DASC-521
Homework 05: Decision Tree Regression

Ender Erkaya

November 2021

1 Introduction

In this homework, we apply decision trees for a regression problem. We use a univariate data set, hence our selected features in all nodes are same as x. For classification problems, we measure impurities on each node to decide the best split. As impurity measures, we used entropy, gini index or misclassification error and decide best split as minimum impurity. Here, in regression problem, instead classification impurity measures we can measure the performance of the split by mean-squared-error. On the contrary to classification problem, we decide y values of the terminal nodes as the mean of its node indices. Different from the lab, we also used pre-pruning parameter to restrict complexity. At the end, we measure rmse values for both test and training data for different p values.

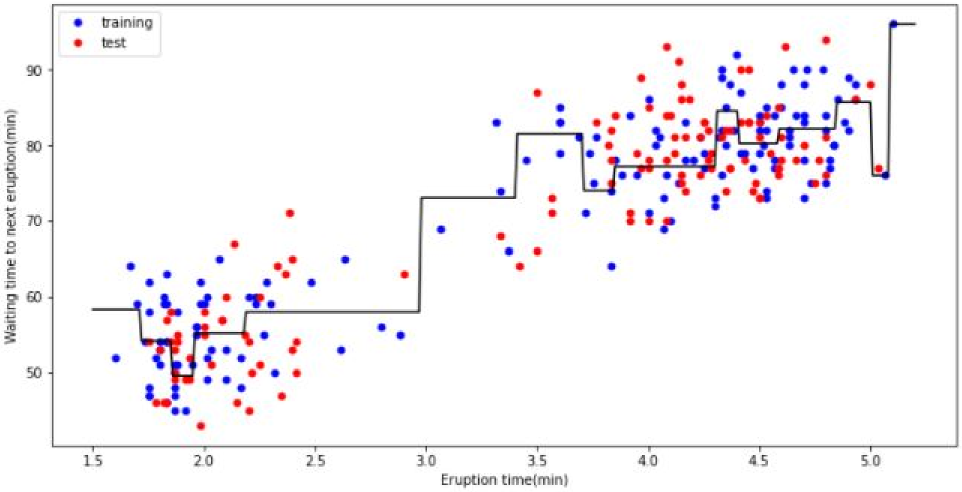


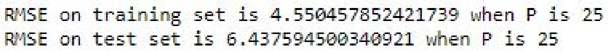
Figure 1: Regression

2 Decision Tree Regressor Visualization

The regression obtained by decision tree together with the observed training and test data is given in the figure 1.

3 RMSE Performance

RMSE performance of the decision tree regressor on the training and test data is displayed as below:



```
RMSE on training set is 4.550457852421739 when P is 25
RMSE on test set is 6.437594500340921 when P is 25
```

Figure 2: RMSE

4 RMSE vs Pre-pruning Parameter

We applied different pre-pruning parameter values p and measure the rmse performance on both training and test data. The obtained plot is given below: **??**

5 Conclusion

To apply decision tree model for regression problem, we used mean squared error measure to decide the best split, where the predicted y values are the means of the data indices of the split. Also, we worked on univariate data for this problem. As p increases we see the inrease of the performance in test data. This adjustment overcomes overfitting problem due to high p value. Healing in overfitting brings better generalization, less fitting to training data as expected.

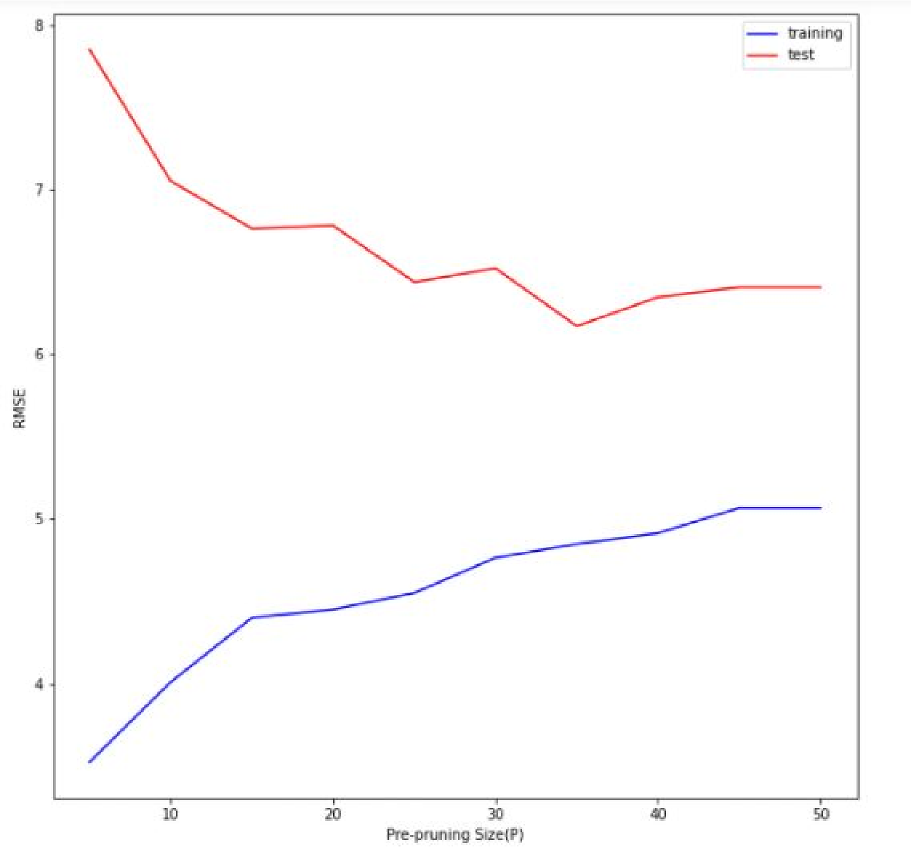


Figure 3: RMSE vs Pre-pruning parameter