DATA MINING & KNOWLEDGE DISCOVERY
(MCAP 2251)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

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

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Frequency of occurrence of an item set is called as _____
(a) Support (b) Confidence
(c) Support Count (d) Rules.

(ii) PCA is a
(a) linear method (b) non linear method
(c) continuous method (d) repeated method.

(iii) What does FP growth algorithm do?
(a) It mines all frequent patterns through pruning rules with lesser support
(b) It mines all frequent patterns through pruning rules with higher support
(c) It mines all frequent patterns by constructing a FP tree
(d) It mines all frequent patterns by constructing an item sets.

(iv) 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

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

(vi) What is the approach of basic algorithm for decision tree induction?
(a) Greedy (b) Top Down
(c) Procedural (d) Step by Step.

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

(viii) What does Apriori algorithm do?
(a) It mines all frequent patterns through pruning rules with lesser support
(b) It mines all frequent patterns through pruning rules with higher support
(c) Both (a) and (b)
(d) None of the above.

(ix) The data is then fed into the model and output from each layer is obtained this step is called _____
(a) Feed forward (b) Feed backward
(c) Input layer (d) Output layer.

(x) _____ is a pooling operation that selects the maximum element from the region of the feature map covered by the filter.
(a) Max Pooling (b) Average Pooling
(c) Global Pooling (d) None of these

Group - B

2. (a) Write down the Apriori Algorithm. Discuss the main drawbacks of Apriori Algorithm. *[(CO2)(Understand/LOCQ)]*

(b) A database has five transactions given as follows.

Transactions	Items
T100	{F,A,C,D,G,M,P}
T200	{A,B,C,F,L,M,O }
T300	{B,F,H,O}
T400	{B,K,C,P}
T500	{A,F,C,L,P,M,N }

Draw a FP Tree for the above data set in which the minimum support count value = 3. *[(CO2)(Apply/IOCQ)]*

6 + 6 = 12.

3. (a) Given the following data, using PCA reduce the dimension from 2 to 1.

Feature	Example-1	Example-2	Example-3	Example-4
x	3	9	12	8
y	10	5	4	15

[(CO3)(Execute/HOCQ)]

(b) Discuss the basic steps for Frequent-Pattern (FP) growth algorithm**.**
[(CO2)(Understand/LOCQ)]

8 + 4 = 12

Group – C

4. Write short notes on the followings: *[(CO1,CO4)(Remember/LOCQ)]*
(i) Naive Bayes.
(ii) Pruning in Decision Tree.
(iii) Precision and Recall.

(4 + 4 + 4) = 12

5. (a) What are the differences between over fitting and under fitting**.**

[(CO1,CO4)(Remember/LOCQ)]

(b) NASA wants to be able to discriminate between Martians (M) and Humans (H) based on the following characteristics: Green ∈ {N, Y }, Legs ∈ {2, 3}, Height ∈ {S, T}, Smelly ∈ {N, Y }. Our available training data is as follows**:**

[(CO4)(Create/HOCQ)]

Sl No	*Species*	*Green*	*Legs*	*Height*	*Smelly*
1	M	N	3	S	Y
2	M	Y	2	T	N
3	M	Y	3	T	N
4	M	N	2	S	Y
5	M	Y	3	T	N
6	H	N	2	T	Y
7	H	N	2	S	N
8	H	N	2	T	N
9	H	Y	2	S	N
10	H	N	2	T	Y

Learn a decision tree by building a decision tree by selecting a best attribute that yields maximum Information Gain (IG). Build the decision tree only for the first two levels (means for the root and the next level).

3 + 9 = 12

Group - D

6. (a) Illustrate the working principle between logistic regression and SVM.

[(CO4)(Remember/LOCQ)]

(b) Summarise the main objective of using SVM-Kernel. What is known as Support Vectors? *[(CO4)(Remember/LOCQ)]*

6 + 6 = 12

7. (a) Illustrate how a hyper plane can be drawn with the help of single layer perceptron for a data fitting problem. *[(CO4)(Remember/LOCQ)]*

(b) Illustrate the role of the activation functions in neural networks. List down the names of some popular activation functions used in neural networks.

[(CO4)(Remember/LOCQ)]

6 + 6 = 12

Group - E

8. (a) Explain some cases where K-Means clustering fails to give good results.

[(CO5)(Analyse/IOCQ)]

(b) How to determine k using the Silhouette method and Elbow method?

[(CO5)(Remember/LOCQ)]

(c) How to initialize of the centroids and improves the quality of the K-Means clustering algorithm. *[(CO5)(Analyse/IOCQ)]*

3 + 6 + 3 = 12

9. (a) Define minimum distance and maximum distances between two clusters. *[(CO5)(Remember/LOCQ)]*

(b) Construct the dendrograms for the following proximity matrix using both minimum distance and maximum distances. *[(CO5)(Analyse/IOCQ)]*

	P1	P2	P3	P4	P5
P1	1.00	0.90	0.10	0.65	0.20
P2	0.90	1.00	0.70	0.60	0.50
P3	0.10	0.70	1.00	0.40	0.30
P4	0.65	0.60	0.40	1.00	0.80
P5	0.20	0.50	0.30	0.80	1.00

2 + 10 = 12

Cognition Level	*LOCQ*	*IOCQ*	*HOCQ*
Percentage distribution	*59.37*	*2292*	*17.71*

Course Outcome (CO):

After the completion of the course students will be able to

MCAP2251.1 : Describe basic concept of data mining and related models.
MCAP2251.2 : Design the kinds of patterns using association rule mining.
MCAP2251.3 : Explore data analysis by dimensionality reduction as well as information compression using PCA.
MCAP2251.4 : Deploy appropriate classification techniques to solving 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.*