U.S.N ☐☐☐☐☐☐☐☐☐☐

P.E.S. College of Engineering, Mandya - 571 401
(An Autonomous Institution affiliated to VTU, Belagavi)
Eighth Semester, B.E. – Computer Science and Engineering
Semester End Examination; May/June - 2018
Data Mining and Analytics

Time: 3 hrs   Max. Marks: 100

Note: Answer **FIVE** full questions, selecting **ONE** full question from each unit.

UNIT - I

1a.	What is data mining? Explain the challenges that motivated the development of data mining.	6
b.	Explain the Knowledge Discovery in Databases (KDD) with neat diagram.	6
c.	List all the data preprocessing techniques and in short explain any three.	8
2 a.	Explain different types of datasets with example.	10
b.	For the following vectors X and Y calculate the cosine, correlation, Euclidian and Jaccord similarity; X: (0, 1, 0, 1) and Y: (1, 0, 1, 0).	10

UNIT - II

3 a.	Write the characteristics of Decision tree induction.	6
b.	Compare rule based ordering scheme with class based ordering scheme.	4
c.	Explain k-nearest neighbor classification algorithm.	10
4 a.	Explain the process of building a classifier model with a neat diagram.	6
b.	Explain why attribute relevance analysis is needed? How it can be performed using Gini index and entropy.	6

c. Consider a binary classification problem with the following set of attributes and attribute values:

Air conditioner = {working, broken}

Engine = {good, bad}

Mileage = {high, medium, low}

Rust = {yes , no}

Suppose a rule based classifier produces the following rule set :

Mileage = high → value = low

Mileage = low → value = high

Air conditioner = working , Engine = good → value = high

Air conditioner = broken → value = low

    8

  i) Are the rules mutually exclusive?   ii) Is the rule set exhaustive?

  iii) Is ordering needed for this set of rules?   iv) Do you need a default class for the rule set?

UNIT - III

5 a. Consider the data set shown in Table 5a

Table 5a

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

10

  i) Estimate the conditional probabilities for P (A|+), P (B|+), P (C|+), P (A|-), P (B|-) and P (C|-)

  ii) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample(A = 0, B = 1, C = 0) using Naïve Bayes approach.

  b. Explain with an algorithm learning perceptron model. 10

6 a. Explain how to estimate conditional probabilities for continuous attributes? 10

  b. State Mercer's theorem. Explain characteristics of SVM. 10

UNIT - IV

7 a. With an algorithm explain frequent item set generation of the Apriori algorithm. 8

  b. Define the following :

  i) Maximal frequent item set   ii) Closed frequent item set

4

  c. Compare the following traversal of item set lattice :

  i) General to specific versus specific to general   ii) Breadth first versus Depth first

8

8 a. Discuss the computational complexity of the Apriori algorithm. 10

  b. Explain frequent item set generation in FP growth algorithm. 10

UNIT - V

9 a. What is cluster analysis? Explain different types of clustering. 10

  b. With an algorithm, explain Bisecting k-means algorithm. 10

10 a. List and explain different types of Clusters. 8

  b. Use the similarity matrix in Table 10b to perform single and complete link hierarchical clustering. Show the result by dendrogram. The dendrogram should clearly show the order in which the points are merged.

Table 10b

12

	P_1	P_2	P_3	P_4	P_5
P_1	1.00	0.10	0.41	0.55	0.35
P_2	0.10	1.00	0.64	0.47	0.98
P_3	0.41	0.64	1.00	0.44	0.85
P_4	0.55	0.47	0.44	1.00	0.76
P_5	0.35	0.98	0.85	0.76	1.00

* * * *