Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (COMPUTER SCIENCE & ENGINEERING)

MAKEUP EXAMINATIONS, JUNE 2017

SUBJECT: DATA WAREHOUSING AND DATA MINING [CSE 4007]

REVISED CREDIT SYSTEM
(20/06/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL FIVE** questions.
- Missing data may be suitable assumed.

1A. What kind of data and patterns can be mined? What makes a pattern interesting? **3M**

1B. Explain the different measures used to assess the dispersion of numeric data. How to check whether a given distribution is normal distribution or not? **3M**

1C. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 18,19,20, 21,22,23,28,29,30,31, 32,33, 60,61,62. Show the results for each of the following preprocessing techniques:

i. Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps.
ii. Transform the data using z-score normalization.
iii. Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, stratified sampling. Use samples of size of 5 and class value "youth"(18-29), "middle-aged"(30-59), and "senior"(60 and above) **4M**

2A. Give Apriori algorithm for discovering frequent itemsets for mining Boolean association rules. How the association rules are generated from frequent itemsets? **4M**

2B. A database with nine transactions is given in Table 2B:

Table 2B

TID	T1	T2	T3	T4	T5	T6	T7	T8	T9
List of Item IDs	A,B, E,	B,D	B,C	A,B, D	A,C	B, C	A, C	A,B, C, E	A,B, C

Generate frequent patterns using Vertical Data Format(Min. Support = 2(22%)) **3M**

2C. The contingency table given in Table 2C summarizes the transactions with respect to Play chess and like Science fiction.

Table 2C

	Play chess	**Not play chess**	**Sum (row)**
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

a. With min. sup = 25% and min. conf. = 50%, is association rule "Ply chess" → "Like science fiction" strong?

b. Based on the given data, is Play chess independent of Like science fiction? If not, what kind of correlation relationship exists between the two? Use both lift and Chi-Square measures.. **3M**

3A. Build a decision tree using information gain as attribute selection measure for the data set(class label: buys_computer) given in Table 3A.

Table 3A

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

3M

3B. Explain the working principle of SVM by considering a linearly separable data. How it can be extended to linearly inseparable data? **4M**

3C. Briefly explain the following techniques in improving classification accuracy:
i. Bagging ii. Boosting iii. Random Forest **3M**

4A. For the data given in Table 4A:

Table 4A

Object ID	**Test-1 Categorical**	**Test – 2 Ordinal**	**Test-3 Numeric**
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	Good	64
4	Code A	Excellent	28

Compute the dissimilaarity matrix by considering: **4M**
i. Test-1 only ii. Test-2 only iii. Test-1, Test-2 and Test-3 attributes

.

4B. In the context of density based clustering, illustrate directly density-reachable, desnity-reachable, density-connected. Also, give pseudocode for DBSCAN algorithm **4M**

4C. Briefly discuss the following methods used in determining the number of clusters.

i. Empirical method ii. Elbow method iii. Cross validation method **2M**

5A. Give a block diagram to show the building blocks of data warehouse. Briefly explain the purpose of each component. **4M**

5B. Suppose that a Sales Facts data warehouse consists of the four dimensions, *product, store, time,* and *promotion*, and the two measures, *Sold Quantity* and *Net Sales*, where Sold Quantity / *Net Sales* is the No. of units sold / profit over the sales of a product in a particular store on a given date. Products may be of different categories and brands. Stores are located in each district headquarters across the sates. Based upon the season, different promotion schemes are used to promote the sales.

(a) Draw a *star schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*product, store, time, promotion*], what specific *OLAP operations* should one perform in order to list total Net Sales of Reliance brand in Karnataka State in the year 2004? **3M**

5C. Illustrate the usage of group by cube in creating the cube on *sales*(*item_name, color, clothes_size, quantity*).
Explain, how the null values in the resultant cube can be addressed by grouping() and decode() functions? **3M**