DATA MINING & KNOWLEDGE DISCOVERY
(CSE3132)

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) Chandrayan-3 has sent you the data about seismic activity in the moon, and you want to predict a magnitude of the next moonquake, this is an example of?
(a) Dimensionality Reduction (b) Supervised Learning
(c) Unsupervised Learning (d) Reinforcement Learning

(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) Suppose that X_1, ..., X_m are categorical input attributes and Y is categorical output attribute. Suppose we plan to learn a decision tree without pruning, using the standard algorithm. The maximum depth of the decision tree must be
(a) less than m+1 (b) greater than m+1
(c) both (a) and (b) can be true (d) None of (a) and (b) are true

(iv) **Statement I:** "A nonlinearly separable training set in a given feature space can always be made linearly-separable in another space."
Statement II: "Using the kernel trick, one can get non-linear decision boundaries using algorithms designed originally for linear models."
(a) Both the statements are FALSE. (b) Statement I is FALSE but Statement II is TRUE.
(c) Statement I is TRUE but Statement II is FALSE (d) Both the statements are TRUE.

(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) 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

(vii) If a transaction set consist of 1000 transactions, 300 transactions contain bread, 350 transactions contain butter, 150 transactions contain both bread and butter. Then the confidence of buying bread with butter (butter ⇒ bread) is
(a) 30% (b) 42.86% (c) 50% (d) 65%

(viii) In Random forest you can generate hundreds of trees (say T_1, T_2 .....T_n) and then aggregate the results of these tree. Which of the following is true about individual (T_k) 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) 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 these

Fill in the blanks with the correct word

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

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

(xiii) The goal in Bayes' classifier is to predict class label using _____ probability.

(xiv) An itemset whose support is greater than or equal to a minimum support threshold is _____.

(xv) Steps in Apriori algorithm are _____ and _____.

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,CO3)(Understand,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) Define Gini Index and gain in Gini index. *[(CO1 & 2)(Remember & Understand/LOCQ)]*

(b) Construct (induct) a decision tree using gain in Gini index from the data provided in the following table. Consider the Gender as the class label. *[(CO4)(Apply/IOCQ)]*

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

2 + 10 = 12

Group - C

4. (a) Suppose a support vector machine for separating pluses from minuses finds a plus support vector at the point $x_1 = (1, 0)$, a minus support vector at $x_2 = (0, 1)$. You are to determine values for the classification vector w and the threshold value b. *[(CO4),Understand/IOCQ)]*

(b) Sometimes data is just nonlinearly separable or data has errors and one wants to ignore them to obtain a better solution. In fact, this is achieved by relaxing the margin, in other words, using a soft margin. Derive the Lagrangian for the optimization problem as defined by linear SVM – soft margin classification. *[(CO1),Remember/LOCQ]]*

3 + 9 = 12

5. (a) Consider the following dataset 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

Use Naïve Bayes' classifier to predict the decision for predicting Inflated having Color = Yellow, Size = Large, Act = Stretch and Age = Adult. *[(CO3)(Apply/IOCQ)]*

(b) In which cases Naïve Bayes classifier may provide indecisive results (hint: posterior probability may become zero)? How can we use Naïve Bayes classifier in those circumstances? *[(CO4)(Analyse/IOCQ)]*

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 considering minimum support count as 2.

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) Construct the frequent itemsets considering the minimum confidence value as 2 and show at least 6 association rules. *[(CO4)(Understand/IOCQ)]*

7 + 5 = 12

7. (a) Define support and confidence in mining frequent pattern. *[(CO1),(Remember/LOCQ)]*

(b) Are these measures symmetric? Justify your answer. *[(CO5)(Understand/IOCQ)]*

(c) Please review the following sales data from a small grocery store. The data in the diagram below shows 8 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), CART4(A,B,C), CART 5(C,E,F). CART 6(B,F) CART 7(B,E) CART 8(A,B).
Construct the FP-tree for the given transaction database and find all frequent item-sets using FP-growth approach considering 2 as the minimum support count. *[(CO4 & 5)(Apply/HOCQ)]*

2 + 2 + 8 = 12

Group - E

8. (a) Consider the data points provided in the table below. Perform hierarchical clustering considering single linkage method (MAX distance) to generate a cover.

Points	X coordinate	Y coordinate
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) Discuss differences between Clustering and Classification. *[(CO3)Understand/IOCQ)]*

(b) Group the following data points using k-means clustering technique, where k=3 and each data point represented in the form of (x_coordinate, y_coordinate). Consider A1, B1, C1 as the initial cluster centres. **Data Points:** A1(2,10); A2(2, 5); A3(8,4); B1(5, 8); B2(7, 5); B3(6,4); C1(1,2); C2(4,9) *[(CO3 & 6)(Remember and Apply/LOCQ)]*

4 + 8 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	44	42	14