DATA MINING & KNOWLEDGE DISCOVERY
(MCAP 2251)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following is required by K-means clustering?
(a) Defined distance metric (b) Number of clusters
(c) Initial guess as to cluster centroids (d) All of the above.

(ii) What are support vectors?
(a) These are the data points which help the SVM to generate optimal hyper plane
(b) It is an intermediate vector generated during calculation of optimal hyper plane
(c) In SVM all the data points are called support vectors
(d) This are predefined vectors used in calculating hyper plane.

(iii) Which of the following is not applicable in Data Mining?
(a) Knowledge extraction (b) Data exploration
(c) Data transformation (d) None of these.

(iv) The Sigmoid function is defined as $f(t) =$
(a) $\frac{1}{\exp(t)+\exp(-t)}$ (b) $texp(-t)$
(c) $\frac{1}{1+\exp(t)}$ (d) $\frac{1}{1+\exp(-t)}$

(v) Which of the following algorithms cannot be used for reducing the dimensionality of data?
(a) t-SNE (b) PCA (c) LDA (d) None of these.

(vi) Which one of the clustering techniques needs the merging approach?
(a) Partitioned (b) Naïve Bayes
(c) Hierarchical (d) Both (a) and (c).

(vii) Which layer in ANN can receive data from external sources?
(a) Input layer (b) Output layer
(c) Hidden layer (d) All the above.

(viii) Slack variable is applicable for

(a) SVM (b) Bayes classifier

(c) K-means (d) None of the above

(ix) DBSCAN cannot be used (with high accuracy) for datasets that are

(a) convex (b) uniform density

(c) non-uniform density (d) none of the above

(x) Which of the following statements is not true about k-Nearest Neighbor classification?

(a) The output is a class membership

(b) An object is classified by a plurality vote of its neighbors

(c) If k = 1, then the object is simply assigned to the class of that single nearest neighbor

(d) The output is the property value for the object.

Fill in the blanks with the correct word

(xi) Support vector machines is _____.

(xii) Frequency of occurrence of an item set is called as _____.

(xiii) KDD refers to _____.

(xiv) The issues like efficiency, scalability of data mining algorithms comes under _____.

(xv) Weighted sums are passed via activation functions and _____ is produced.

Group - B

2. (a) State Aproiri algorithm. Discuss the drawbacks of Aproiri algorithm.

[(CO1,CO3)(Remember/LOCQ)]

(b) Using the following data set using PCA reduce the dimension from 2 to 1.

Feature	example-1	example-2	example-3	example-4
x	4	8	12	8
y	12	5	4	13

[(CO1,CO2)(Remember/LOCQ)]

(3 + 3) + 6 = 12

3. (a) A transaction data set is given below.

Transaction Id	Items purchased
T1	bread, cheese, egg, juice.
T2	bread, cheese, juice.
T3	bread, milk, yogurt.
T4	bread, juice, milk.
T5	Cheese, juice, milk.

Generate association rules using Apriori algorithm in which support is 50% and confidence is 75%.

[(CO3)(Create/HOCQ)]

(b) Illustrate the following terms with relevant example.
(i) Support. (ii) Confidence. (iii) Lift. *[(CO1,CO3)(Understand/LOCQ)]*

8 + 4 = 12

Group - C

4. (a) Why naïve Bayesian classification is called naïve? Briefly outline the major ideas of naïve Bayesian classification. *[(CO4)(Analyse/IOCQ)]*

(b) Construct (induct) a decision tree using information gain from the data provided in the following table. Consider the Gender as the class label.

Sl No	Over 170(CM)	Eye	Hair length	Gender
1	No	Blue	Short	Male
2	Yes	Brown	Long	Female
3	No	Blue	Long	Female
4	No	Blue	Long	Female
5	Yes	Brown	Short	Male
6	No	Blue	Long	Female
7	Yes	Brown	Short	Female
8	Yes	Blue	Long	Male

[(CO4)(Apply/IOCQ)]

4 + 8 = 12

5. Write short notes on the followings:
- Bagging and Boosting.
- Rule Based Classification.
- K-Nearest Neighbour.

[(C01,CO4)(Remember/LOCQ)]

(4 + 4 + 4) = 12

Group - D

6. (a) Differentiate between logistic regression and SVM without a kernel with relevant example. *[(CO4)(Analyse/IOCQ)]*

(b) Illustrate the terms Hard-Margin and Soft-Margin SVM. Explain with diagram that why kernel tricks is used? *[(CO4)(Understand/LOCQ)]*

6 + (3 + 3) = 12

7. (a) Name the different activation functions used in ANN with diagram. *[(CO4)(Understand/LOCQ)]*

(b) Explain with relevant diagram the working principle of a perceptron in ANN. *[(CO4)(Remember/LOCQ)]*

6 + 6 = 12

Group - E

8. Perform K-means clustering on all the points in the following table, where K = 2. Randomly select the initial seeds and perform the algorithm for two iterations.

Points	X co-ordinate	Y co-ordinate
p1	1	9
p2	2	10
p3	7	4
p4	10	3
p5	5	6
p6	6	11
p7	3	4
p8	4	9
p9	8	1
p10	3	12
p11	7	6
p12	11	2

[(CO5,CO6)(Apply/IOCQ)]

12

9. (a) Define Core point, Border Point and Noise point in the perspective of DBSCAN clustering algorithm. *[(CO5)(Remember/LOCQ)]*

(b) Explain why DBSCAN does not work well for the data having varying density. *[(CO5)(Analyse/IOCQ)]*

(c) Briefly describe, a methodology to select the values of the parameters (viz., eps (the radius) and the minpts (the minimum points)) of the DBSCAN Algorithm *[(CO5)(Analyse/IOCQ)]*

3 + 4 + 5 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	51.04	40.63	8.33

Course Outcome (CO):

After the completion of the course students will be able to

MCAP2251.1 Describe basic concepts of data mining and related models.
MCAP2251.2 Explore data analysis by dimensionality reduction as well as information compression using PCA.
MCAP2251.3 Identify patterns using association rule mining.
MCAP2251.4 Deploy appropriate classification techniques to fit the data.
MCAP2251.5 Cluster the high dimensional data for better data organization.
MCAP2251.6 Implement the data mining algorithms for real world data.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*