MACHINE LEARNING
(CSE5131)

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 poor binary classification model for detecting a rare cancer always predicts positive for presence of the disease. What can we infer about the model's performance?
(a) The model has high accuracy, maximum precision but low recall
(b) The model has poor accuracy, poor precision, but maximum recall
(c) The model has poor accuracy, maximum precision and minimum recall
(d) The model has maximum accuracy, maximum precision but minimum recall

(ii) Which of the following methods do we use to best fit the data in Logistic Regression?
(a) Least Square Error (b) Maximum Likelihood
(c) Both (a) and (b) (d) None of the above.

(iii) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?
I. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.
II. In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.
(a) Only I (b) Only II
(c) Both I and II (d) None of I and II

(iv) 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

(v) Which of the following techniques would perform better for reducing the dimensions of a data set?
(a) Removing columns that have too many missing values
(b) Removing columns that have high variance in data
(c) Removing columns with dissimilar data trends
(d) None of these

(vi) PCA finds the directions of maximum variance in the data by
(a) Singular Value Decomposition (SVD)
(b) Eigen decomposition
(c) Correlation analysis
(d) T-distributed Stochastic Neighbour Embedding (t-SNE)

(vii) The maximum number of dichotomies which can be achieved by of a positive ray hypothesis using a training data set having 10 data points is
(a) 10 (b) 11 (c) 9 (d) 12

(viii) Suppose your model is demonstrating high variance across different training sets. Which of the following is not a valid way to try and reduce the variance?
(a) Increase the amount of training data in each training set.
(b) Improve the optimization algorithm being used for error minimization.
(c) Decrease the model complexity.
(d) Reduce the noise in the training data.

(ix) When a model performs well on training data (the data on which the algorithm was trained) but does not perform well on test data (new or unseen data), we say that the model is
(a) Overfitting (b) Generalizing (c) Regularizing (d) None of the above

(x) What is the primary purpose of a Convolutional Neural Network (CNN)?
(a) Object detection (b) Image classification
(c) Text generation (d) Reinforcement learning

Fill in the blanks with the correct word

(xi) When a model performs well on training data (the data on which the algorithm was trained) but does not perform well on test data (new or unseen data), we say that the model is _____.

(xii) In an MLP, the number of nodes in the input layer is 10 (including bias node) and in the hidden layer is 5 (including bias node). The maximum number of connections from the input layer to the hidden layer is _____.

(xiii) In regularization, the _____ parameter controls the trade-off between fitting the training data well and keeping the model parameters small.

(xiv) The VC dimension of a perceptron is _____ than VC dimension of a linear SVM.

(xv) The activation function _____ is commonly used in the convolutional layers of a Convolutional Neural Network.

Group - B

2. (a) Deduce the formula for multi-variate linear regression. *[(CO1)(Remember/LOCQ)]*
(b) The following table provides the marks obtained 10 students in class test and semester examination in machine learning. Estimate the marks a student may obtain in the semester examination when she obtained 20 in class test using linear regression.

Sl No	Marks in Class test	Marks in Semester
1	28	53
2	27	39
3	23	47
4	17	36
5	24	40
6	28	39
7	16	36
8	11	30
9	22	35
10	18	42

[(CO2)(Understand/LOCQ)]

5 + 7 = 12

3. (a) Explain mathematically how the weight and bias expressions are determined for a Linear Regression Model utilizing the Least Squares Method on a single dependant variable. *[(CO1)(Understand/LOCQ)]*

   (b) 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 /IOCQ)]*

   (c) Define overfitting in a decision tree and describe a method to avoid overfitting in a decision tree. *[(CO4)(Apply/HOCQ)]*

4 + 4 + 4 = 12

Group - C

4. (a) Define MIN, MAX and Average distance in the context of hierarchical clustering. *[(CO1)(Remember/LOCQ)]*

   (b) Consider the data points provided in the table below. Perform hierarchical clustering considering complete link method (MAX distance). Try to approximately plot them on a 2D plane and show the nested clusters. Show the dendrogram with merging distance on Y-axis. *[(CO2 & CO3)(Describe & Apply/LOCQ)]*

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

3 + 9 = 12

5. (a) Explain the significance of dimensionality reduction. *[(CO2)(Understand/IOCQ)]*

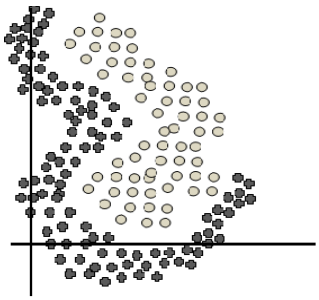
(b) What is PCA? What does a PCA do? *[(CO1)(Remember/LOCQ)]*

(c) Write the steps of a PCA algorithm. *[(CO2)(Desribe/LOCQ)]*

4 + 4 + 4 = 12

Group - D

6. (a) Suppose that we want to build a neural network that classifies two dimensional data (i.e., X = [x1, x2]) into two classes: pluses and circles. We have a set of training data that is plotted as follows:



Draw a network that can solve this classification problem. Justify your choice of the number of nodes and the architecture. Draw the decision boundary that your network can find on the diagram. *[(CO6)(Apply/HOCQ)]*

(b) Explain the Bias-Variance trade off in the context of learning. *[(CO1)(Remember/LOCQ)]*

6 + 6 = 12

7. (a) Explain the importance of Hoeffding's inequality in the feasibility of learning. *[(CO4) (Analyse/HOCQ)]*

(b) Discuss, in brief, the difference between Input space and feature space. Also explain how the linear regression formula can be used for nonlinear cases. *[(CO4)(Analyse/LOCQ)]*

(c) Explain with example the 1-hot representation in the context of pre-processing of data. *[(CO1) (Understand/LOCQ)]*

4 + 6 + 2 = 12

Group - E

8. (a) 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. *[(CO2 & CO3)(Understand/LOCQ)]*

(b) Discuss about the tradeoff parameter between error and margin in the case soft margin SVM. *[(CO4)(Analyse/HOCQ)]*

9 + 3 = 12

9. Write short notes on any three of the followings:
(i) Overfitting in machine learning (ii) Regularization in machine learning
(iii) Validation in machine learning (iv) Weight Decay in Regularization
(v) Deterministic Noise. *[(CO4)(Remember/LOCQ)]*

(3 × 4) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	74	8	18