EC500 - Introduction to Learning From Data

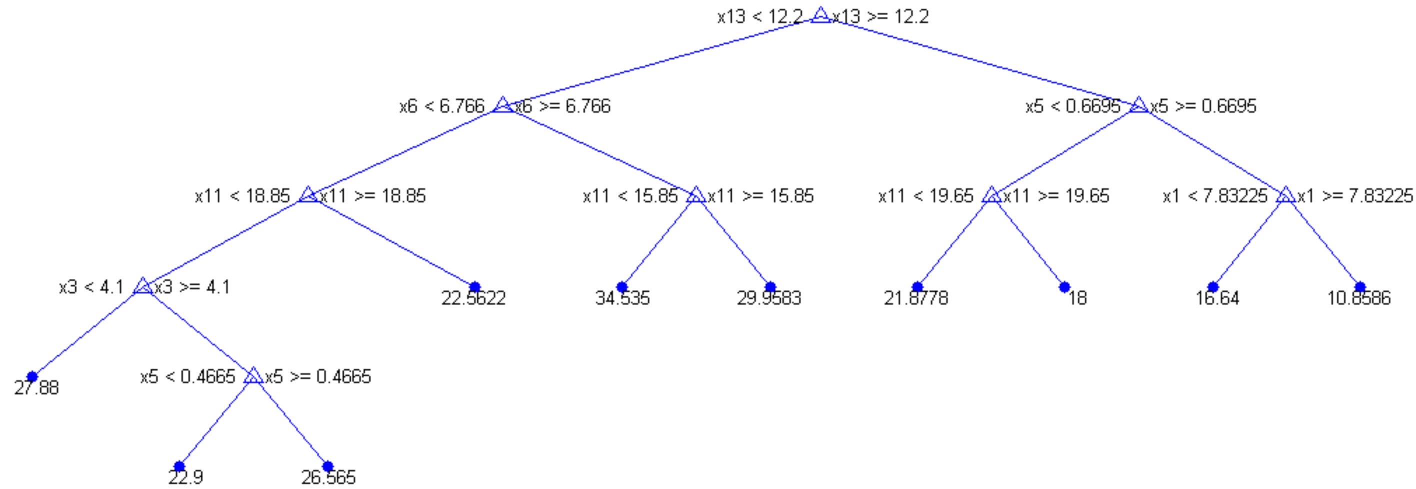
Matlab 3

Mikhail Andreev

November 9, 2015

1 Exploring Boston Housing Data with Regression Trees

The regression tree generated from the training data can be seen here:

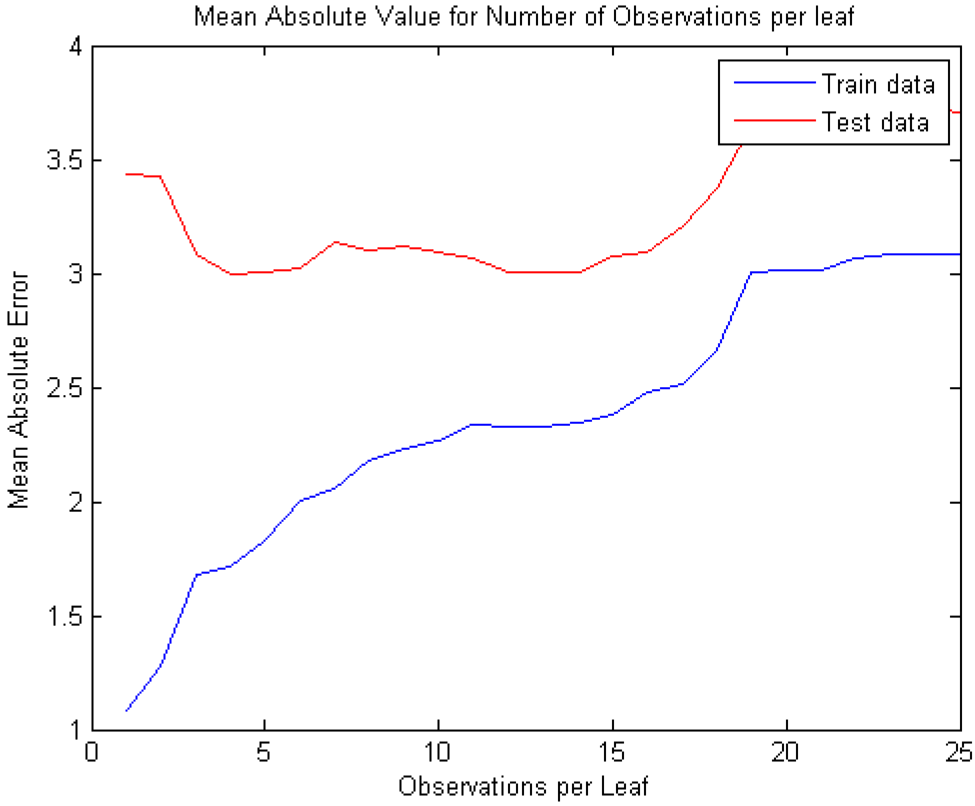


Using the input vector:

$[5, 18, 2.31, 1, 0.5440, 2, 64, 3.7, 1, 300, 15, 390, 10]$

We get an output $MEDV = 22.047619$.

If we plot the Mean Absolute Error for both the training and testing sets, using different numbers of observations per leaf, you get the following graph:



From the graph we can see that when the number of observations are low, we essentially memorize the training data, making the training error almost 0, while having a high testing error. As we increase the number of observations, training error starts to steadily increase, while the testing error drops to a lower plateau. For a large number of observations, the testing error is constant, meaning the number of observations is not a deciding factor in the accuracy of the tree in this case. However, after a certain point, each leaf is making too many observations, generalizing the results too much, causing a large increase in both training and testing error.

2 Ordinary Least Squares versus Robust Linear Regression

When implementing the Ordinary Least Squares method, the input data matrix will in general yield a unique solution. A different input data matrix will return new values of W and b. This occurs because when determining the values of W, they are multiples of the input data matrix.

The value of $w_{OLS} = 1.2476$. The value of $b_{OLS} = 2.528$. The calculated value of $MSE = 1.390425$. The calculated value of $MAD = 0.965729$.

3 Overfitting and Ridge Regression

4 Lasso vs Ridge