Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

II SEMESTER M.TECH. (CSE/CSIS)

END SEMESTER EXAMINATIONS, APRIL 2018

SUBJECT: DATA MINING AND APPLICATION [CSE5236]

REVISED CREDIT SYSTEM
(23/04/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A Define Data Mining and on what kind of data it can be performed? **4**

1B Consider the data given in the contingency Table Q1B. Determine whether Gender independent of the education level? (**Note**: chi –square value for significance level 0.05 and degrees of freedom 3 is equal to 7.815) **4**

Table Q1B

Gender	**High School**	**B.E**	**M.Tech**	**Ph.D**	**Total**
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

1C List the factors which are responsible for high quality data. **2**

2A Construct Frequent Pattern Tree for the database given in Table Q2A and also find all the frequent itemsets with respect to the minimum support count is 2. **5**

Table Q2A

Tid	**Items**	**Tid**	**Items**
1	1,2,5	6	2,3
2	2,4	7	1,3
3	2,3	8	1,2,3,5
4	1,2,4	9	1,2,3
5	1,3		

2B How the hash-based technique can be used to improve the efficiency of Apriori algorithm? Use the same to find all frequent 2-itemsets with respect to the minimum support is equal to 2 transactions for the database given in Table 2QB. Use H2= [x + (y *10)] mod 5 as a hash function. **3**

Table 2QB

TID	**Items**	**TID**	**Items**
1	3,4	**6**	1,3
2	2,3	**7**	2
3	1,2 ,3,5	**8**	1,3
4	2,5	**9**	1,2,3
5	1,2	**10**	1,3

2C List the variations of multilevel association rules mining and then explain any one of them. **2**

3A It is found that Eigen values of the covariance matrix corresponding to the data with two variables x and y are 0 .049 and 1.284 respectively and its Eigen vectors matrix V is **3**

$$\begin{matrix} \mathbf{-0.7351} & \mathbf{-0.67} \\ \mathbf{0.6778} & \mathbf{-.7351} \end{matrix}$$

Find the percentages of the variables explained by the first two principal components.

3B Discuss the steps involved in data classification **2**

3C Consider the tuples corresponding to the weather forecast for cricket match given in the Table Q3A. The class label attribute ‘Play’ has two values ‘YES’ and ‘NO’. Find the best splitting point for the attribute Outlook to construct binary decision tree and also use Bayesian Classification to find the Class to which the following tuples belong in the data set given in the Table Q3C. ? **5**

(i) **X = < rain, hot, high, false >**

(ii) **Y = < sunny, hot, high, false>**

Table Q3C

Outlook	**Temperature**	**Humidity**	**Windy**	**Play ?**
sunny	hot	high	false	NO
sunny	hot	high	true	NO
overcast	hot	high	false	YES
rain	mild	high	false	YES
rain	cool	normal	false	YES
rain	cool	normal	true	NO
overcast	cool	normal	true	YES
sunny	mild	high	false	NO
sunny	cool	normal	false	YES
rain	mild	normal	false	YES
sunny	mild	normal	true	YES
overcast	mild	high	true	YES
overcast	hot	normal	false	YES
rain	mild	high	true	NO

4A Use K-medoids algorithm and Manhattan distance measure to discover two clusters by considering **(3, 4)** and **(7,4)** as cluster medoids for following set of data objects. **5**

{(**2, 6), (3, 4), (3, 8), (4, 7), (6, 2), (6, 4), (7, 3), (7, 4), (8, 5), (7, 6**)}.

Check whether the replacement given in each one of the following on the initial clusters formed is a good replacement or not

(i) **(7, 4) by (8, 5**)

(ii) **(3, 4) by (2, 6**)

.

4B Write a Density Based Clustering Algorithm. **2**

4C List the types of problems that can be solved by text mining and explain any one of them. 3

5A Discuss the different window models to process and mine frequent patterns in data streams **5**

5B What is tokenization? Write an algorithm to generate features from tokens. **3**

5C Write an algorithm for End-of-sentence detection. **2**

****