INTRODUCTION TO DATA MINING
(DSC3101)

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) Which of the following can be appropriately represented as an ordinal attribute?
(a) Gender of a student (b) Ratings in a survey
(c) Hair colour (d) Probability of rain

(ii) In rule-based classification, when multiple rules apply, the system resolves conflicts using:
(a) Random selection (b) Rule ordering schemes
(c) Database indexing (d) Majority voting

(iii) The mathematical representation of Bayes Theorem is
(a) P(H|X) = P(X|H) * P(X) / P(H) (b) P(H|X) = P(X|H) / (P(X) * P(H))
(c) P(H|X) = P(X|H) * P(H) / P(X) (d) P(H|X) = P(X) / (P(X|H) * P(H))

(iv) Overfitting occurs when a model gives
(a) Low accuracy for both training data and new data (b) High accuracy for new data but not for training data
(c) High accuracy for training data but not for new data (d) High accuracy for both training data and new data

(v) Which of the following is not an example of ensemble learning algorithm?
(a) Support Vector Machine (b) Bagging
(c) Boosting (d) Random Forest

(vi) Which of the following is a key characteristic of the bagging technique in ensemble methods?
(a) It trains models sequentially, where each new model corrects the errors of the previous ones.
(b) It assigns higher weights to misclassified instances in subsequent iterations.
(c) It trains multiple models independently on different subsets of the data and averages their predictions.
(d) It is primarily used to reduce the variance in a model's predictions.

(vii) The Apriori principle states that:
(a) If an item set is frequent, then all of its subsets must also be frequent.
(b) If an item set is frequent, then all of its supersets must also be frequent.
(c) If an item set is infrequent, then all of its subsets must also be infrequent.
(d) None of the above.

(viii) In the context of Association Rule Mining, what is the role of the support metric?
(a) It measures the strength of the association between item sets.
(b) It indicates how often a certain item set appears in the dataset.
(c) It is a measure of the predictive power of a rule.
(d) It is used to filter out rules that are not interesting.

(ix) k-means clustering does not depend on
(a) Selection of distance metric (b) Selection of k
(c) Initial guess as to cluster centroids (d) Dimension of data points

(x) Which of the following can act as the best possible termination condition in K-Means clustering algorithm?
(a) For a fixed number of iterations
(b) Assignment of data points to clusters does not change between iterations
(c) Means of cluster changes frequently between successive iterations
(d) Number of clusters is gradually getting equal to k number of initial clusters

Fill in the blanks with the correct word

(xi) Scatter plots show _____ distributions.

(xii) In a binary classification problem, the probability of one class is 0.65. Its entropy is _____

(xiii) In Naïve Bayes Classifier, if P(H1)=0.48, P(H2) = 0.38 and P(H3)=0.14, then it refers to ___ class problem

(xiv) A dataset contains r distinct items. The total number of possible rules, extracted from the dataset is 12. What is the value of r? _____

(xv) One approach to determine the proper value of k in k-means is ____ plot.

Group - B

2. (a) Describe 4 important methods for handling missing values. *[(DSC3101.2)(Remember/LOCQ)]*

(b) What is an Outlier? Give an example where outlier detection may be helpful. *[(DSC3101.2)(Understand/LOCQ)]*

(c) (i) Assume that the minimum and maximum values for the attribute 'Income' are Rs 10,000 and Rs 100,000, respectively. Apply Min-max normalization technique to transform the value Rs. 37,100 for income.
(ii) Assume that the recorded values of A range from −957 to 983. Normalize by decimal scaling to find out the transformed value of A, when the recorded value of A is -750. *[(DSC3101.2)(Apply/IOCQ)]*

4 + 2 + (2 × 3) = 12

3. (a) Extract a rule-based system from the training sample given below.

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

[(DSC3101.2)(Apply/IOCQ)]

(b) For each of the rules generated, compute the accuracy and coverage of the rules. Which rule(s) appear to be better? *[(DSC3101.2)(Apply/IOCQ)]*

7 + (2 × 2 + 1) = 12

Group - C

4. (a) What is the main assumption made while solving the problem of classification using Naïve Bayes Classifier? *[(DSC3101.3)(Understand/LOCQ)]*

(b) "In Naïve Bayes Classification technique, prior probability of each class is required to be computed, whereas not the prior probability of the tuple to be classified" – Explain. *[(DSC3101.3)(Understand/LOCQ)]*

(c) Given the following table, classify the tuple X, who has a long hair, whose height is between 155 and 165 cm, who is in job but cannot cook. Solve the problem using Naïve Bayesian Classifier. *[(DSC3101.3, DSC3101.6) (Apply/IOCQ)]*

Sl. No.	Hair	Height	In job	Can cook	Gender
1	Long	Less than 155 cm	No	No	Female
2	Long	Less than 155 cm	No	Yes	Female
3	Medium	Less than 155 cm	No	No	Male
4	Short	155 –165 cm	No	No	Male
5	Short	Greater than 165 cm	Yes	No	Male
6	Short	Greater than 165 cm	Yes	Yes	Female
7	Medium	Greater than 165 cm	Yes	Yes	Male
8	Long	155 – 165 cm	No	No	Female
9	Long	Greater than 165 cm	Yes	No	Male
10	Short	155 – 165 cm	Yes	No	Male
11	Long	155 – 165 cm	Yes	Yes	Male
12	Medium	155 – 165 cm	No	Yes	Male
13	Medium	Less than 155 cm	Yes	No	Male
14	Short	155 – 165 cm	No	Yes	Female

2 + 2 + 8 = 12

5. (a) Explain with diagrams what you understand by linearly separable data and linearly non-separable data. *[(DSC3101.3)(Understand/LOCQ)]*

(b) Suppose a support vector machine for separating circles from squares finds a circle support vector at the point x1= (2, 0), a square support vector at x2= (0, 2). Determine the classification vector **W** and the threshold value b. *[(DSC3101.3)(Apply/IOCQ)]*

(c) With the help of a diagram, explain what slack variable is. *[(DSC3101.3)(Understand/LOCQ)]*

(d) Name two important Kernel Functions with the formulæ and graphs. *[(DSC3101.3)(Remember/LOCQ)]*

3 + 3 + 2 + 4 = 12

Group - D

6. (a) The following table shows a Market Basket Data. Assume minimum support count = 3.
(i) Construct the FP-tree.
(ii) Prepare a table showing conditional pattern base, conditional FP-tree and generated frequent patterns against the items for which the support count is greater than or equal to minimum support count. *[(DSC3101.4) (Apply/IOCQ]*

Transaction IDs	List of items IDs
1	A, D, F, G
2	C, D, F, H, I, K
3	D, F, G, H, I, K
4	B, D, E, F, I
5	B, F, G, J, K

(b) What is Apriori Property? How does it help in finding the frequent itemsets?

[(DSC3101.4) (Remember/LOCQ]

(3 + 3 × 2) + (2 + 1) = 12

7. (a) Define support, confidence and lift in frequent pattern mining. Are these measures symmetric? Justify your answer.

[(DSC3101.4) (Remember, Understand/LOCQ]

(b) In a fast-food restaurant there are 5 different food items (viz., M1, M2, M3, M4 and M5) available in their menu. Food items ordered in 9 different online transactions/orders are given in the table below:

Order Id	Ordered Food
1	M1, M2, M5
2	M2, M3
3	M2, M4, M5
4	M1, M3
5	M2, M3
6	M1, M2, M3, M5
7	M1, M2, M4, M5
8	M1, M2, M3
9	M1, M3

Compute the support for item-sets {M5}, {M2, M4} and {M2, M4, M5} by treating each Order ID as a market basket.

[(DSC3101.4) (Apply/IOCQ]

(c) Use the above results to compute the confidence of the association rules {M2, M4} → {M5} and {M5} → {M2, M4}.

[(DSC3101.4) (Apply/IOCQ]

(d) If the minimum support is 20% and the minimum confidence is 60%, are the above two rules strong but misleading?

[(DSC3101.4)(Evaluate/IOCQ)]

4 + 2 + 4 + 2 = 12

Group - E

8. Perform K-means clustering (using Euclidean distance as distance function) on 2-dimensional data points, given in the following table. Assume that the initial centroids are P5, P7 and P9. Show the centroids and clusters in first two iterations.

Points	X coordinate	Y coordinate
P1	1	10
P2	10	2
P3	7	3
P4	2	3
P5	9	5
P6	4	11
P7	3	9
P8	3	5
P9	4	3

[(DSC3101.3, DSC3101.6)(Apply/IOCQ)]

12

9. (a) Define minimum distance and maximum distances between two clusters. *[(DSC3101.3)(Remember/LOCQ)]*

(b) Construct the dendrograms for the following distance matrix using both minimum distance and maximum distance. Show all the steps. *[(DSC3101.3, DSC3101.6)(Apply/IOCQ)]*

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.90	0.35	0.80
P2	0.10	0.00	0.30	0.40	0.50
P3	0.90	0.30	0.00	0.60	0.70
P4	0.35	0.40	0.60	0.00	0.80
P5	0.80	0.50	0.70	0.80	0.00

2 + 10 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	29.2	70.8	0