9. A linearly separable dataset is given in Table 4. Predict the class of (0.6, 0.8) using a support vector machine classifier.

x_1	x_2	y	Lagrange Multiplier
0.3858	0.4687	1	65.5261
0.4871	0.611	–1	65.5261
0.9218	0.4103	–1	0
0.7382	0.8936	–1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	–1	0
0.2146	0.0099	1	0

Table 4: Linearly separable dataset

12

DATA MINING AND KNOWLEDGE DISCOVERY
(CSEN 5237)

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) Decision trees are appropriate for the problems where
(a) attributes are both numeric and nominal
(b) target function takes on a discrete number of values
(c) data may have errors
(d) all of the mentioned.

(ii) Which of the following is a predictive model?
(a) clustering (b) regression
(c) summarization (d) association Rule.

(iii) Extreme values that occur infrequently are called as
(a) outliers (b) rare values
(c) dimensionality reduction (d) all of the above.

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

(v) Clustering is considered to be
(a) unsupervised learning (b) supervised learning
(c) semi-Supervised learning (d) reinforcement learning.

(vi) K-means clustering suffers from
(a) bad initialization of centroids (b) bad selection of K
(c) selection of only round shaped clusters (d) all of the above.

(vii) Pick out the hierarchical clustering algorithm
(a) DBSCAN (b) BIRCH
(c) PAM (d) CURE.

(viii) 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.

(ix) Support vectors can be identified by
(a) zero value Lagrangian multipliers (b) class labels
(c) non-zero Lagrangian multipliers (d) proximity to (0, 0).

(x) Average of all shortest path distances for all pair of vertices in a social network is usually around 6. This feature is known as
(a) scale free feature (b) randomness feature
(c) small world feature (d) none of the above.

Group - B

2.(a) Draw a decision tree to predict whether a student will be accepted in the post-graduate program using the data provided in Table 1**.**

ID	GATE qualified	Publications	Written Test qualified	Interview performance	Decision
1	Yes	Yes	No	Bad	Reject
2	No	Yes	Yes	Bad	Reject
3	Yes	No	No	Good	Accept
4	No	No	Yes	Bad	Reject
5	No	Yes	No	Bad	Reject
6	Yes	No	Yes	Good	Accept
7	No	Yes	Yes	Good	Accept
8	Yes	Yes	No	Good	Accept
9	No	No	Yes	Good	Reject
10	No	Yes	No	Bad	Reject
11	No	No	No	Good	Reject
12	No	Yes	No	Good	Accept
13	Yes	Yes	Yes	Bad	Accept
14	Yes	No	No	Bad	Reject
15	Yes	No	Yes	Bad	Accept

Table 1. Decision data for 15 candidates, who applied for post-graduate program.

12

	pepperoni	*pineapple*	*pickledOnion*	*liked*
A	true	true	true	false
B	true	false	false	true
C	false	true	true	false
D	false	true	false	true
E	true	false	false	true

Using Hamming distance throughout, show how the 3-NN classifier with majority voting would classify
{pepperoni = false, pineapple = true, pickledOnion = true}.

6 + 6 = 12

7. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,
*R*1: *A* → + (covers 4 positive and 1 negative examples),
*R*2: *B* → + (covers 30 positive and 10 negative examples),
*R*3: *C* → + (covers 100 positive and 90 negative examples),

determine which is the best and worst candidate rule according to:
(a) Rule accuracy (b) FOIL's information gain and (c) Likelihood ratio statistic.

(3 × 4) = 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	9
p6	7	2
p7	3	8
p8	4	10
p9	8	1
p10	9	3

Describe the major drawbacks of K-means algorithm for clustering.

(8 + 4) = 12

3.(a) Define the mathematical model for Naïve Bayes Classifier.

(b) The following table provides the data of a set of officers. Use Naïve Bayes classifier to classify an officer's gender who is blue-eyed, over 170cm tall and has long hair.

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

4 + 8 = 12

Group – C

4.(a) Define Information gain, Gain Ratio and Gini Index.

(b) Consider the following set of training examples:

Instance	Classification	A1	A2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

What are the information gains of A1 and A2 relative to these training examples? Provide the equation for calculating the information gain as well as the intermediate results.

5 + 7 = 12

5. Consider the data set shown in Table 3.

(i) Compute the support for itemsets *{e}*, *{b, d}*, and *{b, d, e}* by treating each transaction ID as a market basket.

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	$\{a, b, e\}$

Table 3: A market basket dataset

(ii) Use the results in part (a) to compute the confidence for the association rules *{b, d}* → *{e}* and *{e}* → *{b, d}*. Is confidence a symmetric measure?

(iii) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

(iv) Use the results in part (c) to compute the confidence for the association rules *{b, d}* → *{e}* and *{e}* → *{b, d}*.

(4 × 3) = 12

Group – D

6.(a) Draw an FP-growth Tree from the data provided in Table 3 as given in question No.5.

(b) The Pants Pizza Parlour sells pizzas with optional toppings: pepperoni, pineapple and pickled onion. Every day this week you have tried a pizza (A to E) and kept a record of which you liked: