ENGN 2520 Spring 2022
Homework 1

TURNING IN: Upload a PDF file to the course website.

IMPORTANT: Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently. All of the work submitted should be your own. Each student should write on the problem set the set of people with whom they collaborated.

Problem 1

Note that (Q1) and (Q2) below are hypothetical questions. You should answer questions (a), (b), (c) and (d).

Alice and Bob are working together to estimate a function mapping the chemical composition of a solar array to the power output. To collect training data Alice and Bob work in turns in the lab, making new compositions and measuring power output. Suppose Bob is not as carefull as Alice in making his measurements. This leads to some questions:

(Q1) Should Alice and Bob ignore Bob's measurements in estimating their function?

(Q2) How can they incorporate both sets of measurements in a reasonable way?

We can capture the situation with a simple mathematical model. Let $f_w : X \to \mathbb{R}$ be a function defined by a feature map $\phi : X \to \mathbb{R}^M$ and a vector of parameters $w \in \mathbb{R}^M$,

$$f_w(x) = w^T \phi(x)$$

Let T_A and T_B be two sets of training examples. We assume the errors in the training examples are independent but are larger in T_B compared to T_A.

For (x, y) in T_A we assume $y = f_w(x) + e$ with error e distributed according to a Normal distribution $N(0, \sigma_A^2)$. For (x, y) in T_B we assume $y = f_w(x) + e$ with error e distributed according to a Normal distribution $N(0, \sigma_B^2)$. The errors are independent and $\sigma_B^2 > \sigma_A^2$.

Suppose we know σ_A^2 and σ_B^2. What is the maximum likelihood estimate of w?

$$w_{\text{ML}} = \max_w p(T_A, T_B | w)$$

(a) Show that w_{ML} minimizes a sum of *weighted* squared differences. The sum should have one term per example in T_A and one term per example in T_B.

(b) Show how to compute w_{ML} by solving a linear system.

(c) What does this mathematical model say about questions (Q1) and (Q2) above?

(d) Suppose we don't know σ_A and σ_B. How can we estimate w?

Problem 2

In this assignment you will implement the linear basis method for regression with least squares error. The data for this problem is available on the course website. You will estimate functions using the training data and evaluate their generalization on the test data.

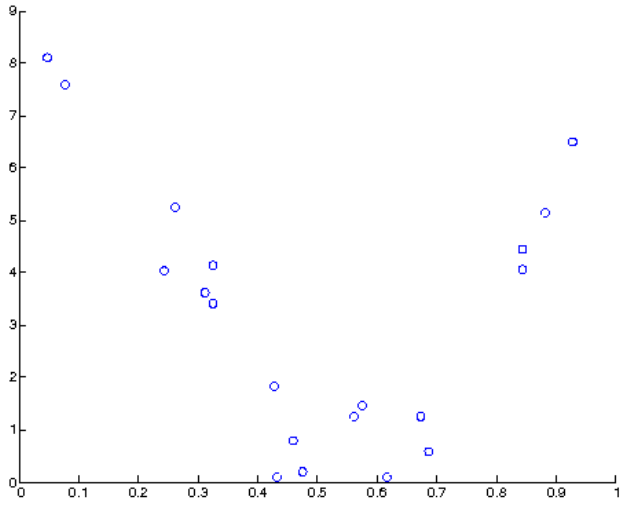
You should solve the problem by setting up a linear system of equations $Aw = b$, where the solution to the linear equations minimizes the sum of squared errors in the training data. To solve a linear system in Matlab you can use `linsolve`. In python you can use `numpy.linalg.solve`.

The picture below was generated by loading the training data into Matlab and plotting the sample points using the command scatter as follows:

> load Xtrain;

> load Ytrain;

> scatter(Xtrain, Ytrain);



For this problem you should use polynomial basis functions:
$\phi_1(x) = 1$, $\phi_2(x) = x$, $\phi_3(x) = x^2$, $\phi_4(x) = x^3$, ...

$$f_w(x) = \sum_{i=1}^{M} w_i \phi_i(x)$$

Let $T = \{(x_1, y_1), \ldots, (x_N, y_N))\}$ be a set with N examples.

The loss of f_w on T is

$$L(f_w, T) = \frac{\sum_{i=1}^{N} (f_w(x_i) - y_i)^2}{N} \tag{1}$$

(a) Use the *training* set to estimate polynomials of degrees 1 through 10. For each degree report the loss of the estimated function on the *training* and *test* sets. Note that a polynomial of degree d is represented using $d + 1$ basis functions.

Submit a plot showing the training and test loss as a function of the degree.

Do you observe overfitting?

(b) Submit a plot showing the training data and the degree 3 polynomial you estimated from that data. Make sure to plot the polynomial by evaluating it on a dense set of x values and not just the x values in the training data.

(c) Submit a plot showing the training data and the degree 10 polynomial you estimated from that data.

(d) Submit your source code along with your homework.