DATA MINING
(AML2101)

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 may be the best way to represent confidence of an association rule (A⇒B), where σ denotes the support count?
(a) σ (A∩B)/ σ (A) (b) σ (AUB)/ σ (A)
(c) σ (A∩B)/ σ (B) (d) σ (AUB)/ σ (B)

(ii) Two documents are said to be very close to each other when the Cosine Similarity measure between their term frequency vectors is
(a) Close to 0 (b) Exactly 0
(c) Close to 1 (d) Exactly 1

(iii) 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

(iv) In a 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

(v) 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

(vi) Sigmoid function is most appropriately represented by which of the following?
(a) $\frac{1}{exp(t)+exp(-t)}$ (b) $t\ exp(-t)$
(c) $\frac{1}{1+exp(t)}$ (d) $\frac{1}{1+exp(-t)}$

(vii) The learner is trying to predict housing prices based on the size of each house. What type of regression is this?
(a) Multivariate Logistic Regression (b) Logistic Regression
(c) Linear Regression (d) Multivariate Linear Regression

(viii) A neuron with input [x1, x2, x3] = [0 · 3, 0 · 5, 0 · 6] and weights [w1, w2, w3] = [0 · 2, 0 ·1, − 0 · 3], the output of linear transformation in this neuron having bias=0.08 will be
(a) 0.07 (b) -0.07
(c) 0.01 (d) -0.01

(ix) 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

(x) Which of the following is finally produced by Hierarchical Clustering?
(a) Final estimate of cluster centroids (b) Tree showing how close things are to each other
(c) Assignment of each point to clusters (d) All of the mentioned

Fill in the blanks with the correct word

(xi) Let X1, ..., Xm 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 _____.

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

(xiii) If a rule R covers 6 out of 14 tuples and if it correctly classifies 4 out of them, then the accuracy of R is _____.

(xiv) A split in the construction of a decision tree is said to be homogeneous if the degree of impurity is _____.

(xv) When True Positive value is 437 and False Positive value is 63, then the precision is _____.

Group - B

2. (a) What do you mean by support and confidence of any association rule? When do you refer a rule strong?

[(AML2101.4) (Remember, Understand/LOCQ)]

(b) Consider the following dataset of transactions with each letter representing an item:

Transaction ID	Items
T1	{E, K, M, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, M}
T4	{C, K, M, U, Y}
T5	{C, E, I, K, O}

Construct the FP-tree for the above dataset and find all frequent item-sets using FP-growth approach considering the minimum support count as 3.

[(AML2101.4) (Apply/IOCQ)]

(2 + 1) + (5 + 4) = 12

3. (a) Explain how box plot can be used to remove the outliers. *[(AML2101.2) (Remember, Understand/LOCQ)]*

(b) Suppose that the data for analysis includes the attribute 'age'. The 'age' values for the data tuples are (in non decreasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Now, answer the following questions:

(i) What is the mean of the data? What is the median?
(ii) What is the mode of the data? Comment on the data's modality.
(iii) What is the mid range of the data?
(iv) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
(v) Give the five-number summary of the data.

[(AML2101.2) (Apply/IOCQ)]

3 + (2 + 2 + 1 + 2 + 2) = 12

Group - C

4. (a) Consider the following dataset:

Instance	*A*	*B*	*C*	Class
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

(i) Estimate the conditional probabilities for P (A = 1|+), P (B = 1|+), P (C = 1|+), P (A = 1|-), P (B = 1|-), P (C = 1|-).
(ii) Use the conditional probabilities in part (i) to predict the class label for a test sample (A = 1, B = 1, C = 1) using the Naive Bayes approach.

[(AML2101.3) (Apply/IOCQ)]

(b) Compare the advantages and disadvantages of eager classification versus lazy classification. *[(AML2101.1) (Remember/LOCQ)]*

(6 + 2) + 4 = 12

5. (a) Compare Information Gain, Gain Ratio and Gini Index as attribute selection measures to create decision tree.

[(AML2101.3) (Understand/LOCQ)]

(b) Find out, using gain in Gini Index, what will be the root node in the decision tree for classifying whether an unknown person is Male or Female. The training data is given below:

Sl. No.	**Over 170 cm (Yes/ No)**	**Eye (Blue or Brown)**	**Hair (Short or Long)**	**Gender (Male/ Female)**
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

[(AML2101.3, AML2101.6) (Apply/ IOCQ, Evaluate/ HOCQ)]

4 + 8 = 12

Group - D

6. (a) What is the advantage of having hidden layers in an Artificial Neural Network (ANN)? *[(AML2101.3)(Analyze/IOCQ)]*

   (b) Consider an ANN where all the neurons are Linear Threshold Units. Train the network using Perceptron learning rule. The set of inputs & desired output training vectors is as follows:
   $(X^{(1)} = [1, -2, 0, -1]^T; d_1 = -1)$, $(X^{(2)} = [0, 1.5, -0.5, -1]^T; d_2 = -1)$, $(X^{(3)} = [-1, 1, 0.5, -1]^T; d_3=1)$. Initial weight vector is $W^{(1)} = [1, -1, 0, 0.5]^T$ and $\eta = 0.1$. Show the learning process step by step for two epochs and report the final weight vector after the completion of two epochs. *[(AML2101.3)(Apply/IOCQ)]*

2 + 10 = 12

7. (a) "Artificial Neural Network can be used as a classifier" – justify the statement. *[(AML2101.3) (Analyze/IOCQ)]*

   (b) Explain the role of ensemble methods in classification. Do you think that an ensemble tends to be more accurate than its base classifiers? Give justifications in support of your answer. *[(AML2101.5) (Analyze/IOCQ)]*

   (c) Briefly explain the working principle of random forests by explicitly mentioning how it improves the accuracy of the decision tree induction. *[(AML2101.5) (Analyze/IOCQ)]*

4 + (2 + 2) + 4 = 12

Group - E

8. (a) Define core point, border point and noise point in DBSCAN algorithm with a diagram. *[(AML2101.3) (Remember/LOCQ)]*

   (b) Consider the following 7 data points:
   A (10, 5), B (1, 4), C (5, 8), D (9, 2), E (12, 10), F (15, 8), G (7, 7).
   Cluster the given points by using Hierarchical clustering using MAX (Complete Linkage) distance. Draw the dendrogram with merging distance and clearly show the merge sequence. *[(AML2101.3) (Apply/IOCQ)]*

(1 + 1 + 1) + 9 = 12

9. (a) Clustering is recognized as an important data mining task with broad applications. Give one application example for each of the following cases:
   (i) An application that uses clustering as a major data mining function.
   (ii) An application that uses clustering as a pre-processing tool for data preparation for other data mining tasks.
   [(AML2101.3)(Remember/LOCQ)]

   (b) Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are: A1 (2,10) ,A2 (2,5) ,A3 (8,4) ,B1 (5,8) ,B2 (7,5) ,B3 (6,4) , C1 (1,2) ,C2 (4,9).
   The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only –
   (i) The three cluster centers after the first round of execution.
   (ii) The three clusters after the second round of execution. *[(AML2101.3)(Apply/IOCQ)]*

2 + (3 + 7) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	19.79	71.88	8.33