Predicting Boston Housing Prices Report

1. Statistical Analysis and Data Exploration

- Number of data points (houses)?

  506

- Number of features?

  506

- Minimum and maximum housing prices?

  Minimum house price: 5. Maximum house price: 50

- Mean and median Boston housing prices?

  Mean Boston housing price: 22.5328. Median Boston housing price: 21.2.

- Standard deviation?

  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?

  Regression Metrics (Mean squared error). The Boston housing price predictions are make predictions on continuous data, it is a regression problem, so the regression metrics is a good choice here.

  The classification is about making prediction on unseen examples and deciding which category new instants belongs. The Boston housing predictions is not belong to this category, so the classification metrics is not appropriate here.

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

  Split data into training and testing data can help to verify the effectiveness of the trained model. If do not split, the model already see all the data, the training model will completely fit the data, so that we do not know how the performance of the model.

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

  The grid search method is used to optimize the model parameters.

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

  Cross-validation is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available. A grid search algorithm must be guided by some performance metric

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?

  The training error is increase as training size increases. The testing error is decrease as training size increases.

- 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?

  Depth 1 suffer from high bias, depth 10 suffer from high variance.

- 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?

  When increase the model complexity, the training error decreases while the test error increases. Depth1 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.

  Predicted house price: 22.45376984

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