DATA MINING & KNOWLEDGE DISCOVERY
(CSEN 3132)

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

(ii) To detect fraudulent usage of credit cards, the following data mining task should be used (Select one):
(a) outlier analysis (b) prediction
(c) association analysis (d) feature selection

(iii) In decision tree, an attribute is selected as root node, where
(a) gain ratio is minimum (b) information gain is maximum
(c) information gain is minimum (d) none of these

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

(v) In SVM, when the C parameter is set to infinite, which of the following holds true?
(a) The optimal hyperplane if exists, will be the one that completely separates the data
(b) The soft-margin classifier will separate the data
(c) Both (a) and (b) are true
(d) None of the above.

(vi) Frequency of occurrence of an itemset is called as
(a) Support (b) Confidence (c) Support Count (d) Rules.

(vii) Boosting is said to be a good classifier because
(a) it creates all ensemble members in parallel, so their diversity can be boosted
(b) it attempts to minimize the margin distribution
(c) it attempts to maximize the margins on the training data
(d) none of the above.

(viii) In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual (Tk) tree in Random Forest?
A. Individual tree is built on a subset of the features
B. Individual tree is built on all the features
C. Individual tree is built on a subset of observations
D. Individual tree is built on full set of observations
(a) A and C (b) A and D (c) B and C (d) B and D.

(ix) Which among the following three properties is/are not satisfied by distance measure?
(a) Symmetry (b) Transitivity
(c) Triangular Inequality (d) Reflexive.

(x) 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.

Fill in the blanks with the correct word

(xi) A lending company wants to estimate the loan amount for a customer who has applied for a possible loan, this is an example of _____.

(xii) The binary entropy for a random binary variable with probability p is maximum when p = _____.

(xiii) "A symmetric matrix is positive definite if all its Eigen values are_____."

(xiv) _____ algorithm mines all frequent patterns through pruning rules with lesser support.

(xv) A good clustering method will produce good quality clusters with _____ intra class similarity.

Group - B

2. (a) Define coverage and accuracy in assessing the rules in rule based classification. *[(CO1)(Remember/LOCQ)]*

(b) What are the issues in the rule based classification? Write the conflict resolution strategies, in detail, to overcome these issues. *[(CO2)(Understand/LOCQ),(CO3) (Analyse/LOCQ)]*

(c) What is confusion matrix? Define Precision and Recall. Explain, in brief, the importance of these two measures to evaluate the performance of a classification model. *[(CO4)(Analyse/HOCQ)]*

2 + 5 + 5 = 12

3. (a) Create a decision tree by using the following dataset 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

[(CSEN3132.3)(Apply/IOCQ)]

(b) What are the differences between overfitting and underfitting? *[(CSEN3132.2)(Understand/LOCQ)]*

10 + 2 = 12

Group - C

4. (a) Why naïve Bayesian classification is called naïve? Briefly outline the major ideas of naïve Bayesian classification. *[(CO1)(Remember/LOCQ)]*

(b) Consider the following dataset of species classification table:

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

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

[(CO3)(Apply/IOCQ)]

(1 + 3) + 8 = 12

5. (a) Construct the Lagrangian for the primal optimization problem in finding the support vectors for a two-class linearly separable classification problem. *[(CO2)(Understand/IOCQ)]*

(b) A linearly separable dataset is given in the following table. 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
1.0	0.6	-1	0
0.8	0.9	-1	0
0.1	0.2	+1	0
0.3	0.3	+1	0
0.9	0.8	-1	0
0.3	0.1	+1	0

[(CO3) (Apply/LOCQ)]

8 + 4 = 12

Group - D

6. (a) Construct the FP-tree for the transaction database provided below and find all frequent item-sets using FP-growth approach.

Transaction ID	List of Items
1	A, B, D
2	A, B, C
3	B, F
4	A, D
5	B, C
6	A, B, D, E
7	A, B, D, F
8	A, C, E
9	A, B, F
10	A, C, E, F

[(CO2)(Apply/LOCQ)]

(b) Describe, in detail, the random forest algorithm for classification. *[(CO4)(Understand/IOCQ)]*

6 + 6 = 12

7. (a) Write the drawbacks of Apriori algorithm in finding the frequent itemsets. Also write the remedies in improving the efficiency of Apriori. *[(CO5, Analysis/HOCQ)]*

(b) Prove that the total number of possible rules extracted from a market basket dataset that contains d unique items is,

$$R = 3^d - 2^{d+1} + 1$$

[(CO6,Analysis/IOCQ]

(2 + 4) + 6 = 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

[(CO2)(Describe/LOCQ)]

(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. *[(CO3)(Apply/IOCQ)]*

7 + (2 + 3) = 12

9. (a) Apply K-means clustering algorithm on all the points given in the following table, where K=2. Randomly select the initial seeds and show the steps 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

[(CSEN3132.3)(Apply/IOCQ)]

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

[(CSEN3132.1)(Remember/LOCQ)]

9 + 3 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	34	54	12