DEC - 2018

[Total No. of Questions : 8] [Total No. of Pages : 3]

M.E. FIRST YEAR (Semester-I)
(Computer Engineering)
Data Mining
(510105B) Elective-I
(2017 Course)

[Time: 3 Hours] **[Max Marks: 50]**

Instructions to the candidates:-

1) **Solve question number 1 or 2, 3 or 4, 5 or 6 and 7 or 8.**
2) **Neat diagram must be drawn whenever necessary.**
3) **Black figures to the right indicate full marks .**
4) **Assume suitable data ,if necessary.**

Q 1) a) Explain the Apriori algorithm for discovering frequent item set for mining Boolean association rules. [5 M]

b) Compare OLAP and OLTP systems. Explain the steps in KDD with a suitable block diagram. [5 M]

OR

Q 2) a) Explain cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables. [5 M]

b) Discuss Data visualization with reference to Data Mining. [5 M]

Q 3) a) Write short notes for the following in detail: [5 M]

(i) Measuring the central tendency
(ii) Measuring the dispersion of data.

b) What is the most appropriate measure of central tendency when the data has outliers? [5 M]

OR

Q 4) a) Explain basic statistical description of Data. [4 M]

b) Data Matrix versus Dissimilarity Matrix [3 M]

c) Explain the Data Types and Distance Metrics. [3 M]

Q 5) a) Why is euclidean distance considered as best to calculate similarity between two feature sets having independent parameters? [5 M]

b) Which distance formula should we use for faster performance, Manhattan distance or Euclidean distance, and why? [5 M]

OR

Q 6) a) What is the difference between Euclidean distance and Hamming distance? [5 M]

b) How to create a dissimilarity matrix for mixed type dataset explain with example. [5 M]

Q7) a) Apply the Naive Bayes Classifier [5 M]

Attributes are Color, Type, Origin, and the subject, stolen can be either yes or no.

Example No.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

b) Discuss the Challenges faced with Imbalanced datasets with example. [5 M]

c) When to use boosting and bagging? What is the difference between bagging and ensembling method? [5 M]

d) Identify the ML methods that are suitable for multi-label classification & how can we apply these methods using WEKA? [5 M]

OR

Q8) a) Why is tree pruning usefull in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? [5 M]

b) How can we evaluating the Accuracy of a classifier or Predictor?. [5 M]

c) Explain Support Vector Machine. Pros and Cons associated with SVM. [5 M]

d) Design Issues of Decision Tree Induction? [5 M]

****