Introduction to Data Science Homework 1

This question will investigate the so-called "Bias-Variance Tradeoff" using a simple model on a simple 1-D data set. Bias-Variance Tradeoff is a crucial concept in data science, and is useful for selecting an appropriate level of model complexity for a problem (and avoid over/under fitting the data). We will find ourselves revisiting this concept again and again as we progress in this course, so we might as well get a simple demonstration of it now.

1. This homework comes with two data files, for training and test data sets respectively. Load the data from these files, and make scatter plots for training and test data sets.

   (Although not directly related to this question, the program used to generate these data sets are also provided. You can take a look if you want.)

2. Fit polynomial of order n to the training data. (This is a straight line for $n = 1$, parabola for $n = 2$, and so on). Use values of n in the range $n = 1, 2, ..., 10$. Plot the training data along with these best-fit polynomials.

   (Hint: You should find that the polynomial fit gets increasingly wilder as n gets larger. Also, you should not have to do the fit from scratch; there are functions in python that you can simply use.)

3. The Mean Squared Error, defined as MSE $= \sum(y_{\text{data}} - y_{\text{model}})^2$, is a measure of the goodness of fit. In other words, a model with small value of MSE is said to "fit the data well".

   For each value of n, calculate the Mean Squared Error (MSE) of the polynomial fit with respect to the *training* data. Plot the MSE as a function of n.

   (Hint: You should find that the MSE in this case is a decreasing function of n in this case. This is by design: the polynomial fit seeks to minimize the MSE of training data in the first place, and higher-order polynomial can wiggle more, therefore giving better fit.)

4. Now, use the polynomial fits that you derived in part 2, and calculate the MSE of these polynomials with respect to the *test* data. Plot the MSE as a function of n.

5. Explain the behavior you found in part 4. What happens at low values of n? How about at high values of n? What value of n is the most appropriate in fitting the data?