 Dapeng Shang

MF850: Advanced Computational Methods

Problem Set 2

Due date: See Blackboard.

Instructions: You submit on Blackboard. You may solve this assigment in groups of two. A submission is constituted by answers to the problems along with the code used. A file called `hw2.{jl,R,py}` should contain your code, or your entry point if you separate your code into multiple files. This file should run without errors from a fresh instance/REPL. In other words, submissions in notebook format are not accepted (but you may of course develop in them before creating the submission).

Please contact the instructor or a TA if you have questions regarding these instructions or if you find the problem formulation unclear.

Problem 2.1 Linear regression is a parametric model. This means that the model captures all information needed to make predictions in a finite set of parameters. For linear regression, this is true because the trained parameters do not depend on the point at which the predictor is evaluated. Indeed, given training data (y, X),

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

and the prediction in a point x is,

$$\hat{\beta} \cdot x,$$

which shows that the parameters of the model do not depend on x.

The goal of this homework is to explore a non-parametric method called locally weighted linear regression and practice parameter tuning.

The CSV files `data{1,2}.csv` contain four columns of data: training data X_{train} (one-dimensional) and y_{train} as well as validation data X_{val} and y_{val}. In both files, the training data is the same, but the range of the X data differs for the validation set. A third dataset is provided for students curious to experiment further, but is not needed to complete the homework.

Consider the regression specification

$$y \sim \beta_0 + \beta_1 x + \cdots + \beta_p x^p$$

for varying values of p.

(a) Fit the regression on `data1.csv` and plot the MSE on the validation set as a function of p. Make sure the range is large enough to capture a (local) minumum.

For what p is the MSE minimized and what is the MSE?

Remark: Compare the minimizing p to the result you get if you minimize the MSE on the training set.

(b) Do the same thing on `data2.csv`. What changes do you observe? How do the optimal p and MSE change?

Now consider a *weighted least squares regression* using weights w by minimizing

$$\sum_{i=1}^{N} w_i (y_i - x_i \cdot \beta_w)^2 = \|\text{diag}(w_i)(y - X\beta_w)\|_2^2,$$

with $\beta_w = (\beta_0, \ldots, \beta_p)$.

(c) Derive a formula for the optimizer $\hat{\beta}_w$.

Hint: Compute the gradient with respect to the coefficients and solve the first order condition equation.

We implement *locally weighted least squares* by making the weights depend on the point x at which we wish to estimate: $w_i = w_i(x)$. In other words, for each x where we want to make a prediction, a separate set of parameters $\hat{\beta}_{w(x)}$ is fitted. For simplicity, consider the weights

$$w_i(x) = \exp(-(x_i - x)^2/2\tau),$$

where τ is a parameter referred to as the *bandwidth*. It is sufficient to use one feature (equal to x) in addition to the intercept, but you may experiment with higher powers.

(d) Use the formula to compute the locally weighted least squares MSE error on the validation set using `data1.csv`.

Plot the error as a function of τ. Make sure the range is large enough to capture a (local) minumum.

For what value of τ is the validation error minimized and what is the validation MSE?

Remark: If you could not find the formula, you may use a library to implement each weighted regression.

(e) Do the same thing on `data2.csv`. What changes do you observe? How do the optimal τ and MSE change? How are the changes different to how the linear regression changed between the two data sets?

This type of technique is a very popular data smoothing method.

