CSE 417T (Machine Learning): Exam 1 Practice Questions

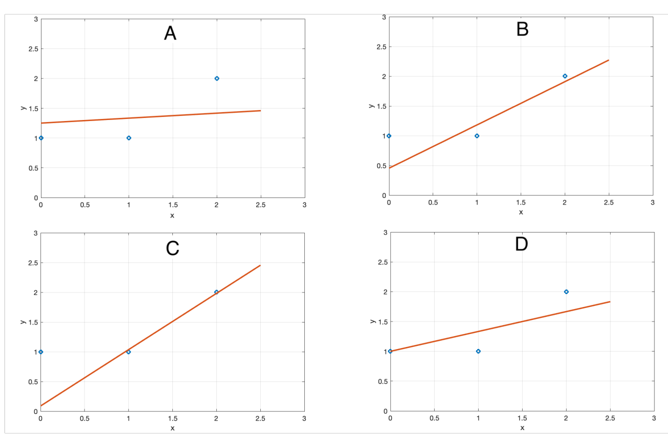
1 Long questions

Expect to see roughly 5 (at most 6) long questions in the exam. There will be 50 points in total for the exam. The long questions will roughly take up 30 out of 50 points.

1. You are a reviewer for the International Mega-Conference on Machine Learning for Everything, and you read papers with the following main claims. Would you accept or reject the paper? Provide a one-to-two sentence justification.
   (a) **accept / reject**. "My algorithm is better than yours since my algorithm has a lower in-sample error!"
   (b) **accept / reject**. "My algorithm is better than yours since my algorithm has a higher VC dimension!"
   (c) **accept / reject**. "My algorithm is better than yours since my algorithm has a low bias!"

2. 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 $D = \{(0,1), (1,1), (2,2)\}$. We applied the following regularizations (E_{in} is the squared error):
   (1) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda w_0^2$, where $\lambda = 1$
   (2) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda w_0^2$, where $\lambda = 10$
   (3) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda w_1^2$, where $\lambda = 1$
   (4) $E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda w_1^2$, where $\lambda = 10$

   The regression results are shown in the plots, however, we got confused about which plot corresponds to which regularization. Please assign each plot to one of the regularization and explain why.



3. Machine Learning Whiz Kid (MLWK) comes up to you with the following proposal for the Best Learner Ever (BLE). Given training dataset $\mathcal{D}$ with binary labels ± 1, BLE learns the following hypothesis $g(\mathbf{x})$: If $\mathbf{x} =$ some $\mathbf{x}_n \in \mathcal{D}$, then $g(\mathbf{x}) = y_n$, else $g(\mathbf{x}) = +1$. MLWK claims that since $E_{\text{in}} = 0$, as N gets large, BLE is guaranteed to get excellent generalization performance because of Hoeffding's inequality. Do you agree with MLWK? If not, then explain why not.

4. You work for Orange, a fictional maker of smartphones, and you have to develop a classifier that predicts whether some input fingerprint matches the fingerprint of a given phone's owner. Suppose classifying an input as $+1$ means that it matches while classifying an input as -1 means that it does not match. Through intensive market research, you know that everytime your classifier incorrectly says two fingerprints do not match when in fact they do, Orange loses 1 cent or 0.01 dollars. However, when your classifier incorrectly says two fingerprints match when in fact they don't, Orange loses 20 dollars.
   (a) Write down the cost matrix (see LFD 1.4.1) for this setting (you can assume that correct predictions incur 0 cost).
   (b) You decide to use logistic regression to predict the probability that an input fingerprint matches the phone owner's fingerprint. You train your model and get some hypothesis g. Suppose for some input, g tells you that the probability of a match is p. What is the expected cost (according to g) of classifying this input as $+1$? What is the expected cost of classifying it as -1?

5. Consider a one dimensional linear regression problem. The training set has N examples $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ where $x_n, y_n \in \mathbb{R}$. Suppose $\mathbf{w}^* = [w_0^*, w_1^*]^T$ minimizes $E_{\text{in}} = \frac{1}{N}\sum_{i=n}^{N}(y_n - w_0 - w_1 x_n)^2$. Which of the following four statements must be true, and why? (There could more than one true statement.)
   1. $\frac{1}{N}\sum_{n=1}^{N}(y_n - w_0^* - w_1^* x_n) y_n = 0$
   2. $\frac{1}{N}\sum_{n=1}^{N}(y_n - w_0^* - w_1^* x_n)(y_n - \overline{y}) = 0$
   3. $\frac{1}{N}\sum_{n=1}^{N}(y_n - w_0^* - w_1^* x_n)(x_n - \overline{x}) = 0$
   4. $\frac{1}{N}\sum_{n=1}^{N}(y_n - w_0^* - w_1^* x_n)(w_0^* + w_1^* x_n) = 0$

   Here $\overline{x}$ and $\overline{y}$ are the sample means in the training data. (Hint: Use what you know about $\nabla E_{\text{in}}(\mathbf{w}^*)$).

2 Multiple-choice questions

Expect to see roughly 10 multiple-choice questions in the exam. The multiple-choice questions will roughly take up 20 out of 50 points.

1. To show that the VC dimension of H is at least $d + 1$, what do we have to prove.
   - ◯ 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.
   - ◯ Every set of $d + 1$ points can be shattered by H.
   - ◯ Every set of $d + 1$ points cannot be shattered by H.

2. In performing updates, the perceptron algorithm does not take into account the distance of a correctly classified example from the current hypothesis $\mathbf{w}$.

   ◯ True

   ◯ False

3. The selection of the initial weight vector does not affect the final output of the perceptron algorithm.

   ◯ True

   ◯ False

4. Suppose I have a dataset with 1000 data points, and I am interested in performing a linear regression, and computing training and test error. For training set size K, I use the methodology of randomly selecting K training examples, and using the remaining $1000-K$ as my test set. In expectation, what would you expect to happen to my training and test errors as K increases?

   ◯ They both increase

   ◯ Training error increases and test error decreases

   ◯ Training error decreases and test error increases

   ◯ They both decrease

5. The VC-dimension of the family of finite unions of positive intervals (i.e., predict $+1$ if the point is inside one of the intervals, predict -1 otherwise) over the real line is

   ◯ 1

   ◯ 2

   ◯ 3

   ◯ ∞

6. In general for most cases, as we increase the amount of training data available to a learning algorithm, what happens to the bias of the learning algorithm?

   ◯ It increases

   ◯ It decreases

   ◯ It stays the same

   ◯ It first increases, then decreases