7. (a) Derive the Lagrangian for the optimization problem (primal) as defined by linear SVM – separable case.

(b) A linearly separable dataset is given in the table below. Predict the class of (0.6, 0.8) using a support vector machine classifier.

X_1	X_2	Y	Lagrange Multiplier
0.3	0.4	+1	5
0.7	0.6	-1	8
0.9	0.5	-1	0
0.7	0.9	-1	0
0.1	0.05	+1	0
0.4	0.3	+1	0
0.9	0.8	-1	0
0.2	0.01	+1	0

6 + 6 = 12

Group – E

8. (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) Explain apriori principle briefly with an example.

9 + 3 =12

9. (a) 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

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

8 + 4= 12

M.TECH/CSE/2ND SEM/CSEN 5237/2018

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) KDD describes the ________.
(a) extraction of data
(b) extraction of information
(c) extraction of rules
(d) whole process of extraction of knowledge from data.

(ii) For a predictive model that needs to be evaluated by the accuracy of the model's performance as well as the ground truth, which of the following should be used as a performance measure?
(a) Accuracy (c) Recall (b) Precision (d) F-measure.

(iii) In a picture, where 7 cats and 10 dogs are present, your dog detection algorithm has detected 9 entities, out of which only 6 are dogs and remaining are cats. What is the precision of your algorithm?
(a) 0.6 (b) 0.66 (c) 6/17 (d) 0.9.

(iv) In Naïve Bayes classifiers, __________ method can provide a bias from a size of m samples with p probability.
(a) Laplacian (c) Lagrangian (b) m-estimate (d) validation.

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

(vi) If T consists of 500 transactions, 20 transaction contain bread, 30 transactions contain jam, 10 transactions contain both bread and jam. Then the support of buying bread and jam is
(a) 2% (b) 20% (c) 3% (d) 30%.

(vii) DBSCAN uses k-nearest neighbour distance to find the parameter —
(a) eps (radius) (b) minPts (c) core points (d) noise points.

(viii) The K-Means algorithm terminates when

(a) a user-defined minimum value for the summation of squared error differences between instances and their corresponding cluster center is seen.

(b) the cluster centers for the current iteration are identical to the cluster centers for the previous iteration.

(c) the number of instances in each cluster for the current iteration is identical to the number of instances in each cluster of the previous iteration.

(d) the number of clusters formed for the current iteration is identical to the number of clusters formed in the previous iteration.

(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) After SVM learning, each Lagrange multiplier α_i takes either zero or non-zero value. What does it indicate in each situation?

(a) A non-zero α_i indicates the data point i is a support vector, meaning it touches the margin boundary.

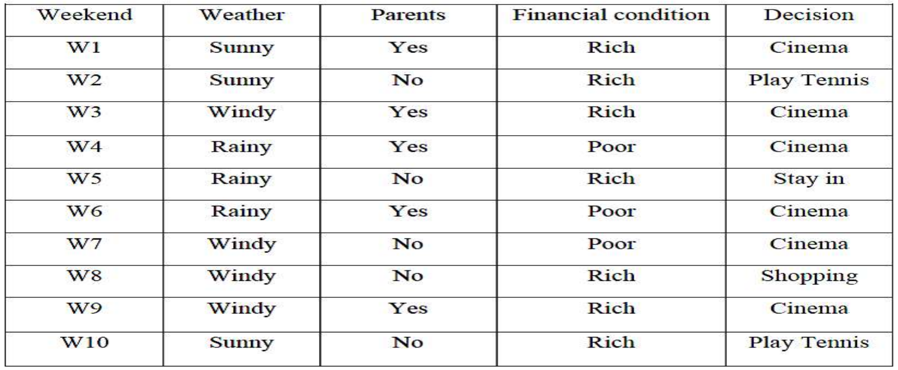
(b) A non-zero α_i indicates that the learning has not yet converged to a global minimum.

(c) A zero α_i indicates that the data point i has become a support vector datapoint, on the margin.

(d) A zero α_i indicates that the learning process has identified support for vector i.

Group – B

2. Create a decision tree by using the given dataset (Table no. 1) that describes what a set of people might decide to do on weekend based on a set of attributes that characterizes the weekends. Here, the weekends are described by the attributes Weather, Parents and Financial Condition. Use entropy as the impurity measure while creating the decision tree.



Weekend	Weather	Parents	Financial condition	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Play Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Play Tennis

Table 1. Decision data for weekend activities over 10 different weekends.

12

3. (a) Define support and confidence in mining frequent pattern.

(b) Are these measures symmetric? Justify your answer.

(c) Please review the following sales data from a small grocery store. The data below shows eight shopping carts (baskets) containing different products (A, B, C, etc.) that customers checked out.

CART-1(A,C,E,F). CART-2(A,F,E), CART-3(C,F), CART-4(A,B,C), CART-5(C,E,F). CART-6(F) CART-7(B,E) CART-8(A,B).

The store manager conducted market basket analysis using the above data and is now considering one of the following two rules to help the store cross-sell products and increase sales. That is, when a customer buys a product, the store would like to recommend to the customer another product that she/he would be most likely to buy as well.

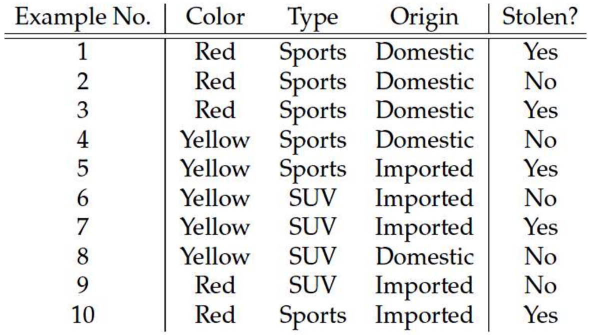
Rule 1: F -> E; Rule 2: E -> F

On the basis of given data, which of the above two rules would be a better predictor of cross-sale than the other? Please justify your answer with *support* and *confidence*.

2 + 2 + 8 = 12

Group – C

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



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

Table 2 **12**

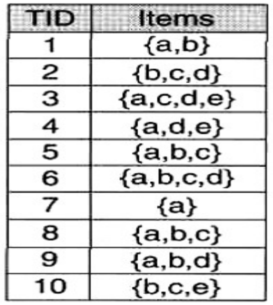
5. (a) Define Information gain, Gain Ratio and Gini Index.

(b) Consider the data provided in Table 2 as set of training examples. What are the information gains of color, type and origin relative to the training examples? Provide the equation for calculating the information gain as well as the intermediate results.

3 + 9 = 12

Group – D

6. Draw the FP-Growth Tree for the following transaction dataset (Table 3). Draw the prefix paths ending with ce and de.



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}

Table 3 **(8 + 4) =12**