MLND - P1 - Vishakh Rayapeta

[1) Statistical Analysis and Data Exploration]
[2) Evaluating Model Performance]
[3) Analyzing Model Performance]
[4) Model Prediction]

1) Statistical Analysis and Data Exploration

- Number of data points (houses)?

  506

- Number of features?

  13

- Minimum and maximum housing prices?

  5.0, 50.0

- Mean and median Boston housing prices?

  22.533, 21.2

- 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?

The Decision Tree Regressor uses the 'Mean Squared Error' to measure the quality of the split. Some of the reasons to use this measure are as follows:

- *All errors are positive and do not cancel each other out*
- *Larger errors are emphasized*
- *This error method is 'differentiable' - so a minimum/maximum point can be derived. Absolute value error is not 'differentiable'.*

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

In the absence of the testing data, it will not be possible to measure and detect overfitting. Overfitting is caused by generating a prediction function that has a seemingly perfect score on the samples but fails to predict on unseen data.

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

Grid search is used to learn estimator parameters. For example: The max_depth parameter in the 'Decision Tree Regressor' can be optimized using an array of parameter values.

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

When the number of samples are scarce, cross validation provides a method to efficiently use the samples needed for learning the model. Another advantage of cross validation is that it iterates the partitions for training & testing sets over the entire sample space. This eliminates the bias that may be introduced by an arbitrary division of the sample space into a single training & testing set.

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?

  As training size increases training error rises and testing error falls.

- 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/underfitting or high variance/overfitting?

  For max_depth = 1 *(test error ~ 56, training error ~ 42)*

The test error is approximately 56, the training error is approximately 42. Both these errors are comparatively higher than the error values when max_depth is 10. This leads me to conclude that this model suffers from bias / underfitting.

For max_depth = 10 *(test error ~ 18, training error ~ 0)*

The training error is almost zero which implies a perfect fit. However, the test error is approximately 18. Also, the relative difference between the training & test error has increased from from 14(56-42) to 18(18-0). This leads me to conclude that the model suffers from high variance/overfitting.

- 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?

The training error reduces exponentially with increasing model complexity. The test error also reduces exponentially as the complexity increases from 0. However, it reaches a minimum value at max_depth = 6 and then fluctuates.

max_depth = 6 seems to best generalize the dataset since the test error is the smallest.

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.

  Most common max_depth = 4

  Prediction = 21.630

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

  The value 21.630 is reasonable - It lies within the min/max range of [5.0 50.0].

  Also, it is within the standard deviation range (9.188) of the mean (22.533).