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) Data Mining is
(a) the actual discovery phase of a knowledge discovery process
(b) the stage of selecting the right data for a Knowledge Discovery process
(c) a subject-oriented integrated time variant non-volatile collection of data in support of management
(d) none of the above.

(ii) 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%.

(iii) Bayesian classifier is
(a) a class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory
(b) any mechanism employed by a learning system to constrain the search space of a hypothesis
(c) an approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation
(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 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.

(vi) When you find noise in data which of the following option would you consider in the implementation of k-NN classifier?
(a) I will increase the value of k (b) I will decrease the value of k
(c) Noise cannot be dependent on value of k (d) None of these.

(vii) 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.

(viii) Statement I: "A non linearly-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.

(ix) Which of the following is required by K-means clustering?
(a) Defined distance metric (b) Number of clusters
(c) Initial guess as to cluster centroids (d) All of the above.

(x) 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.

Fill in the blanks with the correct word

(xi) Clustering can be categorised as a problem of ______________ learning.

(xii) Let X_1, ..., X_m be the categorical input attributes and Y be the 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 _________.

(xiii) DBSCAN cannot be used (with high accuracy) for datasets that are ____________.

(xiv) Data _________ refers to the step of the knowledge discovery process, in which the several data sources are combined.

(xv) In Support Vector Machine, a __________ Lagrange multiplier (α_i) indicates the data point *i* is a support vector, meaning it touches the margin boundary.

Group - B

2. (a) Write the decision tree algorithm for classification. *[(CO1)(Remember/LOCQ)]*

(b) For the dataset given below (Table 1), find the most suitable attribute to be selected for the root node by the attribute selection method using information gain measure.

Table 1

Sl No	Gender	Age	Salary	Purchased (Class)
1	Male	<=25	Low	No
2	Female	> 25	Low	No
3	Male	> 25	Medium	Yes
4	Female	> 25	High	Yes
5	Male	<=25	Medium	No
6	Female	> 25	High	Yes
7	Male	<=25	High	Yes
8	Female	<=25	Medium	Yes
9	Male	> 25	Low	No
10	Female	> 25	High	Yes
11	Male	> 25	Low	Yes
12	Female	<=25	Medium	No

[(CO3)(Apply/IOCQ)]

4 + 8 = 12

3. (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

Group - C

4. (a) What is the difference between Bayes classifier and Naive Bayes classifier? *[(CO2)(Understand/IOCQ)]*

(b) What is Bayes' theorem? Given the dataset in table 1 (Question 2), predict using Naïve Bayes classifier, whether a customer with Gender = Female, Age > 25 and Salary = Medium will purchase or not. *[(CO3)(Apply/LOCQ)]*

(c) Define the decision error in the context of Bayes classifier for the two class problem. *[(CO5)(Analyse/HOCQ)]*

2 + 7 + 3 = 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 below (Table 2). Predict the class of (0.6, 0.8) using a support vector machine classifier.

Table 2

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

(b) You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity we assign the meal items short names (M1 – M5) rather than the full descriptive names (e.g., Big Mac).

Table 3

Meal Item	List of Item
Order: 1	M1, M2, M5
Order: 2	M2, M3
Order: 3	M2, M4
Order: 4	M1, M3
Order: 5	M2, M3
Order: 6	M1, M2, M3, M5
Order: 7	M1, M2, M4
Order: 8	M1, M2, M3
Order: 9	M1, M3

(i) Compute the support for item-sets {M5}, {M2, M4} and {M2, M4, M5} by treating each transaction ID as a market basket. *[(CO2)(Apply/LOCQ)]*

(ii) Use the results in part (b) to compute the confidence for the association rules {M2, M4} → {M5} and {M5} → { M2, M4}. *[(CO2)(Apply/LOCQ)]*

(c) Prove that the total number (R) of possible rules extracted from a dataset that contains d items is, $R = 3^{\mathrm{d}} - 2^{\mathrm{d}+1} + 1$ *[(CO5)(Analyse/HOCQ)]*

3 + (3 + 3) + 3 = 12

7. (a) Construct the FP-tree for the transaction database provided in question 6 and find all frequent item-sets using FP-growth approach. *[(CO2)(Apply/LOCQ)]*

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

6 + 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. Show the associated dendrograms.

Table 4

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

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

(b) Define minimum distance and maximum distances between two clusters. *[(CSEN3132.1)(Remember/LOCQ)]*

8 + 4 = 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.

Table 5

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	46	43	11

Course Outcome (CO):

After the completion of the course students will be able to

CSEN3132.1. Learn and understand basic knowledge of data mining and related models.
CSEN3132.2. Understand and describe data mining algorithms.
CSEN3132.3. Understand and apply Data mining algorithms.
CSEN3132.4. Suggest appropriate solutions to data mining problems.
CSEN3132.5. Analyze data mining algorithms and techniques.
CSEN3132.6. Perform experiments in Data mining and knowledge discovery using real-world data.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*