Homework 1b: Linear Regression part 2.
EE425X - Machine Learning: A Signal Processing Perspective

Homework 1 focused on learning the parameter θ for linear regression. In this homework we will try to understand how to use that information to predict the output for a given query input. We will also understand bias-variance tradeoff and how to decide the model dimension This HW will use a lot of the code form the previous one.

1. Generate $m + m_{test}$ data points satisfying

$$y = \theta^T \boldsymbol{x} + e$$

with θ being ONE fixed n length vector for all of them. Set $n = 100$. Set the values of $\theta = [100, -99, 98, -97...1]'$. Use $E[e^2] = 0.01||\theta||_2^2$. Do this for $n = 100$.

Now suppose you have only $m = 80$ data points.

a. Try to learn θ and explain what happens. Report both the estimation error in θ, $||\theta - \hat{\theta}||_2^2/||\theta||^2$ and a "Monte Carlo estimate" of the prediction error on the test data (test data MSE).

$$\text{Test-MSE} := \mathbb{E}[(y - \hat{\boldsymbol{y}})^2]$$

Compute above by computing $\hat{\boldsymbol{y}} = \hat{\theta}^T \boldsymbol{x}$ for each test data vector (also called query) and computing $\frac{1}{m_{test}} \sum_{i=1}^{m_{test}} (y_i - \hat{\boldsymbol{y}}_i)^2$.

b. What you will conclude is you cannot learn θ correctly because m is even smaller than n.

Let us assume you do not have the option to increase m. What can you do? All you can do is reduce n.

Do an experiment where you start with $n_{small} = 1$ and keep increasing its value, and each time compute Test-MSE. Obtain a plot. Use the plot and what you learn in class to decide what value of n_{small} is best.

Interpret your results based on the Bias-Variance tradeoff discussion. See Section 11 of Summary-Notes.