Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (COMPUTER AND COMMUNICATION ENGINEERING)

END SEMESTER EXAMINATIONS, APRIL/MAY 2019

SUBJECT: DATA MINING AND PREDICTIVE ANALYSIS [ICT 3252]

REVISED CREDIT SYSTEM

(27/04 /2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Compute the frequent itemset for the transactions given in Table Q.1A. using Pincer-Search algorithm. Assume minimum support count as 2. **5**

Table Q.1A.

Transactions	Items
T1	a,b,c,d,e,f
T2	a,b,c,g
T3	a,b,d,h
T4	b,c,d,e,k
T5	a,b,c

1B. Explain the following data pre-processing techniques with an example: **3**

i. Data smoothing

ii. Aggregation

iii. Attribute construction

1C. State the disadvantage of partitioning methods in cluster analysis. Discuss how outliers affect the performance of k-means clustering algorithm. **2**

2A. i. Why is it important to perform tree pruning. Explain the two methods of tree pruning.

ii. Consider the data given in Table Q.2A. and obtain the root node using Gini index as attribute selection measure. **5**

Table Q.2A.

Owns home	Marries	Gender	Employed	Credit rating	Class
Yes	Yes	Male	Yes	A	B
No	No	Female	Yes	A	A
Yes	Yes	Female	Yes	B	C
Yes	No	Male	No	B	B
No	Yes	Female	Yes	B	C
No	No	Female	Yes	B	A
No	No	Male	No	B	B
Yes	No	Female	Yes	A	A
No	Yes	Female	Yes	A	C
Yes	Yes	Female	Yes	A	C

2B. What is temporal data mining? Explain the various temporal data mining tasks. **3**

2C. Consider the transactional dataset given in Table Q.2C. and obtain the frequent itemset using Vertical data format. Assume minimum support count as 2. **2**

Table Q.2C.

TID	Itemset
T1	A,B,E
T2	B,D
T3	B,C
T4	A,B,D
T5	A,C
T6	B,C
T7	A,C
T8	A,B,C,E
T9	A,B,C

3A. Find the dissimilarity matrix for the dataset given in Table Q.3A. Assume rank(A^+)=3, rank(A)=2 and rank(B)=1. For ordinal attribute, use Euclidean distance metric for distance computation. **5**

Table: Q.3A.

Sl No	Gender (Binary symmetric)	Grade (Ordinal)	Activities (Nominal)	Marks (Numeric)
1	Male	A	Dance	8
2	Female	A^+	Music	9
3	Male	B	Dance	6
4	Female	A	Instrument	8

3B. Consider the contingency matrix given in Table Q.3B. and find whether marital status and education are correlated. Assume χ^2 value as 26.22 for the significance level 0.01. **3**

Table Q.3B.

	Middle school	High school	Bachelors	Masters	PhD
Never Married	18	36	21	9	6
Married	12	36	45	36	21
Divorced	6	9	9	3	3
Widowed	3	9	9	6	3

3C. Why are not all strong association rules interesting? Illustrate with an example. **2**

4A. i. Write the pseudo code for prune and candidate set generation functions in Apriori algorithm.

ii. Consider the dataset given in Table Q.4A. and compute the frequent itemsets using Apriori algorithm. Assume minimum support count as 2. **5**

Table Q.4A.

Transactions	Itemset
T1	1,2,5,8
T2	2,3,9,4
T3	1,2,3,5,9
T4	8,9
T5	3,4,5,6,7
T6	7,8,9,1,2
T7	3,5,7,9

4B. Explain the different categories of web mining with a real time example. **3**

4C. How do you obtain clusters using DBSCAN algorithm? Explain. **2**

5A. i. List and explain the techniques of text mining.

ii. Considering the topology given in Figure Q.5A., obtain the PageRank of web pages A, B and C after three iterations. Assume damping parameter, d =0.7, initial page rank of A, B and C as 1 and start node as A. **5**

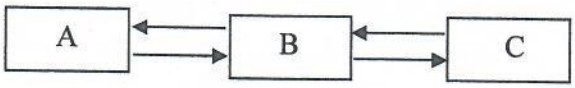

Figure Q.5A.

5B. Obtain the clusters for the dataset given in Table Q.5B. using k-means algorithm. Assume the initial centroids C1 and C2 as object A and B respectively and number of cluster, k=2. **3**

Table Q.5B.

Object Id	X	Y
A	1	1
B	2	1
C	4	3
D	5	4

5C. Define data characterization and data discrimination with an example for each. **2**