Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL 576104

(Constituent College of Manipal University)

SIXTH SEMESTER B.Tech. (IT) DEGREE MAKE-UP EXAMINATION JULY – 2016
SUBJECT: DATA WAREHOUSING AND DATA MINING – ICT 306
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS **/07/2016** **MAX. MARKS: 50**

Instructions to candidates

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

1A. A group of University students took part in a sponsored race. The number of laps completed is given in the table below. Use the information to calculate an estimate for the mean number of laps and draw a boxplot based on the five number summary.

Number of laps	frequency
1-5	2
6-10	9
11-15	15
16-20	20
21-25	17
26-30	25
31-35	2
36-40	1

1B. Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: loose coupling, semi-tight coupling, and tight coupling. State which approach you think is the most popular? Justify.

1C. The mean of 40 data values was 160. It was observed that, the value of 165 was wrongly copied as 125. Find the correct mean. [5+3+2]

2A. Data Set={ T1:1,2,3,4,5,6; T2:1,2,3,7; T3:1,2,4,8; T4:2,3,4,5,9; T5:1,2,3; }
Considering the support count as 2 for the above data set, during the execution of Pincer-Search algorithm, it is found that, when k=1, MFCS={{1,2,3,4,5}}
when k=2, L_2={{1,2},{1,3},{1,4},{2,3},{2,4},{2,5}, {3,4},{3,5},{4,5}} S_2={{1,5}
when k=3, L_3={{1,2,3},{1,2,4},{2,3,4},{2,3,5},{2,4,5}, {3,4,5}} S_3={{1,3,4} MFCS={{1,2,3,4,5}}
Find the followings:
(i) MFCS and MFS when k=2 (ii) MFCS and MFS when k=3 (iii) L_3 after MFS Pruning
(iv) C_4 after Recovery (v) C_4 after MFCS-Prune

2B. Write the pseudo code for Dynamic Itemset Counting algorithm.

2C. Differentiate between roll-up and roll-down OLAP operations. Consider the following cube illustrating temperature of certain days recorded weekly. Show the result of roll-up operation (Temperature) for this cube by assuming levels hot (80-85), mild (70-75), cold (64-69) for Temperature.

Temperature	64	65	68	69	70	71	72	75	80	81	83	85
Week 1	1	0	1	0	1	0	0	0	0	0	1	0
Week 2	0	0	0	1	0	0	1	2	0	1	0	0

[5+3+2]

3A. Find all the frequent patterns for the data set given below by using Partitioning algorithm by considering a partition size of 3 transactions. Show all the steps clearly.
Data Set={ T1:{1,2,3,4}; T2:{1,2,4}; T3:{1,2}; T4:{2,3,4}; T5:{2,3}; T6:{3,4;} [min_sup>= 2]

3B. Briefly describe the join indices. Apply join index for the table Sales and Products given below and write the table showing the result of join index operation.

Sales

Prod_ID	Store_ID	Date	Amount
P1	C1	1	12
P2	C1	1	11
P1	C3	1	50
P2	C2	1	8
P1	C1	2	44
P1	C2	2	4

Products

ID	Name	Price
P1	Bolt	10
P2	Nut	5

3C. Ms. Jepkoech, a Data Analyst at the Smart Swiss Biscuits Ltd, summarized statistics of two data sets of biscuit strength measurements as shown below:
Sample A: 209 129 194 132 173 381 282 283 518 267
Sample B: 203 274 381 282 283 518 267 309 334 417
Draw the quantile-quantile plot for the data given above. [5+3+2]

4A. Explain any one density-based methods by giving an example.

4B. Consider the dataset given below. Let c1(1.5,4.0) be the cluster 1 medoid and c2(1.6,2.0) be the cluster 2 medoid. Apply k-medoid clustering algorithm and check whether swapping the centroid from O4 to O6 would result in better clustering?
O1 (1.5,4); O2(1.5,2.5); O3(1.5,2.5); O4(1.6,2); O5(1.8,1); O6(2,0.5); O7(2.3,2.5)

4C. Discuss the pre-processing steps involved in preparing the data for Classification and Prediction. [5+3+2]

5A. Construct the frequent patterns from the given set of transactions using FP Tree Growth algorithm.
T1: {1,3,4}; T2: {2,3,5}; T3: {1,2,3,5}; T4: {2,5}; T5:{1,3,5}

5B. Apply 3-4-5 rule to generate concept hierarchy of 3 levels for the values:
-34,-26,-22,-22,-20,-20,-18,-18,-16,-16,19,21,24,26,26,28,28,29,32,32,34,35,42,46,46

5C. Discuss the problems faced by today's search tools in finding relevant information on the web. [5+3+2]

6A. Given initial seeds as X1 and X4, obtain clusters for the given dataset by applying k-means algorithm. Dataset = { X1(2,10); X2(2,5); X3(8,4); X4(9,4); X5(5,8); X6(1,2); X7(4,9) }
Also, check whether swapping the initial seeds to X2 and X5 would result in a better clustering.

6B. What is the difference between symmetric and asymmetric binary variables? Consider the relational table where name is an object identifier, gender is a symmetric attribute, and remaining attributes are asymmetric binary. Calculate the distance between each pair of three entities.

Name	Gender	A1	A2	A3
Ben	M	Y	N	Y
Ema	F	Y	N	Y
Joe	M	Y	Y	N

6C. The Probability of playing both cricket and football is 40%. The probability of playing football is 50%. There exists positive correlation between cricket and football. The Correlation measure, Lift between cricket and football is 2. Find the dependent/correlation measures all_confidence and cosine. [5+3+2]