Assignment Report

Project 1: Predicting Boston Housing Prices

Udacity Machine Learning Nanodegree Program
Authored by: Fedor Sulaev
Date: 12/10/2015

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 $\sigma = 9.188$.

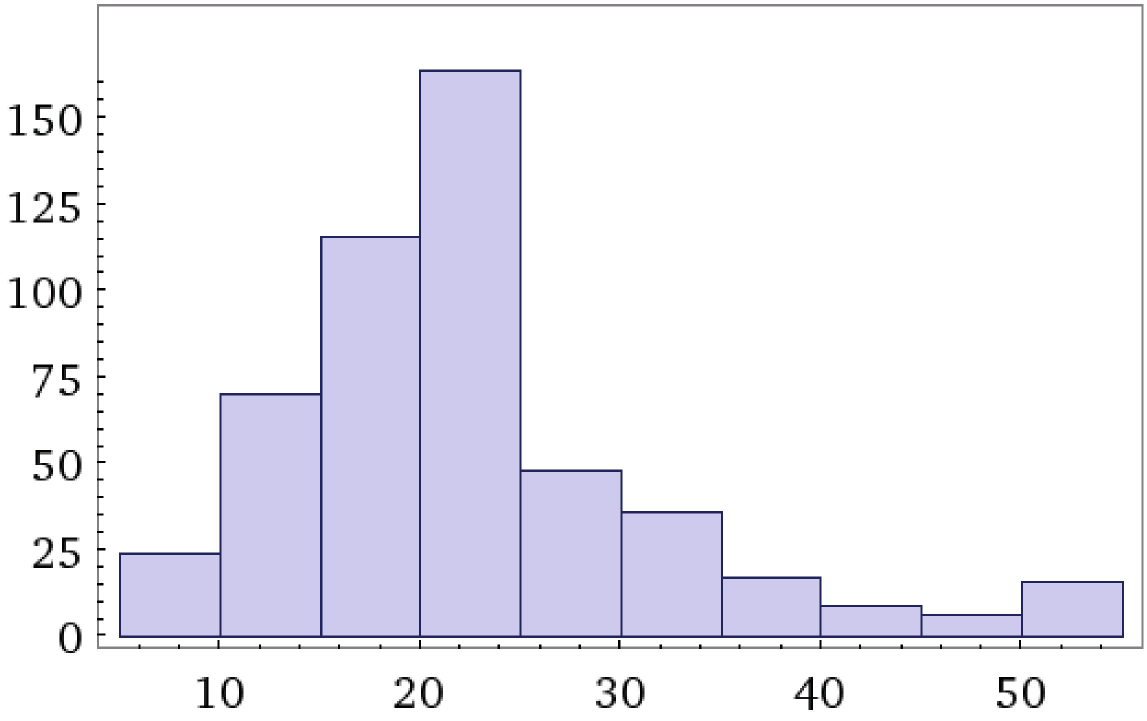

Figure 1 Histogram of housing prices

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 evaluation, this method involves partitioning a dataset into subsets, performing analysis on one subset, and validating the analysis on the other subset. This process is repeated multiple times and results are averaged. Cross validation provides an accurate estimate of model performance for a price of high computation time, and since the dataset for housing prices is relatively small, cross validation is an optimal choice for this 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. Judging by the graph decision tree depth of 4 provides the optimal model: errors are relatively low and the distance between test and training errors is average, which means that this depth provides a compromise between error value and high variance.

Model Prediction

Model with max depth of 4 predicts a price of 21.6 for a house with following features:

Price	Features												
21.6	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 value 21.6 is within one standard deviation from the mean.