

Statistical Analysis and Data Exploration

Size of data (number of houses) = 506

Number of features = 13

Minimum price = 5.0

Maximum price = 50.0

Mean price = 22.5328063241

Median price = 21.2

Standard deviation price = 9.18801154528

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 Boston housing is a regression problem; hence we have to use regression metrics, e. g. explained variation, mean absolute error, mean squared error, median absolute error, the coefficient of determination etc.

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

It allows us to estimate the performance on an independent data set and to check on overfitting. Otherwise, we run into issues evaluating a model because it has already seen all the data.

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

The grid search is a way of systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. We use it since we want to find the best model.

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

Cross validation allows us to avoid the problem of overfitting.

Analyzing Model Performance

- Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?

Usually training and testing error converges 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?

It suffers from high variance/overfitting since the model is getting more complicated.

- 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 is decreasing; the test error is increasing. The model with max depth is no greater than

1 gives the best generalization; after the overfitting takes place – the error is increasing.

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.

Yes, it looks like a quite valid model.