ENGR 421 - HOMEWORK 4
Zeynep Öner
ID: 64912

Decision Tree Regression

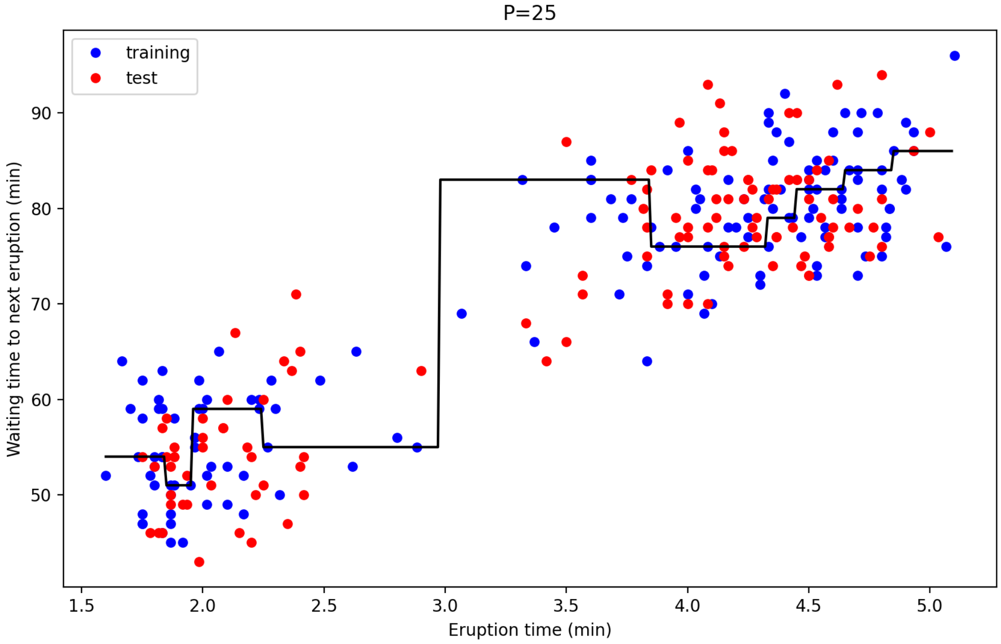
This homework was an implementation of the decision tree algorithm with a pre-pruning parameter P. After running the tree with P = 25, we also try different pre-pruning values and compare RMSEs.

The decision tree is created with 5 dictionaries where we store node indices, whether a node is terminal, whether a node needs splitting, the best split for a node, and a node's frequency. We first initialize the root node as a node that is not terminal and needs splitting. In an infinite loop, we determine all the nodes that need splitting. We break when we have no nodes left to split. We then go through all the nodes that need splitting and determine the best splits for the corresponding nodes. We use the impurity of a split equation we discussed in class, where:

$$-\sum_{s=1}^{S} \frac{N_{ms}}{N_m} \sum_{c=1}^{K} P_{msc} \log_2 P_{msc} \tag{0.1}$$

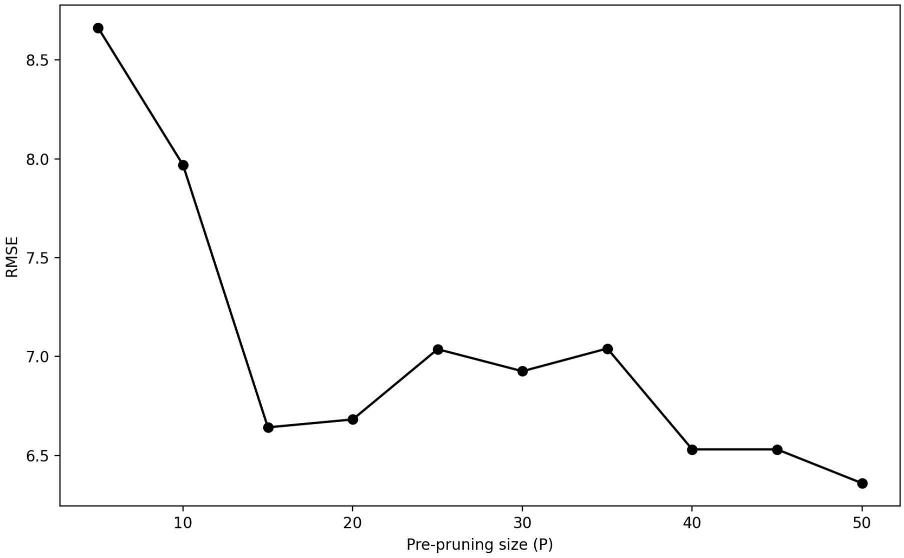
The term inside the class summation is entropy.

The obtained graph from this decision tree containing our training data, test data and estimation of y is below:



Although the graph is a little different from the one in the description, it may be classified as a good algorithm since the RMSE for P=25 is 0.7368. I think it is wrong for me to use the entropy for the impurity calculation, but I tried misclassification error which resulted in worse estimations. Therefore, I stuck with entropy.

The graph that depicts different pre-pruning parameters vs RMSE values is below:



The graph is again a little off, especially with the RMSE calculations for P between 20 and 40. The graph in the description hints at a steady decline of RMSE as P increases, then an increase after 35. We do not see this happening in my graph. Other than that, it does show the fact that overall, RMSE decreases as P increases.