ENGR 421

Homework 05: Decision Tree Regression

Deadline: December 2, 2019, 11:59 PM

Sitare Arslantürk, 57677

In this assignment, Decision Tree Regression for a univariate data set is implemented. The data is divided into two parts for training and testing. The decision tree model is fitted. A regression tree is constructed similar to the classification tree we did on the lab sessions. The impurity measure is replaced by a measure appropriate for regression. In regression, the goodness of a split is measured by the mean square error from $g_m$which is the estimated value in node m. Let's denote X_m is the subset of X reaching node m. In a node, I use the mean of the outputs reaching the node. *Figure 1* shows the overall graphical idea of the decision tree I created.

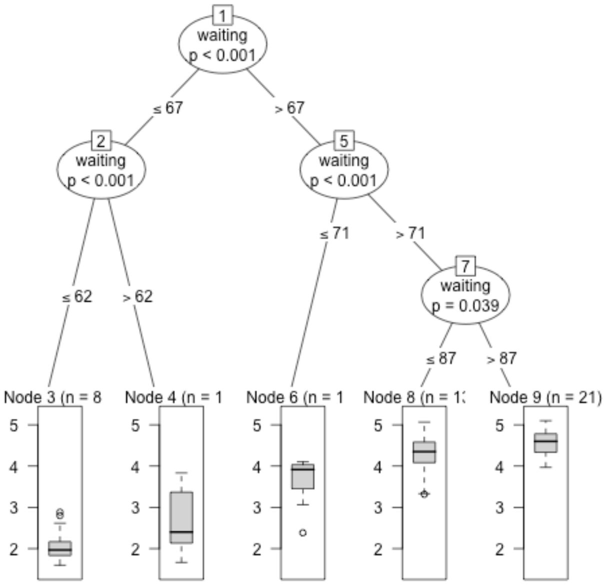
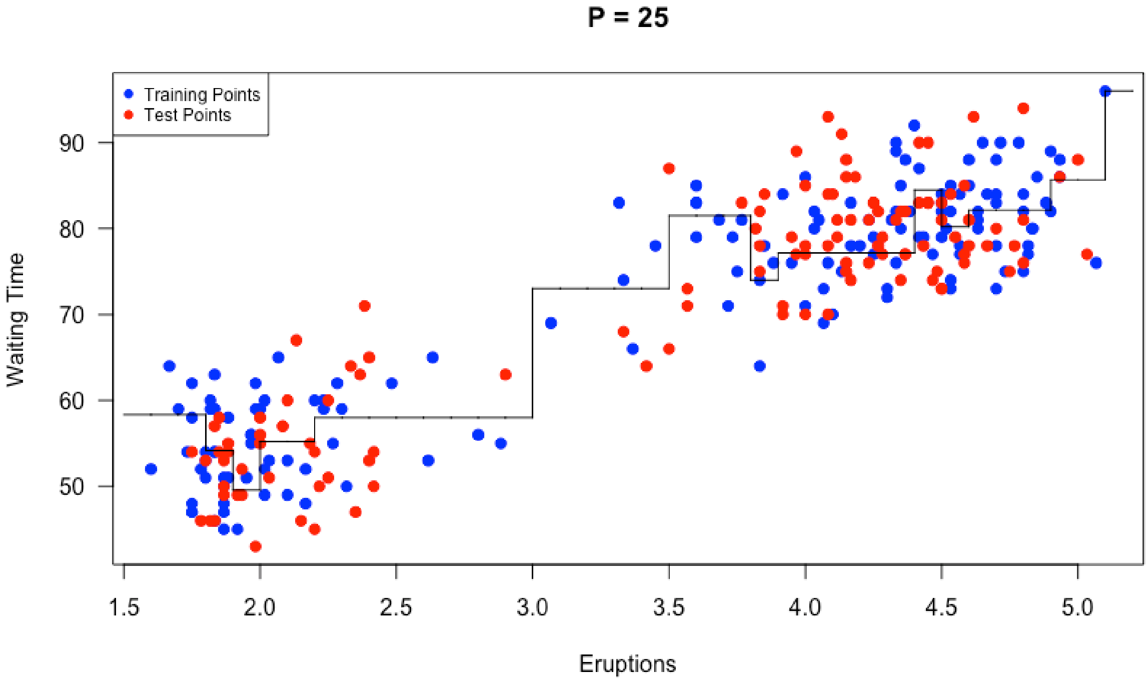


Figure 1

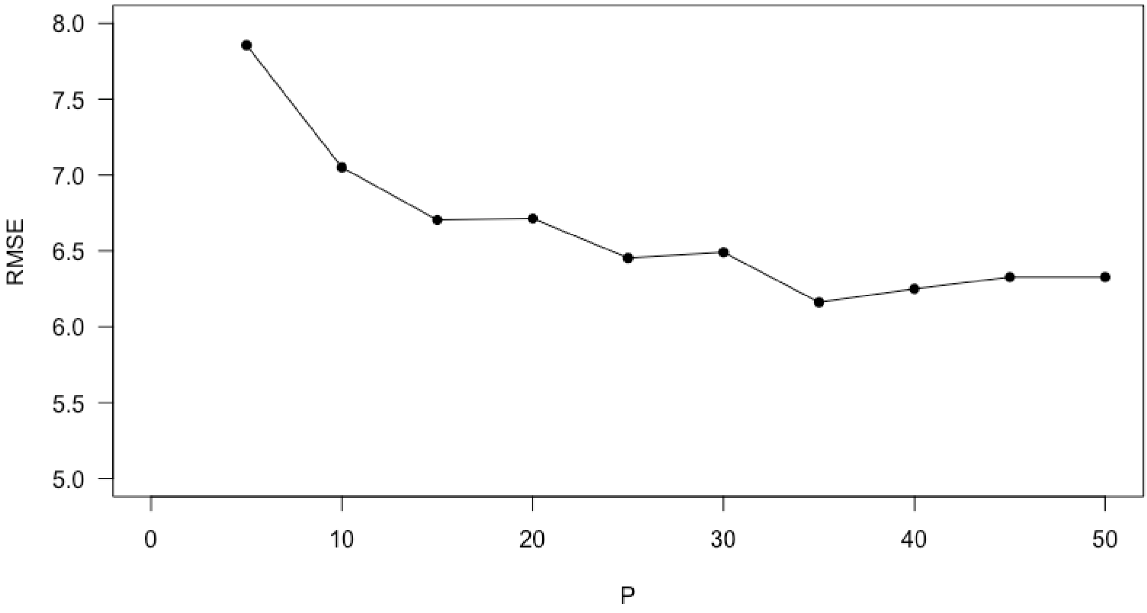
If at a node $E_m < \theta_r$, then a leaf node is created and it stores the $g_m$value. If the error is bigger, the points that reach node m is split further such that the sum of the errors in the branches is minimum. At each node, I look for split threshold that minimizes the error then loop recursively.

$$E_m = \frac{1}{N_m} \sum_t (r^t - g_m)^2 b_m(x^t) \text{ where } N_m = |X_m|$$

A node is not split further if the number of training instances reaching a node is smaller than a certain number P in our case 25, regardless of the impurity or error. A decision based on few instances cause variance and generalization error. Pre-pruning is stopping tree construction early before fully created. In order to prevent the tree from overfitting, I set the pre-pruning parameter P. First, I calculate with P set to 25 and then we calculate based on different P numbers starting from 5 to 50. *Graph 1* depicts the graph of eruptions vs waiting time when the pruning parameter is set to 25. *Graph 2* depicts the pruning parameter vs root mean squared error for P = 5 to P = 50.



Graph 1



Graph 2