Predicting Boston Housing Prices

Jonathan Lee

Questions and Report Structure

- Statistical Analysis and Data Exploration
- Evaluating Model Performance
- Analyzing Model Performance
- Model Prediction

Statistical Analysis and Data Exploration

Query	**Result**
Number of Houses	506
Number of features	13
Minimum Housing Price	5
Maximum Housing Price	50
Mean Housing Price	22.5328063241
Median Housing Price	21.2
Standard deviation	9.18801154528

Evaluating Model Performance (Part I)

- Mean Squared Error is best used to predict Boston Housing data
  - Goal is to minimize this error.
  - This method is most appropriate since we are trying to predict a housing price. The output is continuous and not a binary. Therefore, we do not use precision and recall since we are not trying to classify anything.
  - It also helps emphasize the larger errors compared the absolute error model.
- It is important to split data into training and testing to prevent overfitting
  - For example, if we just trained all the data, then we do not have any data to validate our model to check if it is underfit or overfit. Therefore, by having test data, we can see which model minimizes the error on the test data.

Evaluating Model Performance (Part II)

- Grid search sweeps the grid parameters (max_depth) and trains the model with each of them and finds the one with the best value
  - This needs to be used in order find the best model parameter.
- Cross validation is useful because, by using the default 3-fold cross-validation, we can utilize the full dataset
  - For example, we break the data set into three equal sections. And the one section is used as the validation data. We train with the remaining data. Repeat this processthree times using a different bucket each time. Finally, we average the results.
  - This is used with grid search because we will sweep the model parameter while doing the 3-fold cross-validation for each parameter.

Analyzing Model Performance (Part I)

- General trend of training data error increases as training size increases
- General trend of test data error decreases as training size increases
- For the model with max_depth=1, the model suffers from high bias since the testing error curve flattens out at an high value and adding more data does not help. The model suffers from underfitting because error of the training data is large and increasing, alarmingly, with more data points.
- For the model with max_depth=10, the model suffers from high variance. Although the testing error curve is slightly decreasing with more training data, the gap between testing and training error is huge. This is because the error of the training data is extremely small compared to the testing error due to overfitting.

Analyzing Model Performance (Part II)

- According to the model complexity graph, the training error decreases as max depth increases.
    - For example, you can see that he test error decreases until around the value five where it has a minimum and then slowly increases.
- The max depth is around 5. This best generalizes the dataset since the test error increases around this point due to higher variance.

Model Prediction

- The predicted value is around 20.766 with model complexity around 5
- This could make sense because it's between the max and min. Clearly if it was above or below that then we know there's a problem.
- The value is also within one standard deviation of the mean
- Also checked some some features of the houses in the data set with similar price to see if it was similar and it looked similar enough