CMSC 478 Machine Learning - Spring 2019
Homework Assignment 8
Due at the start of class on May 2^{nd}

Bias/Variance: Consider a dataset generated by sampling x values uniformly from the interval $[-1, 1]$ where $y = f(x) = x^2$. The dataset contains two independent observations, i.e., $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\} = \{(x_1, x_1^2), (x_2, x_2^2)\}$. In this problem we will consider two different ways of fitting a line to this dataset and explore the bias and variance of each approach.

Recall that the expected value of a continuous random variable z is $\int_z p(z) * z$. This can be approximated by sampling a large number of values from $p(z)$ and computing their mean.

(a) Suppose our line if of the form $h(x) = b$, i.e., it is a constant function. We fit the line by setting $h_{\mathcal{D}}(x) = b = (y_1 + y_2)/2 = (x_1^2 + x_2^2)/2$. Describe in words how you could write a program that uses sampling to estimate the mean hypothesis, which is the expected value of b where the expectation is over all datasets:

$$\overline{h}(x) = E_{\mathcal{D}}[h_{\mathcal{D}}(x)]$$

Note that this is just a number because the value produced by any given hypothesis is a constant, independent of the value of x for which it is evaluated.

(b) Write that program and use it to estimate the mean hypothesis. There is no need to turn in code. Just give the single number which is the estimated mean hypothesis. Be sure to base your estimate on a large number of samples.

(c) Describe in words how you could write a program that uses sampling to estimate the bias, which is the expected squared difference between $\overline{h}(x)$ and $f(x)$ where the expectation is over all x:

$$E_x[(\overline{h}(x) - f(x))^2]$$

Note that this definition is different from the one we used in class. Rather than giving the bias as a function of x, we're going to compute the expectation over all x.

(d) Write that program and use it to estimate the bias. There is no need to turn in code. Just give the single number which is the bias. Be sure to base your estimate on a large number of samples.

(e) Describe in words how you could write a program that uses sampling to estimate the variance, which is the expected squared difference between the mean hypothesis and the hypothesis learned for a given dataset, where the expectation is over all x and all datasets $\mathcal{D}$:

$$E_x[E_{\mathcal{D}}[(h_{\mathcal{D}}(x) - \overline{h}(x))^2]]$$

To compute the value above you'll need to repeatedly sample both a dataset and a value for x independently.

(f) Write that program and use it to estimate the variance. There is no need to turn in code. Just give the single number which is the variance. Be sure to base your estimate on a large number of samples.

(g) Now suppose we instead consider a line of the form $h(x) = ax + b$ which we fit by selecting the line that passes through the two points. Modify your code to estimate the new $\overline{h}(x)$, bias, and variance. Comment on how the results change and offer an explanation for why they change in the way they do.

Note that the average hypothesis in this case is a line, not a constant. Given a number of lines in slope/intercept form, you can compute their average by averaging their slopes and their intercepts. It is sufficient to provide these average parameters for $\overline{h}(x)$.