Reg. No.

Manipal Institute of Technology

(A Constituent Institute of Manipal University)

III SEMESTER M. C. A.

END SEMESTER EXAMINATION – NOV/DEC 2015

SUBJECT: DATA WAREHOUSING AND DATA MINING [MCA 5102]

27-11-2015

Time : 3 hours

Max. Marks : 50

Instructions to Candidates

1. Answer ANY FIVE FULL questions.
2. Missing data may be suitably assumed.

1A Define data mining. Explain the sequence of steps in the knowledge discovery process with a neat diagram.

1B Are all data mining patterns interesting? Differentiate between subjective and objective interestingness measures.

1C What is multivariate data analysis? How can it be visualized?

(5 + 3 + 2)

2A Suppose that the data for analysis includes the attribute age. The age values of the 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.

i. Compute the 5 number summary.
ii. Clean the data by finding and eliminating outliers if any.
iii. Draw a box plot for the cleaned data.
iv. Use smoothing by bin means to smooth data using bins of depth size 3.

2B What are concept hierarchies? Generate a concept hierarchy for the attribute 'Date of Birth'.

2C How are data redundancies detection during data integration phase?

(5 + 3 + 2)

3A What is a data warehouse? Explain the terms OLAP & OLTP and compare their features.

3B What are the functions supported by ETL tools in data warehousing?

3C Differentiate between the Drill Up and Drill Down OLAP operations using examples.

(5 + 3 + 2)

4A Consider the transactional data set given below. Let minimum support be 60 %. Find all the frequent item sets only using either the Apriori or the FP Tree method.

TransactionID	Items purchased
100	Bread, Cheese, Eggs, Juice
200	Bread, Cheese, Juice
300	Bread, Milk, Yogurt
400	Bread, Juice, Milk
500	Cheese, Juice, Milk

4B What is the need for performing Correlation Analysis on association rules? Illustrate with an example.

4C How does the partition algorithm improve on the efficiency of the Apriori algorithm?

(5 + 3 + 2)

5A The following table shows the relationship between the amount of fertilizer used and the height of a plant.

i. Calculate a simple linear regression equation using Fertilizer as the descriptor and Height as the response.

ii. Predict the height when fertilizer is 12.3.

Fertilizer	10	5	12	18	14	7	15	13	6	8	9	11	16	20	17
Height	0.7	0.4	0.8	1.4	1.1	0.6	1.3	1.1	0.6	0.7	0.7	0.9	1.3	1.5	1.3

5B Differentiate between the following, with suitable examples.

i. Supervised vs. unsupervised learning

ii. Eager vs. lazy learners

iii. Classification vs. Prediction

5C How is the Information Gain attribute selection measure computed for continuous valued attributes while constructing a decision tree?

(5 + 3 + 2)

6A Consider the following distance matrix and perform agglomerative clustering on the 6 data points. Visualize using a dendrogram.

	A	B	C	D	E	F
A	0	1.0	5.0	9.0	10.0	2.0
B	1.0	0	3.5	8.0	7.0	5.5
C	5.0	3.5	0	3.0	4.0	6.5
D	9.0	8.0	3.0	0	0.5	4.5
E	10.0	7.0	4.0	0.5	0	2.5
F	2.0	5.5	6.5	4.5	2.5	0

6B Given two data points X= (22, 3, 40, 12) and Y= (24, 0 , 46, 8) .Represent them as a distance matrix using

i. Euclidean distance between the data points

ii. Manhattan distance between the data points.

iii. Minkowski distance between the data points using q = 3.

6C Differentiate between Web Content Mining and Web Structure Mining.

(5 + 3 + 2)

-------------------*-------------------