(vi) The complexity of the k-means algorithm would depend on
(a) only K, number of iterations
(b) only Number of points, K
(c) only number of iterations
(d) all of the above combined dimensions.

(vii) DBSCAN uses k-nearest neighbour distance to find the parameter
(a) Eps (radius) (b) MinPts
(c) Core points (d) Noise points.

(viii) Support Vector Machine can be used to classify
(a) linearly separable data only
(b) non-linearly separable data only
(c) both linearly and non-linearly separable data
(d) none of the above.

(ix) Which of the following is finally produced by Hierarchical Clustering?
(a) Final estimate of cluster centroids
(b) Tree showing how close things are to each other
(c) Assignment of each point to clusters
(d) All of the mentioned.

(x) When a rule-set is both mutually exclusive as well as exhaustive, some instance in the training dataset may be covered by,
(a) more than one rule (b) exactly one rule
(c) no rule (d) none of the above.

Group – B

2. (a) Define Information gain.

(b) Construct (induct) a decision tree using information gain from the data provided in the table 1. Consider the Gender as the class label.

Table 1

Sl. No.	Over 170CM	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

3 + 9 = 12

3. Consider the diagram of Fig.1. Justify with reason, which rule a R1, R2 and R3 is the best according to the following measures?
(i) Likelihood ratio statistic
(ii) Laplace
(iii) m-estimate.

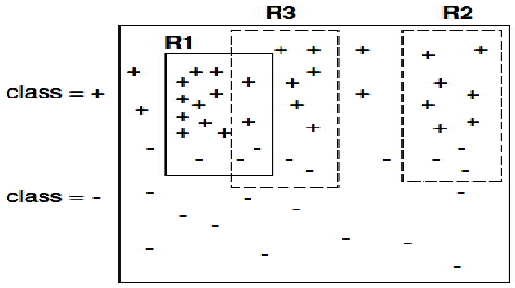

Fig. 1

(3 × 4

Group – C

4. (a) Suppose a support vector machine for separating pluses from mi finds a plus support vector at the point x1 = (1, 0), a minus su vector at x2 = (0, 1). You are to determine values for the classifi vector w and the threshold value b.

(b) Construct the Lagrangian for the primal optimization probl finding the support vectors for a two-class linearly sepa classification problem.

5 + 7

5 Given this dataset of Table 2, can you predict using Naïve Bayes cla whether a Red SUV from Domestic makers will be stolen or not? Use the m-estimate method with m = 3, p = 0.5.

Table 2

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

Group – D

6. (a) Draw the FP-Growth Tree for the following transaction dataset of Table 3.

   (b) Draw the prefix paths ending with *..-c-e* and *..-d-e*.

Table 3

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

8 + 4 = 12

7. (a) Prove that the total number of possible rules extracted from a dataset that contains d items is, $R = 3^d - 2^{d+1} + 1$.

   (b) With an example, briefly explain apriori algorithm.

9 + 3 = 12

Group – E

8. Perform hierarchical clustering method MAX (complete link) on the dataset provided in Table 4 to generate a cover. Try to approximately plot them on a 2D plane and show the nested clusters. Also show the dendrogram with merging distance on Y-axis.

Table 4

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

9. Perform K-means clustering on the dataset provided in Table 4, where K = 3. Randomly generate the initial centroids and perform the algorithm for up to a maximum of four iterations. Show the movement of the centroids and the clusters (for every iteration) by drawing the points and the clusters on the X and Y co-ordinates. Show all the calculations clearly.

12

DATA MINING AND KNOWLEDGE DISCOVERY
(CSEN 4144)

Time Allotted : 3 hrs **Full Mark**

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 practica

Group – A
(Multiple Choice Type Questions)

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

   (i) Cluster is
   (a) group of similar objects that differ significantly from other ob
   (b) operations on a database to transform or simplify data in or prepare it for a machine-learning algorithm
   (c) symbolic representation of facts or ideas from which inforn can potentially be extracted
   (d) none of these.

   (ii) You are given data about seismic activity in Japan, and you w predict a magnitude of the next earthquake, this is an example of
   (a) Dimensionality Reduction
   (b) Supervised Learning
   (c) Unsupervised Learning
   (d) Reinforcement Learning.

   (iii) In a picture, where 7 cats and 10 dogs are present, your dog det algorithm has detected 9 entities, out of which only 6 are dog remaining are cats. What is the precision of your algorithm?
   (a) 0.6 (b) 0.66 (c) 6/17 (d

   (iv) The goal in Naïve Bayes classifier is to predict class label using,
   (a) posterior probability (b) prior probability
   (c) likelihood (d) evidence.

   (v) A lending company wants to estimate the loan amount for a cus who has applied for a possible loan, this is an example of?
   (a) Clustering (b) Classification
   (c) Prediction (d) Association Rule.