P1: Predicting Boston Housing Prices Report

By Himanshu Dongre

1) Statistical Analysis and Data Exploration

- Number of data points (houses)?

  Ans:506

- Number of features?

  Ans:14(Including MEDV feature)

- Minimum and maximum housing prices?

  Ans: Min Value:5.0

  Max Value:50.0

- Mean and median Boston housing prices?

  Ans: Mean =22.5328063241

  Median =21.2

- Standard deviation?

  Ans: Standard Deviation =9.18801154528

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?

  Ans: Mean Squared Error is best to use for predicting Boston Housing data and analyzing the errors. This method is most appropriate because of the benefits of squaring the residual error which automatically converts all the errors as positives, emphasizes larger errors rather than smaller errors, and from calculus is differentiable which allows us to find the minimum or maximum values. Thus it is better suited than

Mean Absolute Error in this case. Also this is a regression problem i.e. we are dealing with a model that makes predictions on continuous data and care about how close the prediction is to the the true value. We are not concerned about how accurate the prediction is. Thus other measurements may not be appropriate in this case.

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

  Ans: It is important to split Boston housing data into training and testing data as it gives us a way to estimate the performance of the model on an independent real dataset. It also helps us prevent over fitting as the model has not seen the data during training and thus won't already know the answers to the presented dataset. If dataset is not split into training and testing data, then it might happen that when the model encounters real data it won't be able to predict it properly because of over fitting.

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

  Ans: Grid search allows us to work through multiple combination of parameter tunes and cross-validation which can help us determine the best performance. It generated a grid of parameters combinations to try over an algorithm which is then used by fit function which tries to fit all the parameter combinations to estimate best fit.

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

  Ans: One of the main goals while creating a model is that we need to maximize both training and testing data to get more effective model. More the training and testing data better is the model. However, one of the limitations of splitting the available dataset into training and testing data is that increasing one decreases another. To overcome this, we use cross-validation. Cross validation helps us to maximize both training and testing dataset using the available dataset. It is used in the grid search to assess the performance of our algorithm over different parameter combinations to give us the best estimator for the available data.

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?

  Ans: From the learning curve graphs we can see that the general trend is that as the training size increases, testing error decreases as our model is trying to learn overtime.

Also as the training size increases initially the training error increases as initially the model is trying to fit the training data. However, after sometime both the training and testing error plateau and the difference between two graphs does not change. That is when we can say that the model has learned as much as it can from the training and testing data.

- 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/over fitting?

  Ans: Learning curve graph with max depth 1 seems to be suffering from high bias/underfitting.Because when the training and testing error almost converge the error is still pretty high (in 40s). Also learning graph with max depth 10 seems to suffer from high variance/over fitting because there is a big gap between training and testing error after the model is fully trained.

- 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?

  Ans: Looking at the model complexity graph one can say that the training error keeps reducing till it almost merges with x line. This makes sense as with increasing complexity model learns to fit training data better. As for testing error, the error first seems to decrease with increasing complexity but after the max depth of 4 the testing error does not decrease further with increase in complexity i.e model starts to overfit.Based on this it appears that a max depth of 4 provides an ideal model which best generalizes the dataset.

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.

  Ans: After a few runs it appears that the most common best max depth is 4.

  Prediction: [21.62974359]

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

Ans: Looking at the learning curve graphs for max depth 1 to 10 and the error vs max depth graphs, it appears that model with max depth 4 provides best generalization for the given data. Below max depth of 4 the model seems to under fit and above max depth of 4 models seems to start over fitting. Also looking at the error vs max depth graph we can see that the error decreases for testing data till a max depth of 4 and after that tends to increase a little. The predicted housing price of 21.62974359 also seems close to the median(21.2).All these evidences prove that max depth of 4 is the best estimator for predicting Boston housing prices.