MACHINE LEARNING
(CSEN 3233)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) An artificial neuron receives n inputs $x_1, x_2, \ldots\ldots, x_n$ with weights $w_1, w_2, \ldots\ldots\ldots, w_n$ attached to the input links. The weighted sum _____ is computed to be passed on to a non-linear filter Φ called activation function to release the output.
(a) $\sum_{i=1}^{n} W_i$ (b) $\sum_{i=1}^{n} X_i$ (c) $\sum_{i=1}^{n} W_i + \sum_{i=1}^{n} X_i$ (d) $\sum_{i=1}^{n} W_i * X_i$

(ii) For a neural network, which one of these structural assumptions is the one that most affects the trade-off between under-fitting and overfitting?
(a) Number of hidden nodes (b) Learning rate
(c) Initial choice of weights (d) Use of constant term unit input.

(iii) 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.

(iv) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?
1. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.
2. 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 1 (b) Only 2
(c) Both 1 and 2 (d) None of 1 and 2.

(v) The effectiveness of an SVM depends upon
(a) Selection of Kernel (b) Kernel Parameters
(c) Soft Margin Parameter C (d) All of the above

(vi) 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.

(vii) Data used to optimize the parameter settings of a supervised learning model
(a) training (b) test (c) verification (d) validation.

(viii) It is given that the VC-dimension of H (d_{VC}) is n. This means:
(a) There are n points we cannot shatter
(b) There are n+1 points we cannot shatter
(c) We cannot shatter any set of n points
(d) We cannot shatter any set of n+1 points.

(ix) Consider a binary classification problem. Suppose you have trained a model on a linearly separable training set, and you get a new labeled data point which is correctly classified by the model, and far away from the decision boundary. If you now add this new point to your earlier training set and re-train, in which cases is the learnt decision boundary likely to change?
(a) When your model is a perceptron
(b) When your model is logistic regression
(c) When your model is an SVM
(d) None of the above.

(x) **Statement 1:** The error of a hypothesis measured over a training set provides a pessimistically biased estimate of true error of the hypothesis
Statement 2: Gradient descent have the problem of falling into local minima
(a) Only Statement 1 is true (b) Only Statement 2 is true
(c) Both are true (d) Both are false.

Group- B

2. (a) What are the in-sample and out-of-sample errors in the context of classification? *[(CO1)(Remember/LOCQ)]*

(b) Define Hoeffding's inequality in the context of feasibility of learning. *[(CO5)(Analyze/HOCQ)]*

(c) Consider the following confusion matrix for a two-class data set on which classification has been done:

	Actual class 1	Actual class 2
Placed in class 1	34	26
Placed in class 2	36	44

Evaluate the overall accuracy of this classification results. Also calculate the Precision and Recall values. *[(CO2)(Understand/IOCQ)]*

(d) Derive the linear regression formula for multiple dependent variables. Also explain how the derived linear regression formula can be used for nonlinear cases. *[(CO2)(Understand/LOCQ)]*

2 + 2 + 2 + (4 + 2) = 12

3. (a) Write the Perceptron Learning Algorithm (PLA) and briefly explain the working principle of the algorithm. *[(CSEN3233.1)(Remember/LOCQ)]*

(b) Explain how the same can be used for non-linear regression.
[(CSEN3233.2)(Understand/LOCQ)]

(c) Derive the linear regression formula for single dependent variables.
[(CSEN3233.3)(Understand/IOCQ)]

4 + 4 + 4 = 12

Group - C

4. (a) Define Dichotomy, Growth function and break point. *[(CO3)(Understand/LOCQ)]*

(b) Prove that the number of dichotomies on N points for a hypothesis with finite break point is polynomial in N. *[(CO4)(Analyze/HOCQ)]*

3 + 9 = 12

5. (a) Define VC dimension. *[(CO1)(Remeber/LOCQ)]*

(b) Explain the importance of VC dimension in machine learning?
[(CO3)(Understand/HOCQ)]

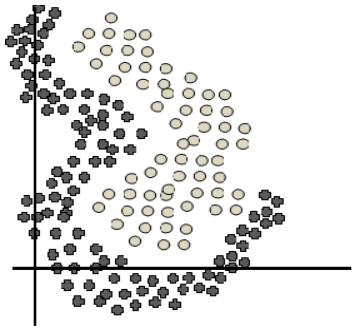
(c) Find the VC Dimension for the following hypotheses:
(i) Positive intervals $F(x) = +1$ for $a \leq x \leq b$; -1 otherwise.
(ii) Perceptron in R^2. *[(CO4)(Analyse/LOCQ)]*

(d) You are given 4 points X_1, X_2 , X_3 and X_4. Calculate the number of dichotomies when break point is 2. *[(CO6)(Apply/IOCQ)]*

2 + 3 + 4 + 3 = 12

Group - D

6. (a) Suppose that we want to build a neural network that classifies two dimensional data (i.e., $X = [x_1, x_2]$) 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/IOCQ)]*

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

6 + 6 = 12

7. (a) Explain how multi-layer perceptron (MLP) can be used as an estimator. Write all the required steps, in detail. *[(CO6)(Apply/LOCQ)]*

(b) Explain the main reasons why a Back-Propagation training algorithm might not find a set of weights which minimizes the training error for a given feed-forward neural network. *[(CO4)(Analyse/HOCQ)]*

(c) Explain the purpose of the momentum term that is often included in the Back-Propagation learning algorithm. *[(CO4)(Understand/IOCQ)]*

6 + 3 + 3 = 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. *[(CO1)(Remember /LOCQ)]*

(b) State the mercer's condition in selecting a kernel function for non-linear SVM. *[(CO5)(Analyse/IOCQ)]*

9 + 3 = 12

9. (a) Write short notes on any two of the followings:
(i) Overfitting
(ii) Regularization in Machine Learning
(iii) Validation in Machine Learning. *[(CO1)(Remember/IOCQ)]*

(b) A linearly separable dataset is given in the following Table. Predict the class of (0.6, 0.8) using a support vector machine classifier. *[(CO6)(Apply/HOCQ)]*

X_1	X_2	Y	Lagrange Multiplier
0.3	0.4	+1	5
0.7	0.6	-1	8
0.9	0.5	-1	0
0.7	0.9	-1	0
0.1	0.05	+1	0
0.4	0.3	+1	0
0.9	0.8	-1	0
0.2	0.01	+1	0

(2 × 4) + 4 = 12

Cognition Level	*LOCQ*	*IOCQ*	*HOCQ*
Percentage distribution	*45*	*33*	*22*

Course Outcome (CO):

After the completion of the course students will be able to

1. Learn and understand the basics of machine learning approaches and paradigm.
2. Understand and describe various machine learning algorithms.
3. Understand complexity of Machine Learning algorithms and their limitations.
4. Mathematically Analyse various machine learning approaches and paradigms
5. Analyse various machine learning techniques to get an insight of when to apply a particular machine learning approach.
6. Apply common Machine Learning algorithms in practice and implementing their own using real-world data.

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