Seat No.: _____ **Enrolment No._____**

GUJARAT TECHNOLOGICAL UNIVERSITY
MCA SEM-IV Examination- Dec.-2011

Subject code: 640005 **Date: 22/12/2011**
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)** 1) What is data warehouse? Briefly explain the key words used in the definition. **04**

2) Draw a labeled diagram of 3-tier Data warehouse architecture. **03**

(b) What do you mean by OLTP and OLAP? Distinguish between them on the following features. **07**

1) Users and system orientation
2) Data content
3) Database design
4) View
5) Access pattern

Q.2 **(a)** What is data mining? Describe the steps involved in data mining process with diagram. **07**

(b) List the various Data Pre-processing methods. Discuss various Data Cleaning techniques for Missing value and Noisy data. **07**

OR

(b) 1) What do you mean by data reduction? Briefly describe two data reduction strategies. **04**

2) Data: 50, 125, 75, 200, 150, 125, 100. **03**

Normalize the above data using following normalization methods

1) Min max normalization with min = 0 and max = 1.
2) Decimal scaling normalization

Q.3 **(a)** What is Apriori property? How the Apriori property is used in finding frequent itemset. Explain Join and Prune step. **07**

(b) Compare and contrast between star and snowflake schema. Create a star and snowflake schema Data warehouse for sales with three dimensions item, time, **07**

location and two measures sales_unit, sales_amount.

OR

Q.3 **(a)** What are the interestingness measures of association rule mining? Explain three interestingness measures giving appropriate examples. **07**

(b) Given a transaction database: **07**

{bread milk butter beer}

{bread butter water jam beer}

{beer diapers bread butter jam}

{butter milk juice}

{diapers beer juice water}

1) For minimal support 0.6 find all frequent itemsets
2) For minimal confidence 0.7 find association rules of the form item1→{item2, item3}

Q.4 **(a)** Consider the following database (D) and calculate **07**

1) Info(D)
2) $Info_{age}(D)$
3) Gain(age)

id	age	income	student	credit_rating	CLASS: buys_computer
1	youth	high	no	fair	No
2	youth	high	no	excellent	No
3	middle_aged	high	no	fair	Yes
4	senior	medium	no	fair	Yes
5	senior	low	yes	fair	Yes
6	senior	low	yes	excellent	No
7	middle_aged	low	yes	excellent	Yes
8	youth	medium	no	fair	No
9	youth	low	yes	fair	Yes
10	senior	medium	yes	fair	Yes
11	youth	medium	yes	excellent	Yes
12	middle_aged	medium	no	excellent	Yes
13	middle_aged	high	yes	fair	Yes
14	senior	medium	no	excellent	No

(b) Write down the decision tree induction algorithm for classification. Explain the working of decision tree induction algorithm. **07**

OR

Q.4 **(a)** What is Bayes theorem? Explain the working of Naïve Bayesian Classifier. **07**

(b) Discuss two Ensemble methods for increasing the accuracy of classifier. **07**

Q.5 **(a)** Cluster the following eight points (with (x, y) representing locations) into three clusters: **07**

A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9).

Assume that initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

Use Manhattan distance for the distance function between two points. Use k-means algorithm to find the three cluster centers after the first iteration.

(b) 1) What is confusion matrix? Define classifier accuracy measures. **04**

2) What do you mean by cross validation? **03**

OR

Q.5 **(a)** Write down the k-mediod algorithm for clustering and explain its working. How it is better than k-mean clustering algorithm? **07**

(b) 1) Differentiate between clustering and classification. **04**

2) Discuss the application of Data Mining in Retail Industry. **03**

****