MACHINE LEARNING
(CSEN 4264)

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) Let's say, a "linear regression" model perfectly fits the training data (training error is zero). Now, which of the following statement is true?
(a) You will always have test error zero
(b) You cannot have test error zero
(c) none of the above
(d) can't say.

(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) Jaccard distance (d) both (a) and (b).

(iii) There is at least one set of four points in R^3 that can be shattered by the hypothesis set of all 2D planes in R^3.
(a) True (b) False
(c) Can't say (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 growth function h(N) for positive intervals (h(X) = 1 when a ≤ X ≤ b and h(X) = -1 otherwise) is
(a) N+1 (b) ½ N(N+1) (c) 2^N (d) ½ N(N+1) + 1.

(vi) The back-propagation algorithm learns a globally optimal neural network with hidden layers.
(a) always true (b) always false
(c) mostly true (d) mostly false.

(vii) H consists of all hypotheses in two dimensions h: $R^2 \rightarrow \{-1, +1\}$ that are positive inside some convex set and negative elsewhere. The break point of H is
(a) N (b) N+1 (c) ∞ (infinity) (d) 2^N.

(viii) **Statement 1:** The error surface followed by the gradient descent back propagation algorithm changes if we change the training data.
Statement 2: Stochastic gradient descent is always a better idea than batch gradient descent.
(a) only statement 1 is true (b) only statement 2 is true
(c) both are true (d) both are false.

(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) The effectiveness of an SVM depends upon
(a) selection of kernel (b) kernel parameters
(c) soft margin parameter C (d) all of the above.

Group – B

2. (a) Derive the linear regression formula for single dependent variables.

(b) Consider the perceptron in two dimensions: $h(x) = \text{sign}(w^T x)$ where $w = [w_o, w_1, w_2]^T$ and $x = [1, x_1, x_2]^T$. Technically, x has three coordinates, but we call this perceptron two-dimensional because the first coordinate is fixed at 1.
(i) Show that the regions on the plane where h(x) = + 1 and h(x) = - 1 are separated by a line. If we express this line by the equation $x_2 = ax_1 + b$, what are the slope a and intercept b in terms of w_o , w_1, w_2?
(ii) Draw a picture for the cases $w = [3, 2, 1]^T$ and $w = -[3 , 2, 1]^T$.
6 + 6 = 12

3. (a) Discuss with example the in-sample error and out-of-sample error.

(b) Classes attended by 10 students in machine learning and marks obtained in the examination are provided in the following table. Estimate the marks a student may obtain in the examination when she attended 20 classes, using linear regression.

Sl No	Attendance	Marks	Sl No	Attendance	Marks
1	28	43	6	28	39
2	27	39	7	26	36
3	23	27	8	21	36
4	27	36	9	22	31
5	24	34	10	28	37

4 + 8 = 12

Group – C

4. (a) Explain the importance of VC dimension in machine learning?

   (b) Find the VC dimension for the following hypotheses:
   (i) Positive intervals
   F(x) = +1 for a ≤ x ≤ b; -1 otherwise.
   (ii) Perceptron in R^2.

   (c) Explain the Bias-Variance trade off in the context of learning.

3 + 3 + 6 = 12

5. (a) Decompose out-of-sample error in terms of bias and variance based on squared error measure.

   (b) Let B(N, k) be the maximum number of dichotomies on N points such that no subset of size k of the N points can be shattered by these dichotomies. By assuming B(N, k) ≤ B(N-1, k) B(N-1, k-1) and defining the necessary boundary conditions, show that

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

6 + 6 = 12

Group – D

6. (a) What does the learning rate do in back-propagation training?

   (b) Describe what is likely to happen when a learning rate is used that is too large, and when one is used that is too small. How can one optimize the learning rate?

   (c) Describe the importance of using bias in neural network.

   (d) 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.

   (e) Explain the purpose of the momentum term that is often included in the back-propagation learning algorithm.

2 + 3 + 2 + 3 + 2 = 12

7. (a) Suppose a one-layered neural network with a single weight w is used to implement a function y = 2x + 3 + c, where x and y are the input and output parameters respectively, whereas c is Gaussian noise (random number). Derive the update equation using gradient descent approach to minimize the mean squared error.

   (b) You are asked to simulate the Boolean function $x_1 \wedge x_2 \vee (\neg X_3)$ using a multi-layer perceptron. Construct the network and explain how your network is able to model the said function.

6 + 6 = 12

Group – E

8. Construct the primal problem and then derive the Lagrangian and its dual for the optimization problem as defined by linear support vector machine (SVM) classification.

12

9. Suppose we have five 1D data points
$x_1=1$, $x_2=2$, $x_3=4$, $x_4=5$, $x_5=6$, with 1, 2, 6 as class 1 and 4, 5 as class 2 $\Rightarrow$ $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$, $y_5=1$.
When a polynomial kernel of degree two ($K(x,y) = (xy+1)^2$) is used and C is set to 100, we get the Lagrange multipliers as follows:
$\alpha_1=0$, $\alpha_2=2.5$, $\alpha_3=0$, $\alpha_4=7.333$, $\alpha_5=4.833$
Identify the support vectors and derive the discrimination function.

12