Reg. No.										

Manipal Institute of Technology, Manipal

(A Constituent Institute of Manipal University)

VII SEMESTER B.TECH (COMPUTER SCIENCE AND ENGINEERING)

END SEMESTER EXAMINATIONS, NOV/DEC 2015

SUBJECT: DATA WAREHOUSING AND DATA MINING [CSE 433]

REVISED CREDIT SYSTEM

DATE: 26 -12-2015

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ANY FIVE FULL** questions.
- Missing data, if any, may be suitably assumed.

1A. Discuss the concept of three tier architecture of the data warehouse with a neat diagram. **3M**

1B.Suppose that a data warehouse for an engineering college consists of four dimensions **(student, branch, semesters, subjects)** and a measure called **marks** which stores the subject wise marks in every semester.

(a) Draw a lattice of cuboids for the above given four dimensions and also find the total number of cuboids in the lattice so constructed.

(b)Starting with the base cuboid **(student, branch, semester, subjects)**, what specific OLAP operations should one perform in order to list the average marks of each computer science student in all semesters? **4M**

1C. Compare the systems OLAP and OLTP based on the user and system orientation, view, database design and contents, user access pattern. **3M**

2A Define data mart and write down the schemas suitable for constructing data warehouse and data mart? **1M**

2B. Consider a database given in Table Q.2B. Find all strong association rules for largest frequent itemset with respect to the minimum sup = 20% and minimum confidence = 60%. **4M**

Table Q.2B

TID	Items	TID	Items
1	3, 4	**6**	1, 3
2	2, 3	**7**	2
3	1, 2 , 3, 5	**8**	1, 3
4	2, 5	**9**	1, 2, 3
5	1, 2	**10**	1, 3

2C Use pincer search algorithm to discover all maximal frequent itemsets for the Customer Basket Database given in Table Q. 2C with respect to 20% of minimum support. **5M**

Table Q.2C

TID	Products
1	Burger, Coke, Juice
2	Juice, Potato-Chips
3	Coke, Burger
4	Juice, Ground-nuts
5	Coke, Ground-nuts

3A. Write a partition algorithm for discovering frequent itemsets. Explain how this algorithm is better than Apriori algorithm? **3M**

3B. Consider a Table Q.3B of tuples corresponding to the weather forecast for cricket match. The class label attribute 'Play' has two values 'YES' and 'NO'. Find the best splitting point for the attribute Outlook to construct binary decision tree. **5M**

Table Q. 3B

Outlook	Temperature	Humidity	Windy	Play ?
sunny	hot	high	false	NO
sunny	hot	high	true	NO
overcast	hot	high	false	YES
rain	mild	high	false	YES
rain	cool	normal	false	YES
rain	cool	normal	true	NO
overcast	cool	normal	true	YES
sunny	mild	high	false	NO
sunny	cool	normal	false	YES
rain	mild	normal	false	YES
sunny	mild	normal	true	YES
overcast	mild	high	true	YES
overcast	hot	normal	false	YES
rain	mild	high	true	NO

3C. Use Bayesian Classification for the data set given in the Table Q. 3B to determine the classes of the following tuples?

(a)X= <rain, hot, high, false >

(b)Y= <sunny, hot, high, false>

2M

4A. Justify Back propagation Algorithm. Write an algorithm to get trained neural network for the inputs**: *D*,** a data set consisting of the training tuples and their associated target values; ***l***, the learning rate; **network,** a multilayer feed-forward network. **3M**

4B. Write a note on support vector machine. **2M**

4C.What are the aspects with which clustering methods can be compared and also list different types of clustering techniques **5M**

5A. List the cases that arise when algorithm K-medoids is used for clustering data objects. **1M**

5B. Consider the following set of data objects

{(2,6),(3,4), (3,8),(4,7),(6,2),(6,4),(7,3),(7,4),(8,5),(7,6)}

Use K-medoids algorithm and Manhattan distance measure to discover two clusters by considering (3,4) and (7,4) as cluster medoids. Check whether the replacement of

(i) (7,4) by (8,5)

(ii) (3,4) by (2,6)

on the **initial clusters** formed is a good replacement or not. **4M**

5C. Discuss the method for removal of redundant categorical attributes in the datasets. Suppose that a researcher studies aggression content in the dreams of men and women. Each participant reports his or her most recent dream then each dream is judged by a panel of experts to have low, medium, or high aggression content. The observed frequencies are shown in the Table Q. 5C. Is there a relationship between gender and the aggression content off dreams?. Test with significance level = 0.01 and chi-squared value = 9.210 with respect to the degrees of freedom = 2 in the distribution table. **5M**

Table Q. 5C

Aggression content/ Gender		**Gender**	
		Female	**Male**
Aggression content	**Low**	18	4
	Medium	4	17
	High	2	15

6A. (i) Discuss different types of tree pruning in decision tree induction?

(ii)When do we use the attribute section measure Gain ratio? **3M**

6B. Use z-score normalization method to normalize the following group of data

1200, 1300, 1400, 1600, 2000 **2M**

6C. Explain the concept of web content mining and web usage mining. **5M**

****