MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

Reg. No.										

II SEMESTER M.TECH. (COMPUTER NETWORKING AND ENGINEERING)

END SEMESTER EXAMINATIONS, APRIL 2017

SUBJECT: PROGRAM ELECTIVE - I DATA WAREHOUSING AND DATA MINING [ICT 5234]

REVISED CREDIT SYSTEM
(25/04/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer ALL the questions.
- Missing data if any, may be suitably assumed.

1A. Construct Pattern Count Tree for the following database and compute the frequent items considering a minimum support count of 3, for the following set of transactions {T1={1,5,6,8}, T2={2,4,8}, T3={4,5,7}, T4={2,3}, T5={5,6,7}, T6={2,3,4}, T7={2,6,7,9}, T8={5}, T9={8} T10={3,5,7}, T11={3,5,7}, T12={5,6,8}, T13={2,4,6,7}, T14={1,3,5,7}, T15={2,3,9}}. **5**

1B. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 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. **3**

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

(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years

1C. Discuss any four requirements for clustering in data mining. **2**

2A. Check whether the decomposition R= (R1, R2) is dependency preserving, where R= (A,B,C,D,E) R1 = (A,B,C), R2 = (A, D, E) and F ={ A→BC, CD→E, B→D,E→A}. **5**

2B. Suppose that a data warehouse consists of three dimensions 'time', 'doctor' and 'patient' and the two measures 'count' and 'charge' where 'charge' is the fees that a doctor charges for a visit and 'count' indicates number of visits. 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 2016? **3**

2C. Define Functional dependency and also find all functional dependency corresponding to the relation R=(A,B,C) given in the Table Q.2C. **2**

Table Q.2C

A	B	C
a1	b1	c1
a1	b1	c2
a2	b1	c1

3A. Compute the frequent items using Dynamic Itemset Counting technique considering a minimum support count of 2 and stops after every three consecutive transactions for the following set of transactions : {T1={I1,I2,I5}, T2={I2,I4}, T3={I2,I3}, T4={I1,I2,I4}, T5={I1,I3}, T6={I2,I3}, T7={I1,I3}, T8={I1,I2,I3,I5}, T9={I1,I2,I3}} **5**

3B. Write the Apriori algorithm to find all frequent itemsets. 3

3C. What is noise? List all the methods available for smoothing of data. 2

4A. Use information gain method to find the first best split attribute in the decision tree for the training set given in the Table Q.4A. 5

Table Q.4A

Income Range	Credit card insurance	Gender	Age	Life Insurance promotion (Class label)
40-50K	No	M	45	No
30-40K	No	F	40	Yes
40-50K	No	M	42	No
30-40K	Yes	M	43	Yes
50-60K	No	F	38	Yes
20-30K	No	F	55	No
30-40K	Yes	M	35	Yes
20-30K	No	M	27	No
30-40K	No	M	43	No
30-40K	No	F	41	Yes
40-50K	No	F	43	Yes
20-30K	No	M	29	Yes
50-60K	No	F	39	Yes
40-50K	No	M	55	No
20-30K	Yes	F	19	Yes

4B. 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} ,MFS={} 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} }. Find the following: 3

(i) MFCS , MFS , when k=2.

(ii) L_2 after MFS pruning.

4C. Distinguish between supervised learning and unsupervised learning. 2

5A. Suppose that the data mining task is to cluster the following points into 3 clusters. A1(2,12), A2(4,6), A3(9,5), B1(5,8), B2(7,5), B3(6,4), C1(2,5), C2(4,8). Suppose initially we assign A1, B1 and C1 as the center of each cluster, solve using k-means algorithm for 2 iterations. 5

5B. Consider the 2 X 2 contingency table given in Table Q.5B, summarizing observed count and the total transactions with respect to type of drinks and snacks, students of an Engineering college preferred. 3

Table Q.5B

		Drinks		
		Milk	*Coffee*	*∑row*
Snacks	*Pizza*	2000	1000	3000
	Burger	1000	1500	2500
	∑col	3000	2500	5500

Use $\chi 2$ test to check the dependency of Snacks and Drinks for degree of freedom n=1, significance level = 0.001and $\chi 2 = 10.828$.

5C. Write the K-medoid algorithm. 2