CSE 417T (Machine Learning): Exam 1

50 points, 75 minutes.
March 10, 2022

Full Name	
Student ID	

Instructions: Write your answers in the space provided. For your convenience, parts of the pages have been designated as "scratch work areas." These will not be graded. We will not accept or grade any work that does not appear in the space provided for answers.

Format: There are two sections. The first has 5 long answer questions, the second has 10 multiple choice questions. Please check that you have all the questions in your exam handout. First do all the questions you're relatively sure of to guarantee those points.

Do not turn the page until you are instructed to start

1 Long questions (30 points)

1. (6 points) You are a reviewer for the International Mega-Conference on Machine Learning for Everything, and you read papers with the following main claims/results. Should you accept or reject the paper? Provide a one-to-two sentence justification.

   (a) **accept / reject**. "My algorithm is better than yours because I used the simplest possible hypothesis set to avoid overfitting!"

   (b) **accept / reject**. "My algorithm is better than yours. Look at the reported test error! (The authors calculate the test errors of multiple learning models on the same test set, and report the one with the smallest error.)

   (c) **accept / reject**. "My algorithm is better than yours. It can fit any training dataset perfectly!"

Scratch work area: Will **not** be graded!

2. (6 points) Suppose you have the following cost-matrix for a classification problem, where h is your hypothesis and f is the true target function:

		f	
		+1	-1
h	+1	C_1	C_2
	-1	C_3	C_4

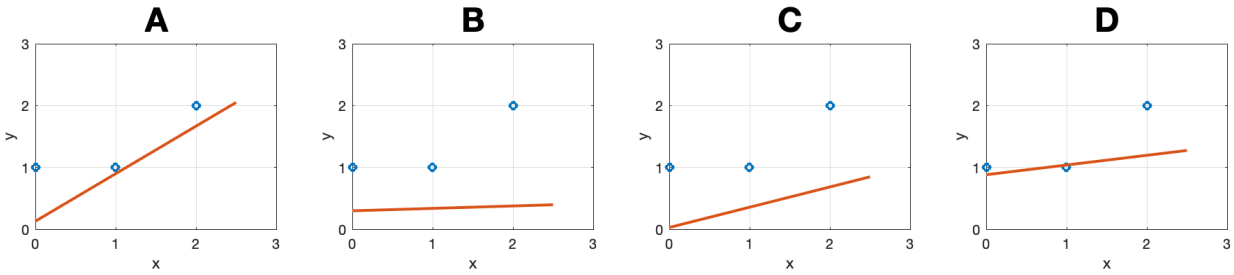
You have trained a logistic regression model and want to used it for classification. Suppose your logistic regression model outputs probability p that a data point is positive (which is your estimate of the probability that the label is +1). Note that in this application, you might still incur cost even if you make correct predictions. Above what threshold (expressed using C_1, C_2, C_3, and C_4) would you choose to classify that data point as positive? Why?

Scratch work area: Will **not** be graded!

3. (6 points) Consider a linear regression with 1-dimensional input, i.e., each data point is represented by (x, y). We performed linear regression on a dataset $D = \{(0, 1), (1, 1), (2, 2)\}$ with the following regularizations (E_{in} is the squared error):

(1) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda(w_0^2 + 10w_1^2)$, where $\lambda = 1$
(2) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda(w_0^2 + 10w_1^2)$, where $\lambda = 10$
(3) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda(10w_0^2 + w_1^2)$, where $\lambda = 1$
(4) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda(10w_0^2 + w_1^2)$, where $\lambda = 10$

The regression results are shown in the plots below. However, we got confused and forgot which plot corresponds to which regularization method. Please map each plot to one of the regularizations and explain why.



Your mapping (2 out of 6 points): (1): _____ (2): _____ (3): _____ (4): _____
(Input A to D for each space).

Explanations (4 out of 6 points):

Scratch work area: Will **not** be graded!

4. (6 points) You are given a dataset with 1-dimensional input $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ to perform a regression problem. Your hypothesis set contains hypotheses of the form:

   $$h(x) = w_0 + w_1^2 x$$

   Note that each hypothesis h can be characterized by $\vec{w}$ with two elements w_0 and w_1. We want to minimize the sum of the following point-wise error function: $e(h(x), y) = (h(x) - y)^4 = (w_0 + w_1^2 x - y)^4$. Derive a gradient descent algorithm that minimizes the in-sample error (average error in the training dataset) for this regression problem. You only need to write the update step of the gradient descent, i.e., in the form $\vec{w}(t+1) \leftarrow \vec{w}(t) + \cdots$. Show your derivations. (You may write the update for w_0 and w_1 separately if you find that easier.)

Scratch work area: Will **not** be graded!

5. (6 points) Suppose you have the following training dataset for a binary classification problem.

x_1	x_2	y
+1	+1	-1
+1	-1	+1
-1	+1	+1
-1	-1	-1

Which of the following feature transformations can be applied to this dataset such that the transformed data is linearly separable? Select all that apply.

(a) $\Phi(\vec{x}) = \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \end{bmatrix}$ (b) $\Phi(\vec{x}) = \begin{bmatrix} 1 \\ x_1 - x_2 \end{bmatrix}$ (c) $\Phi(\vec{x}) = \begin{bmatrix} 1 \\ |x_1 - x_2| \end{bmatrix}$ (d) $\Phi(\vec{x}) = \begin{bmatrix} 1 \\ x_1 x_2 \end{bmatrix}$

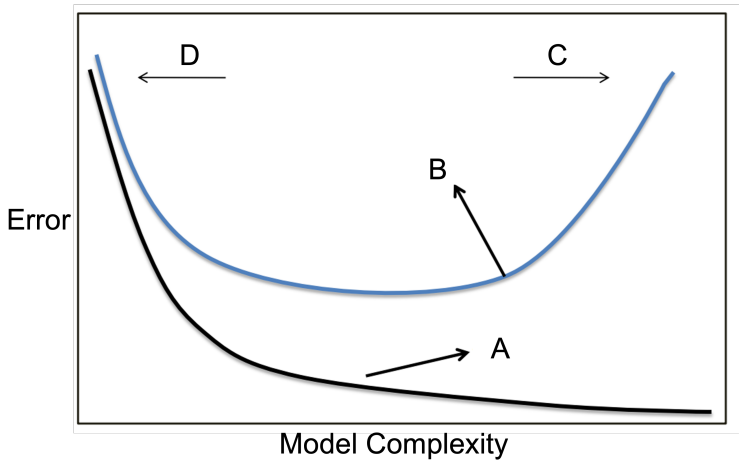
For each option that you selected, write down a separator (linear in the transformed space) of the form $h(\vec{x}) = sign(\vec{w}^T \Phi(\vec{x}))$ that separates this dataset.

Scratch work area: Will **not** be graded!

2 Multiple choice questions (20 points)

Instructions: Select the correct answer for each problem. You will get 2 points for a correct answer. There is no penalty for an incorrect answer. Please make sure the selection is clear and not ambiguous.

Please see the following figure for the first two questions. A learner has split the dataset into training and validation (not used for model selection). Assume the learner has followed the procedure we discussed in class and obtained training and validation errors with increasing model complexity. The below figure depicts the error curves.



1. Which of the curve is more likely to be the training error and which is more likely to be the validation error?

   ◯ A: Training; B: Validation

   ◯ A: Validation; B: Training

   ◯ A: Both training and validation

   ◯ B: Both training and validation

2. Compare the regions of C and D. What are the relative bias / variances in each region?

   ◯ C: low bias / low variance; D: high bias / high variance

   ◯ C: low bias / high variance; D: high bias / low variance

   ◯ C: high bias / low variance; D: low bias / high variance

   ◯ C: high bias / high variance; D: low bias / low variance

---

Scratch work area: Will **not** be graded!

3. For any finite hypothesis set $\mathcal{H}$ such that $|\mathcal{H}| = M$, the tightest correct upper bound for $d_{\text{VC}}(\mathcal{H})$ is:

   ◯ $M \ln M$

   ◯ $\log_2 M$

   ◯ M

   ◯ 42

4. To show that the VC dimension of H is at most $d + 1$, what do we have to prove.

   ◯ Every set of $d + 2$ points can be shattered by H.

   ◯ Every set of $d + 1$ points cannot be shattered by H.

   ◯ Every set of $d + 2$ points cannot be shattered by H.

   ◯ There is a set of $d + 1$ points that can be shattered by H.

   ◯ There is a set of $d + 1$ points that cannot be shattered by H.

5. The VC-dimension of axis-parallel linear separators (i.e., each hypothesis in the hypothesis set is a linear separator that is parallel to one of the axis) in two dimensions is

   ◯ 1

   ◯ 2

   ◯ 3

   ◯ 4

Scratch work area: Will **not** be graded!

6. Suppose I were to train a logistic regression model on the following (one-dimensional) dataset of (x, y) values: $\{(-10, -1), (30, -1), (30, 1), (-10, 1)\}$. What would the returned values of w_0 and w_1 be?

   ◯ 10 and 10
   ◯ 0 and 10
   ◯ 10 and 0
   ◯ 0 and 0
   ◯ 10 and 1

7. You are given dataset of (x, y) values: $\{(-1, 0), (0, 1), (1, 0)\}$ and asked to choose between the following two models: $H_0 = \{h_0(x) = b\}$ and $H_1 = \{h_1(x) = ax + b\}$. The error measure is squared error. Which model would you choose, assuming you use leave-one-out cross validation for model selection?

   ◯ H_0
   ◯ H_1

8. Consider V-fold cross-validation. Let's consider the tradeoffs of larger or smaller V (the number of folds). Generally speaking, on average, with a higher number of folds, the cross-validation error will be

   ◯ Higher
   ◯ Lower
   ◯ Same

Scratch work area: Will **not** be graded!

9. Overfitting is generally a symptom of high **bias/variance** (choose one in (a)), and underfitting is generally a symptom of high **bias/variance** (choose one in (b)).

   ◯ (a) bias, (b) bias

   ◯ (a) bias, (b) variance

   ◯ (a) variance, (b) bias

   ◯ (a) variance, (b) variance

10. If you are given m data points, and use half for training and half for testing, for any hypothesis set, the difference between training error and test error decreases as m increases.

    ◯ True

    ◯ False

Scratch work area: Will **not** be graded!