CS 189 Introduction to Machine Learning
Spring 2018 HW0

Your self-grade URL is http://eecs189.org/self_grade?question_ids=1_1,1_2,2,3_1,3_2,3_3,4_1,4_2,4_3,4_4,4_5,5.

This homework is due **Thursday, June 21 at 10 p.m.**

2 Sample Submission

Please submit a plain text file to the Gradescope programming assignment "Homework 0 Test Set":

1. Containing 5 rows, where each row has only one value "1".
2. No spaces or miscellaneous characters.
3. Name it "submission.txt".

3 Linear Regression and Adversarial Noise

In this question, we will investigate how the presence of noise in the data can adversely affect the model that we learn from it.

Suppose we obtain a training dataset consisting of n points (x_i, y_i) where $n \geq 2$. In case of no noise in the system, these set of points lie on a line given by $y = w_1 x + w_2$, i.e, for each i, $y_i = w_1 x_i + w_2$. The variable x is commonly referred to as the covariate [1] and y's are referred to as the observation. Suppose that all x_i's are distinct and non-zero. Our task is to estimate the slope w_1 and the y-intercept w_2 from the training data. We call the pair (w_1, w_2) as the true model.

Suppose that an adversary modifies our data by corrupting the observations and we now have the training data $(x_i, \tilde{y}_i)$ where $\tilde{y}_i = y_i + \epsilon_i$ and the noise ϵ_i is chosen by the adversary. Note that the adversary has access to the features x_i but *can not* modify them. Its goal is to trick us into learning a wrong model $(\hat{w}_1, \hat{w}_2)$ from the dataset $\{(x_i, \tilde{y}_i), i = 1, \ldots, n\}$. We denote by $(\hat{w}_1, \hat{w}_2)$ the model that we learn from this dataset $\{(x_i, \tilde{y}_i), i = 1, \ldots, n\}$ using the standard ordinary least-squares regression.

(a) Suppose that the adversary wants us to learn a particular wrong model (w_1^*, w_2^*). If we use standard ordinary least-squares regression, can the adversary *always* (for any choice of w_1^* and w_2^*) fool us by setting a particular value for exactly one ϵ_i (and leaving other observations as it is, i.e., $\tilde{y}_j = y_j, j \neq i$), so that we obtain $\hat{w}_1 = w_1^*$ and $\hat{w}_2 = w_2^*$? If yes, justify by providing a mathematical mechanism for the adversary to set the value of the noise term as a function of the dataset $\{(x_i, y_i), i = 1, \ldots, n\}$ and (w_1^*, w_2^*)? If no, provide a counter example.

[1] Besides covariate, some other names for x include feature, regressor, predictor.

Solution: The answer is no. Intuitively, the adversary is only in control of one degree of freedom. We now present a counterexample with just two data points.

Remark: Note that all x_i's are non-zero and distinct and any counter example should be using these two pieces of information. The reason behind assuming this information was to avoid two corner cases: (1) vertical lines (infinite slope), (2) rank deficiency of the covariate matrix (discussed in Food for thought below).

Let the two pairs of observations be $(x_1, y_1) = (1, 0), (x_2, y_2) = (-1, 1)$. Let $\mathbf{X}$ be the feature matrix:

$$\mathbf{X} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

and let $\mathbf{y}$ denote the observation vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The corrupted observation vector can take one of the forms $\tilde{\mathbf{y}} = \begin{bmatrix} 0 + \epsilon_1 \\ 1 \end{bmatrix}$ or $\tilde{\mathbf{y}} = \begin{bmatrix} 0 \\ 1 + \epsilon_2 \end{bmatrix}$. Standard OLS in this case simplifies to finding the solution to the equation $\mathbf{X}\mathbf{w} = \tilde{\mathbf{y}}$ and hence we have

$$\mathbf{X}\mathbf{w} = \tilde{\mathbf{y}} \Rightarrow \hat{\mathbf{w}} = \mathbf{X}^{-1}\tilde{\mathbf{y}} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\tilde{\mathbf{y}} \Rightarrow \hat{\mathbf{w}} = \frac{1}{2}\begin{bmatrix} \tilde{y}_1 - \tilde{y}_2 \\ \tilde{y}_1 + \tilde{y}_2 \end{bmatrix}.$$

Note that we can take inverse because $\mathbf{X}$ is a square matrix. For a more general case, refer to part (b). Thus the OLS solutions are of the form

$$\hat{\mathbf{w}} = \frac{1}{2}\begin{bmatrix} \epsilon_1 - 1 \\ \epsilon_1 + 1 \end{bmatrix} \quad \text{or} \quad \hat{\mathbf{w}} = \begin{bmatrix} -1 - \epsilon_2 \\ 1 + \epsilon_2 \end{bmatrix}$$

and it is immediately clear that this adversary **CANNOT fool us in choosing** $\hat{\mathbf{w}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ **or say** $\hat{\mathbf{w}} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$. In fact, there are infinitely many $\mathbf{w}^*$ that the adversary cannot trick us into learning. The reason being that, in both cases, if the adversary uses a certain value of noise to trick us to hallucinate a particular w_1^*, it can no longer choose w_2^* of its choice, as the OLS solution determines it automatically.

(b) Repeat part (a) for the case when the adversary can corrupt two observations, i.e., for the case when the adversary can set up at most two of the ϵ_i's to any non-zero values of its choice.

Solution: Yes. Intuitively, the adversary now has control over two degrees of freedom.

The adversary needs control of two points that have different x coordinates. Consider n points $\{(x_i, y_i)\}_{i=1}^n$, where linear regression recovers $\hat{w}_1, \hat{w}_2$. We can start by examining the closed form solution for least-squares. Without loss of generality, let us assume that the first two points are the ones that the adversary is going to want to corrupt. Let us construct the standard