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, because it is more sensitive to outliers than absolute error. The mean squared error function is also differentiable, which allows us to find the minimum error values with calculus.

- 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 you can't determine if your model is overfit. In order to assess whether our model is overfit or not, we need to have an independent test set.

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

Gridsearch tests different parameter values of your model to help you find the ideal value of parameters to optimize the model.

- 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. As stated in the earlier answer, this helps you assess if your model is overfit or not. If you only have one training and one testing data set, your model might overfit if the testing dataset is imblanaced, and your model won't optimize all of the data (since some of it is held for the testing set). K-fold cross validation is a process in which you divide the data into multiple samples to maximize data usage and reduce the chance of overfiting.

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 4 best generalizes the data, since there is limited or no benefit for the models with greater depth than 4. It is the simplest model with the lowest error on the test set. As model complexity increases past a max depth of 4, the data is not generalized better.

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

  After running the program several times, it was confirmed that the model with max_depth 4 is the model with the best parameters.

The prediction for the house for the vector [11.95, 0.0, 18.1, 0, 0.659, 5.609, 90.0, 1.385, 24, 680.0, 20.2, 332.09, 12.13] is 20.76598639. This appears to be a reasonable prediction. It is 0.19 standard deviations away from the dataset's mean (22.532806324). Additionally, the 10 nearest neighbors to this data point have a mean of 21.52, which is close to the prediction.