Assignment Report

Project 1: Predicting Boston Housing Prices

Udacity Machine Learning Nanodegree Program
Authored by: Fedor Sulaev
Date: 12/04/2015

Statistical Analysis and Data Exploration

Provided dataset consists of $n = 506$ entries with $k = 13$ features. Housing prices range from 5.0 to 50.0 with mean $\mu = 22.5328$, median $\tilde{x} = 21.2$ and standard deviation $\sigma = 9.188$.

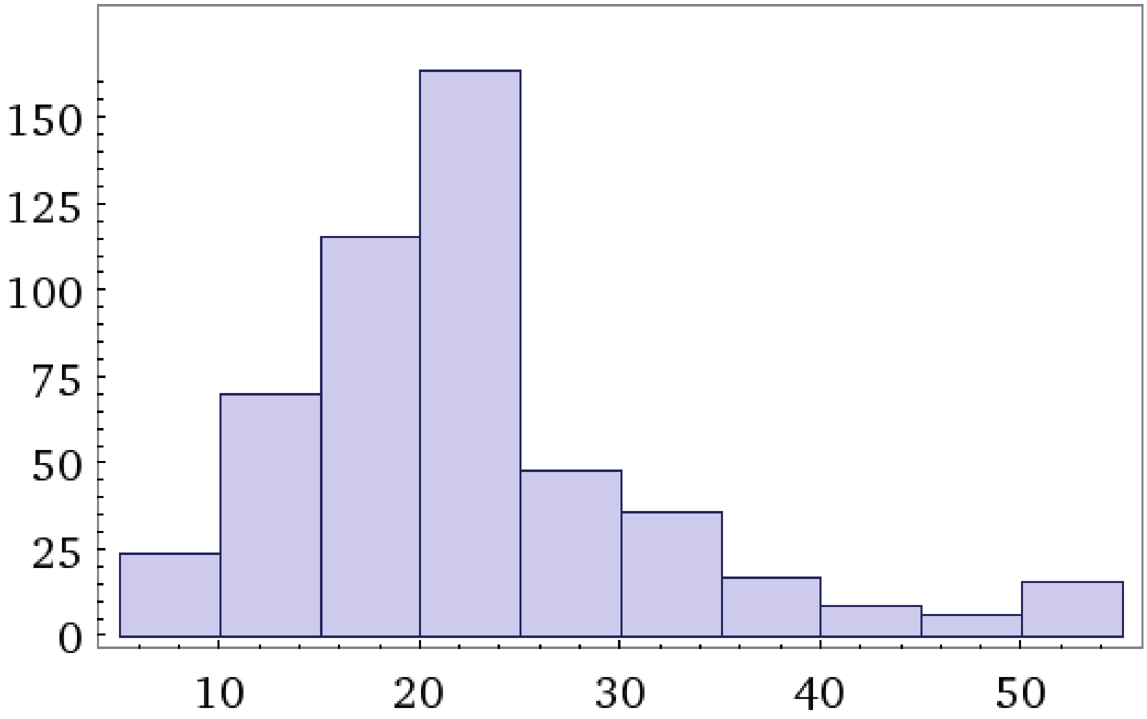

Figure 1 Histogram of housing prices

Evaluating Model Performance

Mean squared error is used as a model performance metric for predicting Boston housing data. MSE is a common metric for regression problems, it converts all the errors to positives and emphasizes larger errors. Another alternative to MSE is the mean absolute error which is more robust to outliers however, as there are no outliers in the housing dataset, MSE is more appropriate for the task.

Data is split into training and testing sets, this technic is used to prevent overfitting and give an estimate of performance on an independent dataset. When a prediction function uses one dataset for learning and testing it might lead to overfitting, the function will have a perfect performance score on the initial dataset but will fail to make accurate predictions on new datasets.

K-fold cross-validation is used for model validation, this method provides maximum accuracy and since the dataset is relatively small, computing performance advantages of conventional validation method are not useful for this particular problem.

Grid search is used for model optimization; it systematically works through different combinations of parameters in order to maximize model performance. Hyperparameter optimization technics like grid search ensure that the model does not overfit its data.

Analyzing Model Performance

Learning curve graphs show that training and testing errors converge with increase of the training size. As training size increases test error decreases and flattens out and training error slightly increases.

On the first learning curve graph the training and testing errors converge and are relatively high which means that the model suffers from high bias. On the last graph the training and testing errors have a gap in between, that means the model has high variance and suffers from overfitting.

Model complexity graph shows that increase in model complexity in general leads to decrease in test and training error but after a certain point training error flattens out and test error increases slightly or flattens out. Max depth bigger than 10 does not seem to affect the training error and only increases variance as the test error grows with increase of max depth.

Model Prediction

Model with max depth of 10 predicts a price of 19.9 for a house with following features:

Price	Features												
19.9	11.95	0.0	18.1	0.0	0.659	5.609	90.0	1.385	24.0	680.0	20.2	332.09	12.13

The dataset contains several houses with similar price:

Price	Features												
19.9	0.627	0.0	8.14	0.0	0.538	5.834	56.5	4.499	4.0	307.0	21.0	395.62	8.47
19.9	3.837	0.0	18.1	0.0	0.77	6.251	91.1	2.296	24.0	666.0	20.2	350.65	14.19
19.9	3.164	0.0	18.1	0.0	0.655	5.759	48.2	3.067	24.0	666.0	20.2	334.4	14.13
19.9	4.349	0.0	18.1	0.0	0.58	6.167	84.0	3.033	24.0	666.0	20.2	396.9	16.29

The same performance metric that is used for learning can be used for evaluation of the prediction. Feature sets of real houses with the same price have following error values when compared to the target feature set:

Entry #	Mean squared error
1	11149.116
2	47.154
3	156.356
4	346.958

Error values appear to be relatively small for most of the entries and predicted price is reasonable for the target set of features.