Predicting Boston Housing Prices Report

1. Statistical Analysis and Data Exploration

- Number of data points (houses)?

  506

- Number of features?

  13

- Minimum and maximum housing prices?

  Minimum house price: 5. Maximum house price: 50

- Mean and median Boston housing prices?

  Mean Boston housing price: 22.5328. Median Boston housing price: 21.2.

- Standard deviation?

  Standard deviation: 9.188

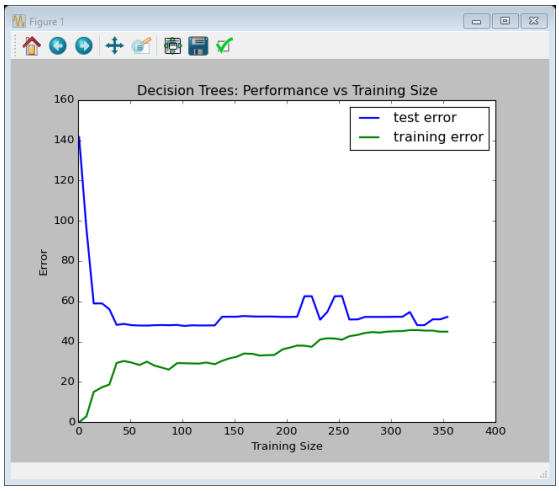
2. Evaluating Model Performance

- Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?

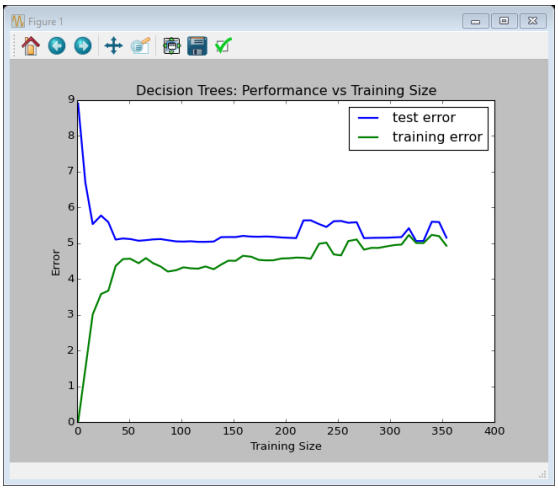
  Regression Metrics (Mean squared error). The Boston housing price predictions are make predictions on continuous data, it is a regression problem, so the regression metrics is a good choice here.

  The classification is about making prediction on unseen examples and deciding which category new instants belongs. The Boston housing predictions is not belong to this category, so the classification metrics is not appropriate here.

  In the regression metrics, I choose the mean squared error, this measurement squared the distance between predict value and the true value. It ensures the value is positive and emphasis the large error. The following two figures show the mean squared error and mean absolute error. From these figure, we can see the mean squared error are larger than mean absolute error. It means the difference are emphasized, but I do not untendered why we need mean squared error.



Mean Squared Error



Mean Absolute Error

- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?

  Split data into training and testing data can help to verify the effectiveness of the trained model. If do not split, the model already see all the data, the training model will completely fit the data, so that we do not know how the performance of the model.

- What does grid search do and why might you want to use it?

  The grid search method is used to optimize the model parameters.

- Why is cross validation useful and why might we use it with grid search?

  When we have limited dataset, the training model may not accuracy, it is easy to understand more dataset will result better model. The cross-validation is an iterative process where train/test sets are randomly generated multiples times in order to evaluate the algorithm at each split, the results are then averaged over the splits. Use cross validation in grid search can generate the best parameters, But when I try to set CV=10 in GridSearchCV, the model is different from not use cross validation in grid search. So how to decide which one is the best model? What is the criteria?

3. Analyzing Model Performance

- Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?

  The training error is increase as training size increases. The testing error is decrease as training size increases.

- Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained, does it suffer from either high bias/under fitting or high variance/overfitting?

Depth 1 suffer from high bias, depth 10 suffer from high variance.

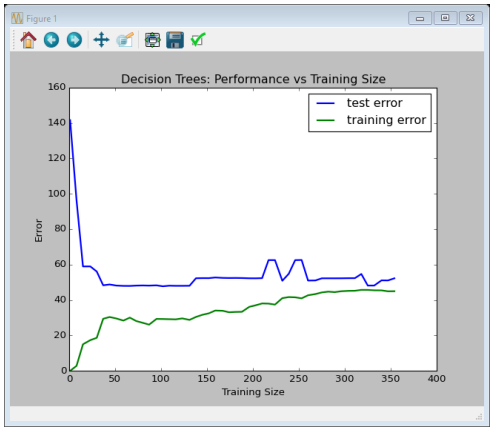


Figure 1.Max Depth 1

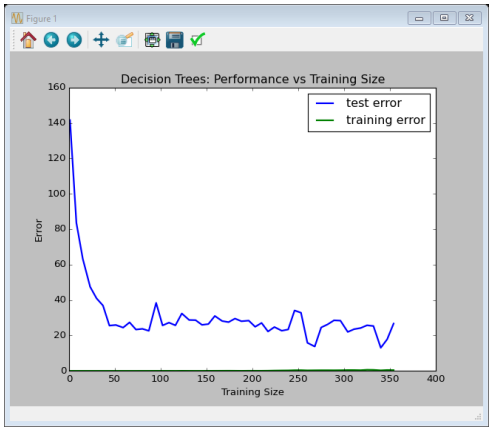


Figure 2. Max Depth 10

Bias occurs when a model has enough data but is not complex enough to capture the underlying relationships. The figure1 is the error and training size figure when max depth is one and the figure 2 is max depth is ten.

From these two figures, we can find out the error of the max depth 1 is almost twice of max depth is 10. That is because when we construct the regression decision tree, we only separate the train set one time. There are many underlying features of the data is not found. That results high bias.

In the contrast, figure two has the max depth of 10. We train the model to a very detailed level, so that, we can find the test error is decrease compare to the figure 1. However, we also notice the test error is not stable in this situation. That is because the data is overly trained, the training error close to zero but when we apply on test error, many test error deviate from the average error very much. That is because when we train the model, to fit as many training data as we can, we applied some unique feature on some data, those feature may only exist on only one data sample and makes the model unable to generalize its predictions to the larger population.

- Look at the model complexity graph. How do the training and test error relate to increasing model complexity? Based on this relationship, which model (max depth) best generalizes the dataset and why?

According to the figure above. When increase the model complexity, the training error decreases. The test error will decrease within a certain max depth. After a threshold, here is the max depth four; the testing error will increase again. Based on the relationship, we should choose max depth 4 to generalize the dataset. Because after max depth 4, the model is over fit the dataset and under fit before the max depth 4.

4. Model Prediction

- Model makes predicted housing price with detailed model parameters (max depth) reported using grid search. Note due to the small randomization of the code it is recommended to run the program several times to identify the most common/reasonable price/model complexity.

Best model parameters: max depth is four. I ran the program several times, depth 4 appeal most times, but I do not understand why there are different results. Predicted house price: 21. 62974359

- Compare prediction to earlier statistics and make a case if you think it is a valid model.

According to the results above, the mean price is 22.5328, standard deviation is 9.188, and the median price is 21.2. My predicated house price is 21.63; it is near the mean and median price and less than one standard deviation. However, I do not understand why less than one standard deviation makes this result reasonable, how to quantify this?