Homework #5 (BST 263, Spring 2019)

1. In this exercise, you will investigate the bias of cross-validation (CV). Consider the simulated data generating process in the R code (`cv.r`) accompanying Lecture 9 (i.e., $X_i \sim \text{Uniform}(0, 5)$, and $Y_i \sim \mathcal{N}(\cos(X_i), 0.3^2)$ given X_i). You can copy chunks of code from `cv.r` (e.g., the univariate KNN regression function) to do the following exercises.

   (a) Generate a test set of 1000 points.

   (b) For $r = 1, \ldots, 25$:

      i. Generate 20 training points: $(x_1, y_1), \ldots, (x_{20}, y_{20})$.

      ii. For $n = 1, \ldots, 20$:

      Compute the MSE of KNN with $K = 1$ on the test set, using $(x_1, y_1), \ldots, (x_n, y_n)$ as the training set (that is, only use the first n training points). Let's call it $\widehat{\text{MSE}}(n, r)$.

   (c) For each $n = 1, \ldots, 20$, average the estimated test MSEs over the 25 runs:

$$\widehat{\text{MSE}}(n) = \frac{1}{25} \sum_{r=1}^{25} \widehat{\text{MSE}}(n, r),$$

   and plot $\widehat{\text{MSE}}(n)$ versus n.

   (d) Describe what trend you see as n increases, and explain why it makes sense. What was the point of averaging over 25 runs (that is, why not just plot one run)?

   (e) How does your plot explain why CV estimates are biased upward (i.e., why does CV tend to overestimate test MSE)?

   (f) From your plot, do you expect the bias of 2-fold CV to be higher when $n = 10$ or when $n = 20$? Why?

2. In this exercise, you will investigate the variance of cross-validation. Use the R code in the section of `cv.r` titled "Choosing the number of folds" to do the following exercises. Modify `nreps` and `K` (# neighbors in KNN) to be `nreps=1000` and `K=1`.

   (a) Plot the (estimated) variance of the CV estimates versus the number of folds. (Just remove the MSE and bias^2 from the current plot and fix the axes/labels.) Run the code five times, with a different training data set each time, by using 10, 20, 30, 40, and 50 for the random number generator seed. Show the five plots.

   (b) What is the (estimated) variance of the LOO-CV estimates? Is this a fluke, or does it make sense? Provide an explanation for what you observe.

   (c) Do you see anything surprising happening around `nfolds=10`? Can you come up with a conjecture for why this might be happening? (Just try your best.) Hint: Try running the code a few times with $n = 14$, $n = 16$, and $n = 18$.

3. Read ESL section 7.10.2, "The Wrong and Right Way to Do Cross-validation". Write a paragraph summarizing this section in your own words.