Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

M. E. Sem. – IInd - Examination – June/July- 2011

Subject code: 1720108

Subject Name: Data Mining & Data Warehousing

Date:29/06/2011 **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. Differentiate between data, information & knowledge. **03**

2. What are major research challenges of data mining in spatiotemporal data analysis **04**

(b) Define each of the following data mining functionalities: characterization, discrimination, association, classification, prediction, clustering, evolution analysis **07**

Q.2 **(a)** In real world data, tuples with missing values for some attributes are common occurrence. Describe various methods for handling this problem. **07**

(b) 1. Suppose that the data for analysis includes the attribute age. The age values for data tuples are 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. **04**

(i) Use min-max normalization to transform value 35 for age onto the range [0.0, 1.0].

(ii) Use z-score normalization to transform value 35 for age, where standard deviation of age is 11.49 years.

2. Briefly explain methods of concept hierarchy generation for categorical data. **03**

OR

(b) Explain sampling method for data reduction. **07**

Q.3 **(a)** 1. Define: data warehouse & data mart. **02**

2. What do you mean by concept hierarchy? Draw concept hierarchy for location(country, state, city, street) and time(year, quarter, month, week, day). **05**

(b) Suppose data warehouse consists of three dimensions time, doctor & patient and two measures count & charge. **07**

(i) Enumerate three classes of schemas that are popularly used for modeling data warehouse.

(ii) Draw snowflake schema for above given data warehouse.

(iii) 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?

OR

Q.3 **(a)** **07**

TID	List of Item Ids
T1	I1, I2, I5, I6
T2	I2, I5
T3	I1, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3, I4
T7	I1, I3
T8	I1, I2, I5
T9	I1, I2, I3

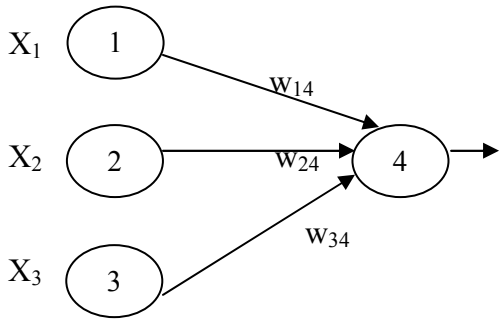
Convert given data in vertical data format & find frequent item sets having min_sup = 2 and min_conf = 70%

(b) What do you meant by multilevel association rule? Give different methods to find frequent item sets at different levels by suitable example. **07**

Q.4 **(a)** What is difference between association & correlation? For given contingency table find type of correlation between items using any three methods. **07**

	Game	$\overline{\text{Game}}$	Total
Video	4,000	3,500	7,500
$\overline{\text{Video}}$	1,000	1,500	2,500
Total	5,000	5,000	10,000

(b) **07**



Initial inputs X_1=1, X_2=0, X_3=1 Weights w_{14} = -0.3, w_{24} = 0.4, w_{34} = -0.5
Bias θ_4 = 0.2
Using the feed forward neural network of above, show the weight values after one iteration of backprapogation algorithm.

OR

Q.4 **(a)** Define: supervised learning, training set, testing set, accuracy of classifier, sensitivity, specificity, regression. **07**

Q.4 **(b)** Training data set is given below. **07**

RID	Income	Credit_rating	Buys_computer
1	Youth	Fair	No
2	Youth	Excellent	No
3	Middle_aged	Fair	Yes
4	Senior	Fair	Yes
5	Senior	Fair	Yes
6	Senior	Excellent	No
7	Middle_aged	Excellent	Yes
8	Youth	Fair	No
9	Youth	Fair	Yes
10	Senior	Fair	Yes
11	Youth	Excellent	Yes
12	Middle_aged	Excellent	Yes
13	Middle_aged	Fair	Yes
14	Senior	Excellent	No

Let Buys_computer is class-label attribute. Find lowest impurity attribute which can be placed at root level of decision tree using Gain Ratio attribute selection measure.

Q.5	**(a)**	Explain various methods of evaluating accuracy of classifier.	**07**
	(b)	Suppose that data mining task is to cluster the following eight points into three clusters. A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9). The distance function is Euclidian distance. Suppose initially we assign A1, B1 & C1 as center of each cluster, respectively. Use k-mean algorithm to show only three cluster centers after the first round execution.	**07**

OR

Q.5	**(a)**	Explain BIRCH algorithm of clustering.	**07**
	(b)	What do you mean by text mining? Explain various issues involved in it.	**07**

****