DATA MINING AND KNOWLEDGE DISCOVERY
(CSEN 4144)

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) _____ is generally used to analyse unlabelled data set.
(a) Supervised learning (b) Unsupervised Learning
(c) Reinforcement Learning (d) None of these

(ii) Which of the following is not a decision tree node?
(a) Root node (b) Internal node
(c) Predicted node (d) Leaf node.

(iii) Which of the following is a unsupervised learning technique?
(a) k-means (b) Decision Tree
(c) Association Rule (d) Naive Bayes.

(iv) Which of the following is used in Rule based classification?
(a) Agglomerative clustering (b) k-means
(c) PRISM (d) DBSCAN.

(v) If T consist of 500000 transactions, 20000 transaction contain bread, 30000 transaction contain jam, 10000 transaction contain both bread and jam. Then the confidence of buying bread with jam is
(a) 33.33% (b) 66.66% (c) 45% (d) 50%.

(vi) _____ is not an example of ensemble learning algorithm?
(a) Random Forest (b) Boosting
(c) Bagging (d) SVM

(vii) Agglomerative clustering is a _____ technique.
(a) partitional clustering
(b) bottom up hierarchical clustering
(c) top-down hierarchical clustering
(d) classification

(viii) A collection of interesting and useful patterns in database is called _____
(a) Knowledge (b) Noise
(c) Data (d) Algorithm.

(ix) Association rules are always defined on_____
(a) binary attribute (b) single attribute
(c) relational database (d) multidimensional attributes.

(x) Which of the following can act as the best possible termination condition in K-Means clustering algorithm?
(a) For a fixed number of iterations
(b) Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum
(c) Means of cluster changes frequently between successive iterations
(d) Number of clusters is gradually getting than k number of initial clusters.

Group – B

2. (a) Define Information gain and gain in the Gini index.

(b) Consider the following data set for a binary class problem.

Sl No.	Color	Size	Act	Age	Inflated
1	Yellow	Small	Stretch	Child	T
2	Yellow	Small	Stretch	Child	T
3	Yellow	Small	Stretch	Child	T
4	Yellow	Small	Stretch	Child	T
5	Yellow	Small	Stretch	Adult	T
6	Yellow	Small	Stretch	Child	F
7	Purple	Large	Dip	Adult	F
8	Purple	Large	Dip	Child	F
9	Purple	Small	Stretch	Adult	T
10	Purple	Small	Stretch	Child	F
11	Purple	Small	Dip	Adult	T
12	Purple	Small	Dip	Child	T
13	Purple	Large	Stretch	Adult	F
14	Purple	Large	Stretch	Child	F
15	Purple	Large	Dip	Adult	F
16	Purple	Large	Dip	Child	T

Calculate the information gain when splitting on different attributes. Which attribute would the decision tree induction algorithm choose?

2 + 10 = 12

3. Write short notes on any 3 (three) of the followings: **(3 × 4) = 12**
(i) Supervised and Unsupervised Learning
(ii) Pruning in Decision Tree

9. Perform K-means clustering on two dimensional data points as given in the following table, where, K = 3. Randomly generate the initial centroids and perform the algorithm for four iterations. Show the movement of centroids and clusters in each iteration, by displaying clusters on the X and Y co-ordinates.

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

12

(iii) Precision and Recall
(iv) Mercer's Condition.

Group – C

4. (a) Explain how to compute and maximize the margin in SVM?

   (b) What is the use of kernel function in Support Vector Machine? Discuss about the notion of dual problem in the context of SVM.

5 + (2 + 5) = 12

5. (a) Briefly explain the Bayes theorem in the context of classification.

Body Temperature	Species Name	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
Warm	Human	yes	no	no	yes	no	Mammal
Cold	Python	no	no	no	no	yes	Reptile
Warm	Whale	yes	yes	no	no	no	Mammal
Cold	Frog	no	semi	no	yes	yes	Amphibian
Cold	Komodo	no	no	no	yes	no	Reptile
Warm	Bat	yes	no	yes	yes	yes	Mammal
Warm	Pigeon	no	no	yes	yes	no	Bird
Warm	Cat	yes	no	no	yes	no	Mammal
Cold	Leopard	yes	yes	no	no	no	Fish
Cold	Turtle	no	semi	no	yes	no	Reptile
Warm	Penguin	no	semi	no	yes	no	Bird
Warm	Porcupine	yes	no	no	yes	yes	Mammal
Cold	Eel	no	yes	no	no	no	Fish
Cold	Salamander	no	semi	no	yes	yes	Amphibian

   (b) Consider the following dataset of species classification table:
   Using Naive Bayes Classifier on the above data set of species classification, find the class label of the species called Salmon having following attribute values:

Body Temperature	Species Name	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates
Cold	Salmon	no	yes	no	no	no

4 + 8 = 12

Group – D

6. An example of a market basket dataset is given by dataset in the following table.

Customer ID	Transaction ID	Items Bought
1	0001	{a. d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}

3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

(a) Considering each customer_id as a market basket, compute the support for itemsets for the above dataset
(i) {e}, {b, d} (ii) {b, d, e}

(b) Considering each customer_id as a market basket compute the confidence for the association rules
(i) {b, d} → {e} (ii) {e} → {b, d}

(c) Suppose s1 and c1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s2 and c2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between the followings:
(i) s1 and s2 (ii) c1 and c2.

3 + 3 + 6 = 12

7. (a) Explain Random Forest learning method in the context of ensemble learning.

(b) Explain the steps of Frequent Pattern (FP) Growth Method with an example.

7 + 5 = 12

Group – E

8. (a) Define, with example, Core point, Border point and Noise point in the perspective of DBSCAN clustering algorithm.

(b) Describe the DBSCAN Algorithm.

(c) Describe the process of selecting the parameters Eps (radius that defines the neighbourhood of a point) and MinPts (minimum number of points in the neighbourhood of the core point) in DBSCAN.

(d) Explain why DBSCAN does not work well for the data having varying density.

3 + 3 + 4 + 2 = 12