Homework 8

Math 87

Due: November 10 2023

1 How effective is regression?

In this exercise, you will investigate the quality of an linear least squares regression as a function of the amount of noise in your data.

- Consider a linear equation $y = mx + b$. Precisely explain how one could randomly sample points satisfying this equation, under the constraints that the samples $0 \leq x \leq N$ for some N. Suppose you then performed least linear squares regression on the data. With enough samples, you should be able to recover the equation of the line. What is the minimum number of samples needed in order to do this?
- Suppose that we have introduced some noise to our equation, i.e. our samples are drawn from $y = mx + b + \eta$, where η is drawn from a normal distribution with mean 0 and variance σ. One would hope that, with enough samples, one could recover the equation of the original line. Predict how σ affects the number of samples required to approximate the line.
- For concreteness, let $m = 3, b = 4, M = 10$. Sample and plot (on separate plots) 100 points sampled from $y = mx + b + \eta$ for $\sigma = 0.01, 0.1, 1, 5, 30$ and plot the original (i.e. no noise) line over the data.
- Perform least linear squares on the data obtained in the preceding part. The solution will give you the coefficients of a line. On each plot, plot this line over the data.
- Run an experiment to determine how many samples are needed in order for the least squares solution to be within 0.01 of the coefficients of the original line for each value of σ (you do not need to find the minimal such value, just some number of samples that works). For each value of σ, plot the error in slope (i.e. the difference between the slope of the original line and the one computed by least squares regression) for the following number of samples: $20, 200, 2000, 20000, 200000$.

2 A polynomial fitting problem

In class, we showed how to extend linear regression to quadratic curve fitting. In this problem, we will apply this to higher order polynomials.

- Determine a general matrix equation formula for fitting a series of observations to an n degree polynomial.
- To test your formula out, plot the data set with Python and determine best fit polynomials of degree 1,2,3,4 and 5. Record the coefficients of these polynomials, as well as the residual error at each point (i.e. the difference in your predicted value and the true value).

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	3.66711	4.06035	4.41999	4.69781	4.82905	4.54028	4.38693	3.17306	1.40012	-1.0748