Ted Samore

Predicting Boston Housing Prices

This report involves analyzing and predicting housing prices in Boston based on a dataset built into scikit learn. The dataset contains the prices of 506 homes, and 13 features describing each home. The following describes the dataset:

- The minimum price is: 5.0
- The max price is 50.0
- The mean price is 22.5328063241
- The median price is 21.2
- The standard deviation is 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?

  Since the data output is continuous, a regression metric is needed. I chose to use mean squared error. A classification metric would not be appropriate.

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

  The problem if you don't split your data into training and testing set is that your model won't generalize as well. The model would mistake signal for noise, and overfit to the training data.

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

Grid search will help you find the ideal value of parameters to optimize the model. It compares the result of your model with different parameters.

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

Cross validation is the process by which you have an independent test dataset to test the validity of the model. Grid search is used to find the optimal parameter values for the model. You would use cross validation with grid search to make sure that the model parameters are ideal for both the training and test datasets.

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 increase, training error increases, and testing error decreases, until a certain point when they both level off.

- 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?

I think the first graph (with max depth 1) suffers from high bias, because it has a relatively high error rate for both the training and set datasets. The last graph (with max depth 10) suffers from high variance because there is a huge discrepancy between the error of the test set and training set with the test set having a higher error.

- 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 model with max depth 5 best generalizes the data, since there is limited or no benefit for the models with greater depth than 5. It is the simplest model with the lowest error on the test set.

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.

  The result I get is max_depth=None.

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

If my code is correct, I don't think this is a valid model, since my earlier statistics suggested that the ideal max depth is around 5.