24

SRN | | | | | | | | | |

REVA UNIVERSITY
Bengaluru, India

Sixth Semester B.Tech CSE Semester End Examination May/June 2018
Course (Subject): Data Mining Techniques
Course Code: BTCS15F6410

Time: 3 Hours **Max. Marks: 100**

Note: *Answer ONE FULL question from each unit.*

UNIT – I

1. a) Knowledge Discovery in Databases(KDD) is very important to know the hidden information in the databases, Explain the process of KDD with a neat diagram. 10
   b) In traditional data analysis techniques, we often encounter practical difficulties in meeting the challenges posed by new data sets, List and explain such motivating challenges 10
   c) Data Mining tasks are broadly divided into 2 categories. Explain the two major Data Mining Tasks in brief. 5

OR

2. a) There may be problems in the data due to human errors, limitations of measuring devices, or flaws in the data collection process. List and brief out the different Measurement and Data Collection Errors/Issues. 15
   b) Knowledge Discovery in Databases(KDD) is very important to know the hidden information in the databases, Explain the process of KDD with a neat diagram. 10
      i. X=(1,0,0,0,0,0,0,0,0,0) ,Y=(0,0,0,0,0,0,1,0,0,1) Find SMC and J.
      ii. X=(3,2,0,5,0,0,0,2,0,0),Y=(1,0,0,0,0,0,0,1,0,2) Find Cos(X,Y).

UNIT – II

3. a) Consider the following transaction database for a super market in the given table:Generate Apriori rules by assuming the minimum support of 30% (atleast three transactions) and minimum confidence of 60%. 15

Customers	Items
C1	Milk,egg,bread,chip
C2	Egg,popcorn,chip,beer
C3	Egg,bread,chip
C4	Milk,egg,bread,popcorn,chip,beer
C5	Milk,bread,beer
C6	Egg,bread,beer
C7	Milk,bread,chip
C8	Milk,egg,bread,butter,chip
C9	Milk,egg,butter,chip

   b) The number of frequent Itemset produced from a transaction data set can be large. It is useful to identify a small representative set of Itemset from which all other frequent Itemset can be derived. Explain the two-compact representation of frequent itemset, with example. 10

OR

25

4. a) The use of Support for Pruning candidate Itemset is guided by Apriori Principle. With the Algorithm explain the candidate Itemset generation using Apriori Principle 10

   b) For the below table: 10

TID	T100	T200	T300	T400	T500	T600	T700	T800	T900
Item-IDS	I1,I2, I5	I2,I4	I2,I3	I1,I2, I4	I1,I3	I2, I3	I1,I3	I1,I2, I3,I5	I1,I2,I3

   i. Construct an FP Tree

   ii. Generate the list of frequent Itemset ordered by their corresponding suffixes.

   c) Support and Confidence measures are very important in Association Analysis. How are they contributing to Association Analysis? Explain. Also, Find out the Support and Confidence for the following transaction details. 5

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

UNIT – III

5. a) Classification technique is used in many fields like detecting spam e-mails, classifying Galaxies based upon their shapes etc. Brief about classification technique. Also, explain general approach for solving a classification problem 10

   b) Justify with the characteristics and an example why the Decision Tree Classifier is widely used classifier? 10

   c) Illustrate with an example, the working of Decision Tree Classifier. 5

OR

6. a) Decision tree induction algorithms must provide a method for expressing an attribute test condition and its corresponding outcomes for different attribute types. Explain in detail. 10

   b) Write an algorithm for decision tree induction, also explain each step in detail. 10

   c) Explain the different strategies for growing a classification rule. 5

UNIT – IV

7. a) Clustering aims to find useful groups of objects, where usefulness is defined by the goals of the data analysis. List and explain different types of clusters. 10

   b) K-means is a prototype based, partitional clustering technique which attempts to find a user-specified number of clusters(K). Write and explain basic K-Means algorithm to find K-clusters. 10

   c) Some clustering technique use hierarchical based approaches main clustering techniques among them are agglomerative and divisive clustering. Highlight the important characteristics of the above algorithms with an illustrative example. 5

OR

26

8. a) Density based clustering locates regions of high density that are separated from one another by regions of low density. Describe DBSCAN algorithm with an example. 10

b) Differentiate between K-means and DBSCAN. 5

c) Clustering involves grouping data based on similarities. Clustering is being used in market segment analysis for grouping customers based on the amount of money spent. Given a set of data points and distances between each data points. Apply Agglomerative clustering to cluster the given data points. Use MIN – link and MAX - link for proximity measure 10

Points	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

***