

Predicting Boston Housing Price

Kosuke Fukui

1. Statistical Analysis and Data Exploration

1.1 Statistics

Number of data points: 506

Number of features: 13

Minimum and maximum housing prices: min 5.0, max 50.0

Mean and median prices of Boston housing prices: mean: 22.53, median: 21.2

Standard deviation: 9.19

2. Evaluating Model Performance

2.1 Performance metrics

I used mean squared error as performance measurement. This error can be divided into Bias and Variance. These values are tradeoff, therefore finding the minimum mean squared error means that finding the best combination of Bias and Variance.

2.2 Data splitting

I split the data into 336 training data and 167 testing data. If we don't split the data, we can't confirm the generalizing capability of the model.

2.3 Cross validation

I used grid search method for cross validation. Grid search is a method to find best parameters for the model in particular interval. Grid search's basic idea is full search. In this case, the search interval is only 10 points(max depth 1~10), therefore I used grid search to optimize parameters. In my program, I used 10-fold cross validation for grid search.

3. Analyzing Model Performance

As figure 1 and 2 shows, as training size increases, the test error decreases

and the training error increases. In addition, when the model complexity is low, the model is suffered from high bias because the model can't learn the trend of the dataset due to the less features. On the other hand, when the model complexity is high, the model is suffered from high variance because over fitting occurs.

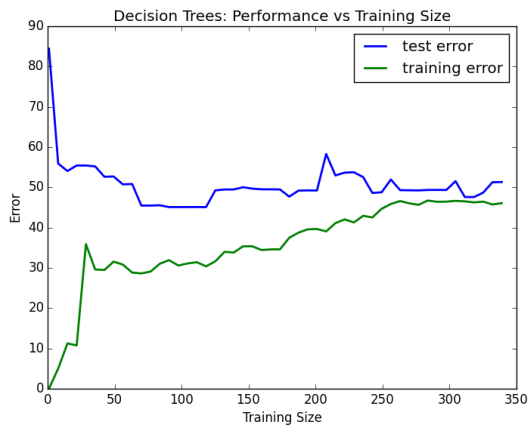


Figure 1 Performance depth:1



Figure 2 Performance depth:10

As figure 3 shows, training error decreases and test error increases after decrease as model complexity increase. It seems that max depth:5 because test error is converged around max depth 5.

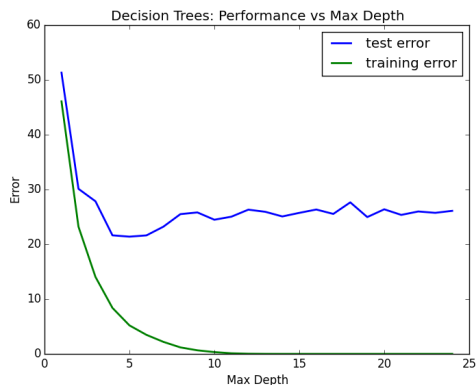


Figure 3 Performance vs Max depth

4. Model Prediction

Parameters: 11.95, 0.00, 18.100, 0, 0.6590, 5.6090, 90.00, 1.385, 24, 680.0, 20.20, 332.09, 12.13

Prediction Price: **20.76598639**

The house price of our client seems to be lower than average price in Boston because the mean of the housing price in Boston is 22.53. However, the 95% confidential interval of this dataset' mean is [21.73, 23.33]. Therefore, we can't say that this 22.77 is lower than Boston Housing average price at 5% significance level.