of the clusters. When a new point is added in a cluster, the new center is computed as follows: (New Center) = Mean of (Old Center) and the (New Point Added). In case of a conflict, choose the cluster for which $\|x_1 - C_x\|$ and $\|y_1 - C_y\|$ are closer to each other, where $(C_x, C_y)$ are the coordinates of the new center. | |
| | | **OR** | |
| **Q.4** | **(a)** | A database has four transactions. Let min_sup = 60% and min_conf = 80%.<br>TID date items_bought<br>T100 18-Jan-2011 {A, B, D, K}<br>T200 18-Jan-2011 {A, B, C, D, E}<br>T300 19-Jan-2011 {A, B, C, E}<br>T400 22-Jan-2011 {A, B, C, D}<br>Find all frequent itemsets using Apriori algorithm. List all of the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g. "A", "B", etc.):<br>For all X ϵ transaction, buys(X, $item_1$) ^ buys(X, $item_2$) => buys(X, $item_3$) [s, c] | **07** |
| | **(b)** | Consider the 8 points given above in Q. 4 (b) main part. Use the distance function as Manhattan distance, and initially take $A_1$, $B_1$, and $C_1$ as the center of each cluster, respectively. If the similarity threshold is given as distance <= 4, which points will get marked as outliers?<br><br>If the initial center of each cluster is assigned as $A_1$, $A_3$, and $C_1$ then which points will get marked as outliers? | **07** |
| | | | |
| **Q.5** | **(a)** | The following table consists of training data from an employee database. Instead of repeating a tuple having the same values for "department", "status", "age", and "salary", only one instance of the tuple is included with the repeat count mentioned under the column (i.e. attribute) "count".<br>department status age salary count<br>sales senior 31…35 E 30<br>sales junior 26…30 A 40<br>sales junior 31…35 B 40<br>systems senior 31…35 H 05<br>systems junior 26…30 E 03<br>systems junior 21…25 E 20<br>systems senior 41…45 H 03<br>marketing senior 36…40 E 10<br>marketing junior 31…35 D 04<br>marketing senior 46…50 E 04<br>marketing junior 26…30 C 06<br>(i) Draw a decision tree taking "department" as the root node, and taking "status" at the next level.<br>(ii) Derive the decision rules with the associated probabilities. For example, if all the tuples at the leaf node belong to only one class, the probability will be 1 (i.e. 100%). However, if, say, 80% of the tuples belong to a particular class, the probability will be 0.8 (i.e. 80%). | **07** |
| | **(b)** | Bayesian theorem is stated as follows:<br>P(H \| X) = P(X \| H) P(H) / P(X)<br>Where X is a data sample whose class label is unknown. H is the hypothesis such as that the data sample X belongs to a specified class C. | **03** |

| | | | |
|---|---|---|---|
| | | The above equation can be re-written as:<br>$P(X \mid H) = P(H \mid X)\, P(X) / P(H)$<br>Can we use any one of the above-stated two equations in the Bayesian theorem? Justify your answer. | |
| | **(c)** | Describe Hold-out method and k-fold cross-validation method for assessing classifier accuracy. | **04** |
| | | **OR** | |
| **Q.5** | **(a)** | Training data tuples from a customer database are given below:<br>RID age income student credit_rating Class:buys_laptop<br>1 <=30 high no excellent yes<br>2 <=30 medium no fair yes<br>3 <=30 low yes fair yes<br>4 <=30 low no fair no<br>5 31…40 high no fair no<br>6 31…40 medium no excellent yes<br>7 31…40 low yes fair yes<br>8 31…40 low yes excellent no<br>9 >40 high no excellent yes<br>10 >40 medium no fair no<br>11 >40 low no fair no<br>12 >40 medium no excellent no<br>13 >40 low no fair no<br>(i) Draw a decision tree taking "credit_rating" as the root node, and taking "age" at the next level.<br>(ii) Derive the decision rules with the associated probabilities. For example, if all the tuples at the leaf node belong to only one class, the probability will be 1 (i.e. 100%). However, if, say, 80% of the tuples belong to a particular class, the probability will be 0.8 (i.e. 80%). | **07** |
| | **(b)** | Let the number of (independent) attributes in a database be 5 and let there be 4 class labels. If we have to use multi-layer, feed-forward artificial neural network (ANN) with back-propagation, how many nodes (neurons) will be in the input layer, in the output layer, and in one hidden layer. In case any assumption and / or any rule-of-thumb is used, it should be stated clearly. | **03** |
| | **(c)** | Describe Bagging (or Bootstrap aggregation) and Boosting for increasing classifier accuracy. | **04** |

**********