ENGR 421 - Machine Learning

Mehmet Ustek

HW5 Report

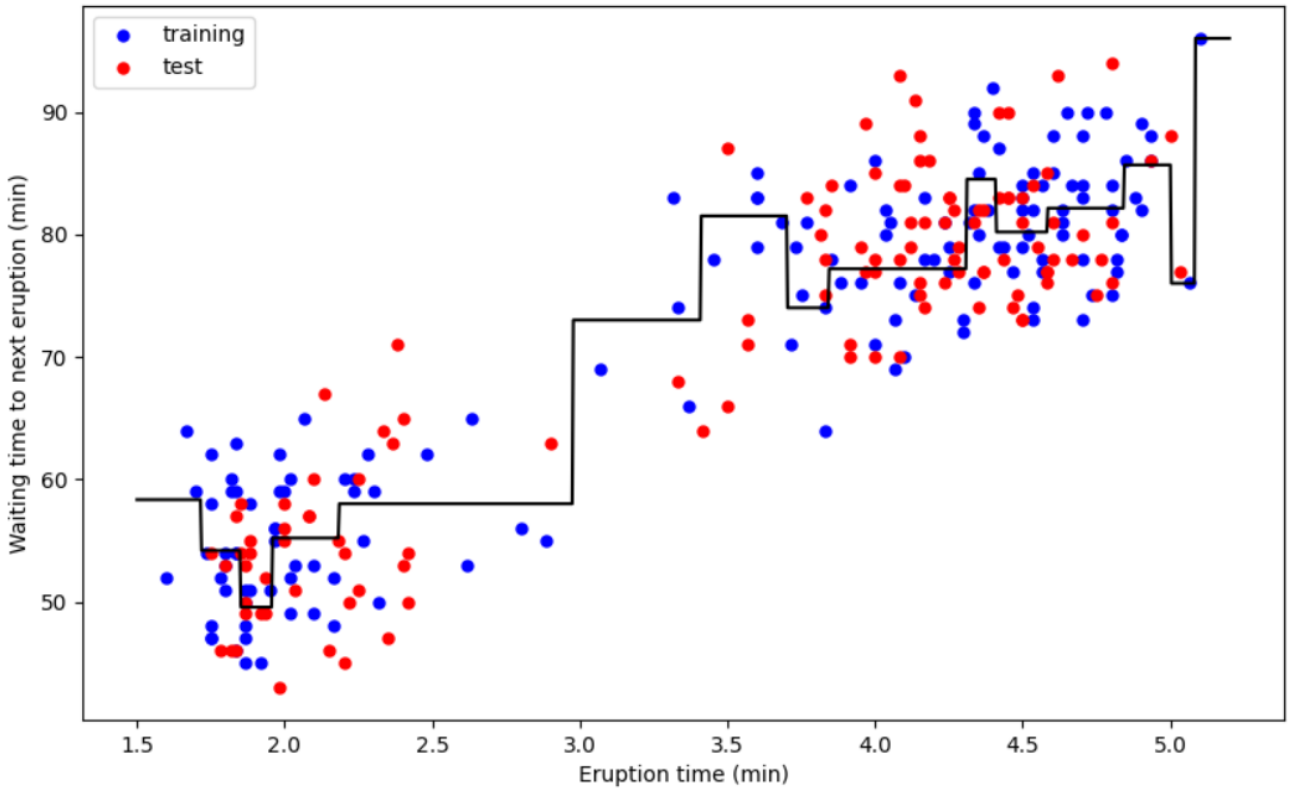
Completeness of the project:
All requirements are done with correct outputs regarding the homework description.

First of all, I divided the data into sizes 150, 122 for training and testing respectively. Then I created the learn(P) function that takes pre-pruning constant P. I modified the lab code to accept this change of pre-pruning. The incoming node size should be less than this constant. Thus, I made the following change in the code:
if len(data_indices) <= P
Furthermore, if a node is terminal, then I need to get the mean of node values in that threshold. For example, if my dataset is 3,4,5,6, and I split the data into two halves with one half carrying data 3,4 and the other 5,6, the mean of these two halves will be 3.5 and 5.5 respectively. Since I am not purifying the whole data points as we did it into decision trees, I need to have this mean value to achieve the regression tree value evaluation.
Having completed these, I changed the split_scores function. As discussed in class, I applied squared error. My implementation is as follows:

- Get mean of left indices
- Get mean of right indices
- For each point in left_indices, find and add $[y(i)-y\_hat(i)]^2$ to the total sum.

The rest of the code for the learn(P) function is the same as lab session, except for now we have only one feature.

Then I created a ‘predict’ function, which takes is_terminal, node_splits and means as parameters. It basically stops and returns means[index] if the node is a terminal node. If not, it goes to the children in depth until it finds a terminal node for predicting the y value.
Having completed these alterations, I plot my data with y_predicted values list as we did it in hw04. The plotted data is as follows:

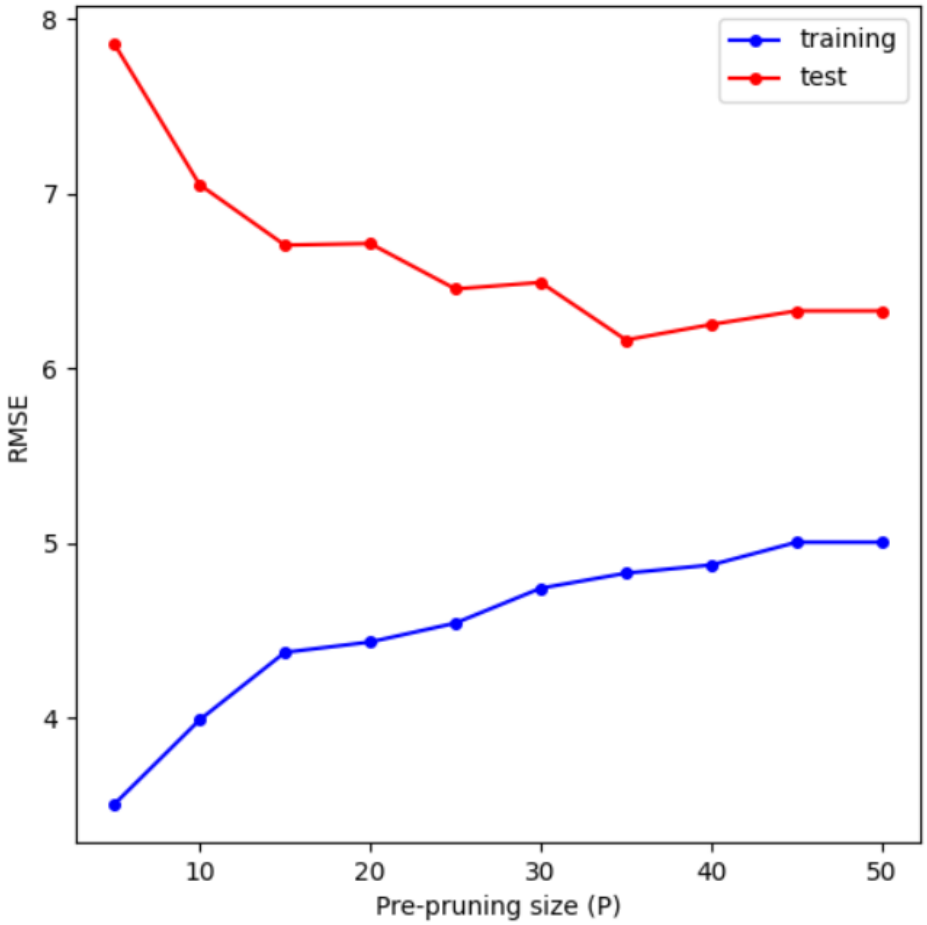


Then, I calculated RMSE first with P = 25, and then for values ranging from 5 to 50. The result for P = 25, was exactly the same as hw description, and the values are as follows:
RMSE on training set is 4.541214189194451 when P is 25
RMSE on test set is 6.454083413352087 when P is 25

The values for training and test data ranging from 5 to 50 is as follows:
RMSE on test set is 7.857603084243197 when P is 5
RMSE on training set is 3.9865873007973 when P is 10
RMSE on test set is 7.051576571621315 when P is 10
RMSE on training set is 4.373539502719482 when P is 15
RMSE on test set is 6.705082196461261 when P is 15
RMSE on training set is 4.432917644878319 when P is 20
RMSE on test set is 6.714228091718779 when P is 20
RMSE on training set is 4.541214189194451 when P is 25
RMSE on test set is 6.454083413352087 when P is 25
RMSE on training set is 4.740887430745958 when P is 30
RMSE on test set is 6.491555539784374 when P is 30
RMSE on training set is 4.826960020838137 when P is 35
RMSE on test set is 6.162109543315439 when P is 35
RMSE on training set is 4.874194088972034 when P is 40
RMSE on test set is 6.2508380958776035 when P is 40
RMSE on training set is 5.004146782988638 when P is 45
RMSE on test set is 6.328291469197311 when P is 45

RMSE on training set is 5.004146782988638 when P is 50
RMSE on test set is 6.328291469197311 when P is 50

And the graph for these values which is exactly the same as hw description is as follows:



Overall, this homework contributed to my understanding of decision trees and regression trees. I solidified my learning on implementation-wise and theoretical-wise decision tree concepts.

Acknowledgements:
I understand the university rules for plagiarism and I have never shared or used any code or slice of code while doing this project. Thus, the effort belongs only to me.