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 2018

SUBJECT: DATA MINING & PREDICTIVE ANALYSIS [ICT 3252]

REVISED CREDIT SYSTEM
(20/04/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. A database has five transactions as given in Table Q.1A. Let minimum support = 60% and minimum confidence = 80%. Find all frequent itemsets using Apriori and FP_Growth algorithms respectively. **5**

Table Q.1A

TID	Items_bought
T100	M,O,N,K,E,Y
T200	D,O,N,K,E,Y
T300	M,A,K,E
T400	M,U,C,K,Y
T500	C,O,O,K,I,E

1B. What are the two common approaches to tree pruning? Explain with an example. **3**

1C. A classification model may change dynamically along with the changes of training data streams. This is known as concept drift. Explain why decision tree induction may not be a suitable method for such dynamically changing data sets. Is naive Bayesian a better method on such data sets? Explain your reasoning. **2**

2A. Write the k-means algorithm. Suppose that the data mining task is to cluster the following eight points (with (x; y) representing location) into three clusters. **5**
A1(2,10); A2(2,5); A3(8,4); B1(5,8); B2(7,5); B3(6,4); C1(1,2); C2(4,9)
The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster respectively. Use the k-means algorithm to find
 i. The three cluster centers after the first round of execution and
 ii. The final three clusters

2B. Given Minimum support=3 and M = 2, apply Dynamic Itemset Counting algorithm on the dataset given in Table Q.2A to obtain the frequent patterns. 3

Table Q.2A

T_ID	Itemsets
T1	1,2,4
T2	1,3,4,5
T3	1,4,5,6
T4	2,5,6
T5	1,2,4,5,6

2C. A database contains 80 records on a particular topic .A search was conducted on that topic and 60 records were retrieved. Of the 60 records retrieved, 45 were relevant. 2
Calculate the precision and recall scores for the search. Also, show that accuracy is a function of sensitivity and specificity

3A. Consider a Table Q.3A of tuples which tells whether a person will default his loan or not. Predict using naïve bayes classification whether Mr.Ratri would default his loan if he doesn't own a house and is married with a job experience of 3years. 5
Also show, how the Laplacian correction is used to avoid computing probability values of zero?

Table Q.3A

Home Owner	Marital Status	Job Experience	Default?
Yes	Single	3	NO
No	Married	4	NO
No	Single	5	NO
Yes	Married	4	NO
No	Divorced	2	YES
No	Married	4	NO
Yes	Divorced	2	NO
No	Married	3	YES
No	Married	3	NO
Yes	Single	2	YES

3B. The contingency table in Table Q.3B summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, $\overline{\textit{hot dogs}}$ refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and $\overline{\textit{hamburgers}}$ refers to the transactions that do not contain hamburgers. 3

Table Q.3B

	hot dogs	$\overline{\textit{hot dogs}}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{\textit{hamburgers}}$	1000	1500	2500
Σ_{col}	3000	2000	5000

i. Suppose that the association rule "hot dogs → hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

ii. Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?

3C. What is temporal data mining? What are the different types of temporal data? 2

4A. Explain the four clustering methods with an example for each 5

4B. Briefly describe the three key components of Web Mining. Give one application for each component respectively 3

4C. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning. (**Example:** Age in years. **Answer:** Discrete, quantitative, ratio) 2

i. Brightness as measured by a light meter
ii. Angles as measured in degrees between 0° and 360°.
iii. Bronze, Silver, and Gold medals as awarded at the Olympics.
iv. Military rank.

5A. 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. 5

i. What is the mean, median and mode of the data? Comment on the data's modality.
ii. Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
iii. Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
iv. Use smoothing by bin means to smooth the above data, using a bin depth of 3.

5B. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 3

5C. With a sample scenario, explain why accuracy alone is a bad measure for classification tasks? How can it be resolved? 2