ENGR 421 / Homework 4: Decision Tree Algorithm

Umur Berkay Karakaş

April 26, 2021

In homework 4, we are given 272 data points abot the duration of the eruption and waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park.

First, I created xtraining variable from the first 150 data points in the data set and xtest variable from the last 122 points in the data set. I assigned their corresponding class labels to ytraining and ytest. I also calculated number of data points and minimum value and maximum value of X.

Then, I defined a learnTree function which takes xtrain, ytrain and P as inputs. For each node, if the node has less than P data points, the node is terminal. Otherwise, we pick the best split by calculating average error. After selecting the best split, we create the new right and left children and move on to the next index for the next iteration. Also, I defined a predict function which predicts the values of a given input x in the fit by using learn tree algorithm.

For P = 25, I got the following plot:

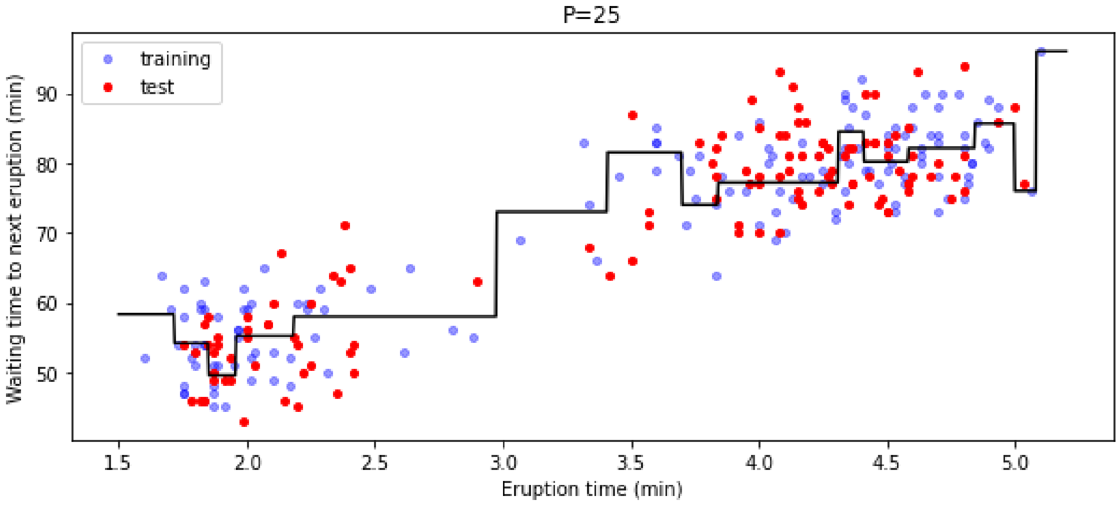

Figure 1: Decision tree fit plot

I calculated RMSE for P = 25, and I got the value which was given in the PDF. Then, I learned trees for P = 5,10,15,..,50 and calculated their RMSEs and got the following plot:

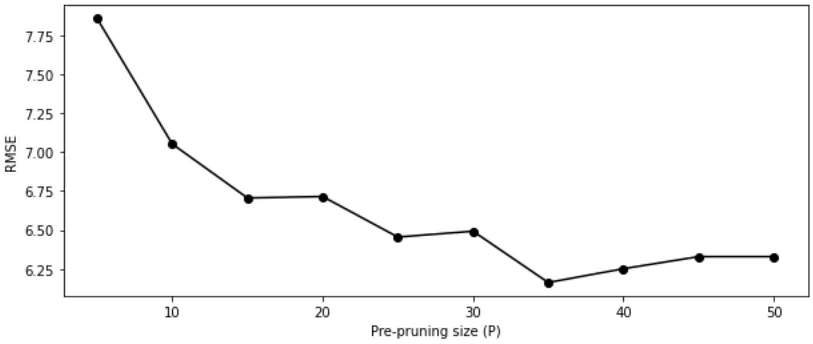

Figure 2: RMSE plot