Reg No.									

MANIPAL INSTITUTE OF TECHNOLOGY
(Constituent Institute of Manipal University)
MANIPAL-576104

SEVENTH SEMESTER B.TECH (CSE) DEGREE END SEMESTER EXAMINATION
NOV./DEC. 2012
DATA WAREHOUSE AND DATA MINING (CSE 405.2)

DATE: 05-12-2012

TIME: 3 HOURS MAX.MARKS: 50

Instructions to Candidates

- Answer **any five** full questions.

1A What are the advantages of In advance approach over Lazy approach? Where exactly these methods are suitable?

1B Distinguish between Snow Flake and Star Schemas with the help of one example

1C 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. (2+3+(2+3))

2A Why do you say that support and confidence frame work is not sufficient for filtering strong association rule? How do you overcome this issue.? Explain with the help of one example.

2B. Consider the following sample database. Calculate the support and confidence for the association rules given below.

(i) Milk → {Bread, Butter} (ii) {Bread, Butter}→ {Egg, Milk}

(iii) Milk→{Butter, Egg} (iv) {Egg, Bread} →Butter

TID	Milk	Bread	Butter	Egg
1	1	1	0	0
2	0	0	1	0
3	1	1	0	1
4	1	1	1	1
5	0	1	1	1

2C. Write a pseudo code for Maximal Frequent Candidate generation procedure. During the execution of Pincer-search Algorithm, it is found that when **k=1, L_1 ={ {A} , {B}, {C}, {D, {E} }, S_1 ={ }, MFCS={A,B,C,D,E}** when **k=2, L_2= { {A,B}, {A,C}, {A,E}, {B,C}, {B,D}, {B,E} }** and **S_2 = { {A,D}, {C,D},{C,E}, {D,E} }.** What is C_3 when **k=3**?. Indicate all steps. (3+2 +5)

3A. List and discuss the three levels of Data modeling in connection with data warehouse.

3B What is the drawback of information gain and how do you overcome this?

3C A database has five transactions. Let min sup = 60% and min conf = 80%.

TID	Item_bought
1	**M, O, N, K, E, Y**
2	**D, O, N, K, E, Y**
3	**M, A, K, E**
4	**D, U, C, K, Y**
5	**C, M, O, K, I, E**

(i)Find all frequent itemsets by constructing FP- Tree.

(ii) List all the strong association rules form

{Item1, Item2} ➔ {Item3}, where $\mathbf{Item_i}$ denotes variables representing items (e.g., "A", "E", etc.) from the frequent itemsets discovered above.

(2+2+(4+2))

4A. Explain any three OLAP operations with the help of one example.

4B. Discuss the concept of granularity and levels of granularity in data warehouse modeling and explain why is so important?

4C. With a neat diagram, explain all the steps required to discover knowledge from the large volumes of data. (3+4+3)

5A How do you overcome the problem of over fitting the data in decision tree? Explain.

5B. Consider a following training database containing the tuples corresponding to the weather forecast for cricket match. The class label attribute 'Play' has two values 'YES' and 'NO' .

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

(i) Find the best splitting point for the attribute **Outlook** to construct binary decision tree.

(ii) Use Bayesian Classification to find the **Class** to which the tuples (a)X=<rain, hot, high, false >

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

(3+(5+2))

6A. Justify the name given to 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.

6B Let the learning rate be 0.9. The initial weight and bias values of the network are given in following table, along with the first training tuple ***X*** = (1, 0, 1), whose class label is 1. Use an appropriate multilayer feed-forward neural network to find the net input, output and error for 4, 5, 6 and also update weight and bias . (Carry out one iteration)

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	–0.3	0.4	0.1	–0.5	0.2	–0.3	–0.2	–0.4	0.2	0.1

Initial input, weight, and bias values.

(5+5)

****