DATA MINING AND KNOWLEDGE DISCOVERY
(MCA1233)

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) Frequency of occurrence of an item set is called as
(a) Support (b) Confidence (c) Support count (d) Rules

(ii) Which statements are correct about the Principal Component Analysis?
(a) PCA is a supervised learning algorithm
(b) PCA is a unsupervised learning algorithm
(c) PCA is a Clustering algorithm
(d) PCA is a dimensionality reduction technique

(iii) What are two steps of pruning in decision tree?
(a) Pessimistic pruning and Optimistic pruning
(b) Post pruning and pre pruning
(c) Cost complexity pruning and time complexity pruning
(d) High pruning and low pruning

(iv) What is the formula for Bayes' theorem? Where (A & B) and (H & E) are events and P(B), P(H) & P(E) ≠ 0.
(a) P(H|E) = [P(E|H) * P(E)] / P(H) (b) P(A|B) = [P(A|B) * P(A)] / P(B)
(c) P(H|E) = [P(H|E) * P(H)] / P(E) (d) P(A|B) = [P(B|A) * P(A)] / P(B)

(v) 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

(vi) 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

(vii) What is true about single layer associative neural networks?
(a) Performs pattern recognition
(b) Can find the parity of a picture
(c) Can determine whether two or more shapes in a picture are connected or not
(d) None of the mentioned

(viii) Which is conclusively produced by Hierarchical Clustering?
(a) Final estimation of cluster centroids
(b) Tree showing how nearby things are to each other
(c) Assignment of each point to clusters
(d) all of these

(ix) Which of the following is/are valid iterative strategies for treating missing values before clustering analysis?
(a) Imputation with mean
(b) Nearest Neighbour assignment
(c) Imputation with Expectation Maximization algorithm
(d) All of the above

(x) Which clustering algorithm can be used to identify overlapping clusters in the data?
(a) K-Means (b) DBSCAN (c) Agglomerative (d) K-means++

Fill in the blanks with the correct word

(xi) A collection of one or more items is called as_____ .

(xii) The FP growth algorithm has _____ phases.

(xiii) _____ network models capture both conditionally dependent and conditionally independent relationships between random variables.

(xiv) _____errors are the expected errors generated by a model because of unknown records.

(xv) To remove sub-nodes of a decision node, _____is used.

Group – B

2. (a) Briefly describe the different steps with relevant example for PCA analysis.
[(CO2)(Understand/LOCQ)]

(b) A database with 5 transactions are given below.

Transactions	Items
T1	X,C,F,Y,G,P,Z
T2	C,B,X,O,F,L,Z
T3	O,B,F,H
T4	B,K,C,P
T5	Z,N,X,F,C,L.P

Draw a FP tree for the above data set for which minimum support count value=3.
[(CO3) (Apply/IOCQ)]
(3 + 3) + 6 = 12

3. (a) Illustrate the benefits of Market Basket Analysis(MBA) with some example. Describe different measures of Market Basket Analysis. *[(CO2) (Understand/LOCQ)]*

(b) Consider the following Table:

Transaction	T1	T2	T3	T4	T5	T6	T7	T8	T9
List of Item_Id	I2, I4,I5	I1,I2,	I2,I3	I2,I4	I1,I3	I1,I2,I3	I1,I3	I1,I2,I3,I5	I1,I2,I3

Generate all frequent item sets by using Aproiri Algorithm where the minimum support Count =2. *[(CO2) (Create/HOCQ)]*

4 + 8 = 12

Group - C

4. (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**:**

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). *[(CO4,C06)(Create/HOCQ)]*

3 + 9 = 12

5. (a) What is the difference between Bayes classifier and Naive Bayes classifier? *[(CO4)(Analyse/IOCQ)]*

(b) What is Bayes' theorem? Given the data set in table(Question 12), predict using Naïve Bayes classifier, whether a customer with Gender = Female, Age > 25 and Salary = Medium will purchase or not. *[(CO4)(Apply/LOCQ)]*

(c) Define the decision error in the context of Bayes classifier for the two class problem. *[(CO4) (Analyse/HOCQ)]*

2 + 7 + 3 = 12

Group - D

6. (a) Can we apply the Kernel-trick in Logistic regression? Why it is not used in practice? *[(CO4)(Analyse/IOCQ)]*

(b) Explain the derivation of weight update equation of a feed forward Multi Layered Perceptron model(i.e., Neural network) , based on back propagation learning algorithm. *[(CO4)(Remember/LOCQ)]*

6 + 6 = 12

7. (a) Construct the Lagrangian for the primal optimization problem in finding the support vectors for a two-class linearly separable classification problem. *[(CO4)(Analyse/HOCQ)]*

(b) Illustrate the following terms with diagram.
(i) Support vectors. (ii) Marginal Planes. (ii) Linear and non-linear separable *[(CO4)(Remember/LOCQ)]*

6 + (2 + 2 + 2) = 12

Group - E

8. (a) Consider the data points provided in the table below. Perform hierarchical clustering considering complete link method (MAX distance) to generate a cover.

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)]

(b) Try to approximately plot them on a 2D plane and show the nested clusters. Also show the dendrogram with merging distance on Y-axis. *[(CO5,CO6)(Apply/IOCQ)]*

7 + 5 = 12

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

(b) Describe the DBSCAN Algorithm. *[(CO5)(Remember/LOCQ)]*

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

(d) 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 + 2 + 3 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	40.63	32.29	27.08