HW4: Learning Concepts

Stat 154, Fall 2019

Problem 1

During lecture, we discussed the formula of the bias-variance decomposition. As we saw, given a data set $\mathcal{D}$ of n points, and a hypothesis $g(x)$, the expectation of the Squared Error for a given out-of-sample point x_o, over all possible training sets, is expressed as (assuming a noiseless target):

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(x_0) - f(x_0)\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(x_0) - \bar{g}(x_0)\right)^2\right]}_{\text{variance}} + \underbrace{\left[(\bar{g}(x_0) - f(x_0))^2\right]}_{\text{bias}^2}$$

The target function is represented by $f(x)$, and the average hypothesis is represented by $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]$.

Now, when there is noise in the data we have that: $y = f(x) + \epsilon$. If ϵ is a zero-mean noise random variable with variance σ^2, **show that the bias-variance decomposition becomes**:

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(x_o) - y_o\right)^2\right] = \text{bias}^2 + \text{var} + \sigma^2$$

Problem 2

This is problem 1, from secion 2.4 in ISL. For each of the following parts, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

a) The sample size n is extremely large, and the number of predictors p is small.

b) The number of predictors p is extremely large, and the number of observations n is small.

c) The relationship between the predictors and response is highly non-linear.

d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

Problem 3

This is problem 3, from section 2.4 in ISL. We now revisit the bias-variance decomposition.

a) Provide a sketch of typical (squared) bias, variance, training error, and test error, and irreducible error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

b) Exlpain why each of the five curves has the shape displayed in part (a).

Problem 4

This is problem 5, from section 2.4 in ISL. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred? Explain.

Problem 5

This is problem 6, from section 2.4 in ISL. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression (as opposed to a non- parametric approach)? What are its disadvantages?

Problem 6

Consider a simplified learning scenario similar to the one discussed in lab-04. Assume that the input dimension is one. Assume that the input variable x is uniformly distributed in the interval $[-1, 1]$. The training data set consists of 2 points $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$ and, the target signal function is $f(x) = x^2$. Also, assume that the output values have noise, that is: $y = f(x) + \epsilon$, where:

- $\epsilon \sim N(0, \sigma^2)$; zero mean and constant variance.

Consider two learning models:

1) For $\mathcal{H}_0$, we choose the constant hypothesis that best fits the data (the horizontal line at the midpoint $b = \frac{y_1+y_2}{2}$).

2) For $\mathcal{H}_1$, we choose the line that passes through the two data points (x_1, y_1) and (x_2, y_2).

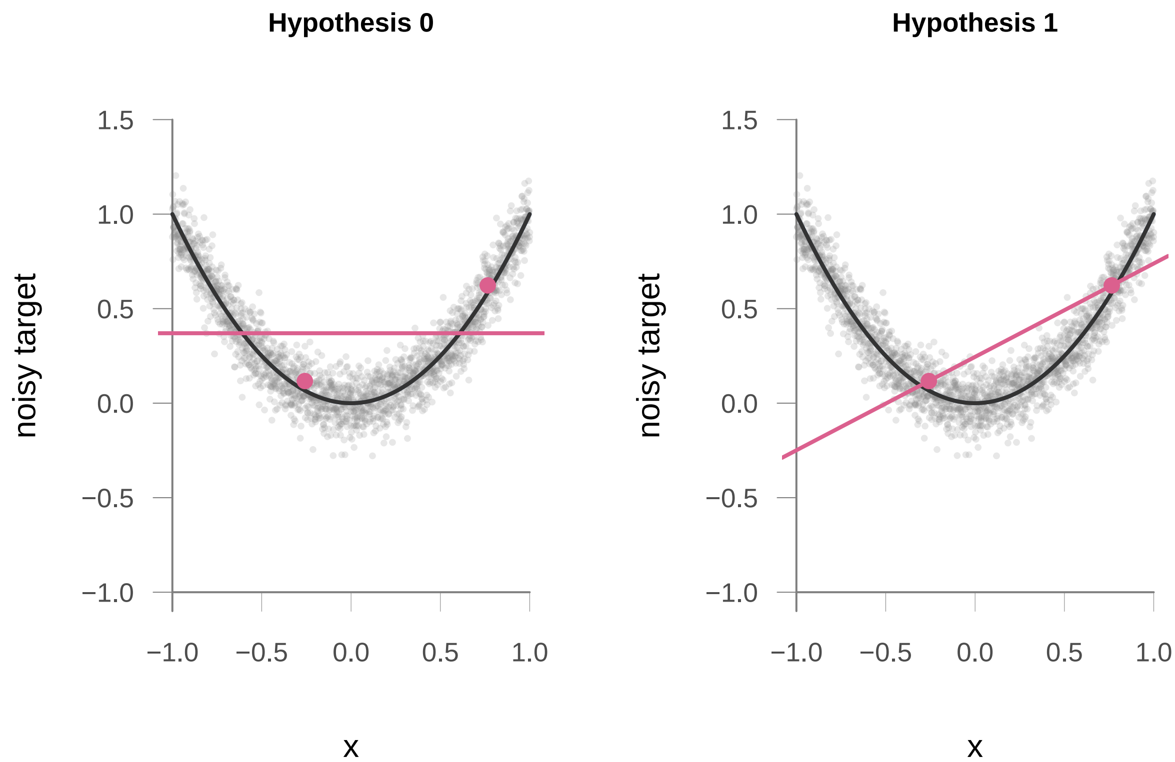

Figure 1: Hypothetical plots of the signal (in black), the noisy y-values (gray dots), and two training points with their fits (in red)

We are interested in assessing the performance of our learning systems in terms of the **overall expected out-of-sample MSE**, that is:

$$\text{overall expected test MSE} = \mathbb{E}_{\mathcal{X}} \left\{ \mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(x_0) - y_0 \right)^2 \right] \right\}$$

a) Describe an experiment that you could run to determine (numerically) $\bar{g}_0(x)$, $\bar{g}_1(x)$, as well as the overall expected test MSE for both hypotheses.

b) Run your experiment, by considering a noiseless scenario (i.e. $\sigma^2 = 0$), and a noisy scenario (i.e. $\sigma^2 \neq 0$).

c) Provide plots for each learning hypothesis, that let you visualize: 1) the signal, 2) some out-of-sample points, multiple fits (corresponding to different training sets), and the average hypothesis.

d) Report the results, and comment on what you obtained. Likewise, provide descriptions for each plot.