P1: Predicting Boston Housing Prices

Statistical Analysis and Data Exploration

Number of data points?	506
Number of features?	13
Minimum house price?	5.0
Maximum house price?	50.0
Mean Boston house price?	22.5328063241
Median Boston house price?	21.2
Standard deviation?	9.18801154528

Note: Prices are in US$1000's.

Evaluating Model Performance

Which measure of model performance is best to use for predicting Boston housing data? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?

As we are predicting on continuous data this is a regression problem. The performance metrics for regression problems are mean absolute error, mean squared error, median absolute error and r squared.

The mean squared error performance measure captures the average difference bwtween the expected atrget values in the test set and the values predicted by the model. Mean squared error values fall into the range [0, infinity], with smaller values indicating better model performance. Using a squared term does result in the mean squared error over-estimating error when the data contains outliers. For that reason, I choose to use the mean absolute error (MAE). The mean absolute error has no such squared term so is not as susceptible to outliers. Another advantage it has is that it is in the same units as the target feature, so can be used to give an intuitive feel of how well a model performs. (Note: you can take the square root of the mean squared error to get back into the units of the target feature but that takes an extra step and computation).

Other measures of performance metric like accuracy, precision, recall and F-score are not appropriate as they are used to evaluate performance of classification models.

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

You need to split your input data into separate training and testing sets inorder to preserve an independent set of data to verify that the model will generalise well when presented with new data. If you train the model with all your data you will end up overfitting and when presented with as yet unseen data would make a prediction that would not be reliable and that you could have no confidence in.

Which cross validation technique do you think is most appropriate and why?

As the amount of input data is relatively small the best way to maximize the training and testing sets is to use k-fold cross validation.

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

Grid search systematically tries multiple combinations of specified estimator (regressor or classifier) parameters inorder to find the combination of parameters and their values that gives the best score for the estimator.

You might want to use grid search to automate the discovery of optimal parameter value combinations for a model particularly if the model you were building had lots of parameters and parameter values and you were unsure of the best values to pick.

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 general trend as you increase training size is that the curves for training and testing error will converge toward a level representing the amount of irreducible error in the data.

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?

Looking at the learning curve for max depth of 1, we observe high error rates on both the training and test data. This indicates high bias, bias occurs when you have an inadequate model, which leads to low accuracy in predictions.

Looking at the learning curve for max depth of 10, we observe that the error rates of both the training and test data have decreased when comparing to the curve for max depth of 1. The error rate on the training data has decreased to a minimum. However, there is a much wider gap between the training and test learning curves, with the test error being greater than the training error, this indicates high variance. Error due to variance indicates overfitting, this means the model is too complex and/or there is not enough data to support it.

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?

With increasing model complexity the general trend is that more variability exists in the model for a fixed set of data.

From my graph the model that best generalizes the dataset has max depth=5. Its at this point where the model fit for the test data is at its best and the difference between training and test errors at one of the smallest values (there are smaller differences but they occur when test errors are larger).

Model Prediction

Compare prediction to earlier statistics

My prediction for the supplied data point is 20.76598639 (amount in US$1000's).
The aim of this project was to create a model to be used to find the best price a client can expect to sell their house for, the predicted price compares favourably with the mean and median prices calculated earlier, falling easily within 1 standard deviation of the mean. The predicted price is very resaonable, in that its not too high, which might put-off prospective purchasers, and not too low, which would not benefit the seller.