PREDICTING BOSTON HOUSING PRICES

DIEGO ROQUE MONTOYA

Abstract. We use Decision Trees with model selection to predict the prices of houses in the Boston market using data from ICS[1].

1. Statistical Analysis and Data Exploration

First we calculate some properties of the dataset.

1.1. **What is the number of data points?** There are 506 data points.

1.2. **What is the number of features?** There are 13 features.

1.3. **What are the minimum and maximum housing prices?** The minimum and maximum housing prices are 5 and 50 thousand respectively

1.4. **What are the mean and median of housing prices?** The mean of the housing prices is 22.532 thousand and the median price is 21.2 thousand.

1.5. **What is the standard deviation of the housing prices?** The standard deviation of the housing prices is 9.188 thousand.

Table 1. Summary

Measurement	Boston Dataset
Data Points	506
Features	13
Minimum Price (thousands)	5
Maximum Price (thousands)	50
Mean Price (thousands)	22.532806
Median Price (thousands)	21.2
Standard Deviation (thousands)	9.188012

Date: January 29, 2016.

2. Evaluating Model Performance

We use a mean square error, as we care about predicting things closely to truth. Furthermore, this is the metric that makes most sense, as internally it's what the Decision Tree Regressor uses for splits. If it's optimizing over a metric, it should be measured over the same metric.

We split the data in training and test data to obtain a reasonable estimate of how well our algorithm would fare in the real world, with new data, assuming we trained over a representative sample of the real world data. Otherwise we would not be able to get a reportable score, that's comparable with other results.

Decision Trees can be prone to overfitting. To avoid that, we will tune the parameter of the maximum depth allowed to prevent overfitting while minimizing error, to get a model with better generality.

Grid Search evaluates every combination of the parameter options and selects the one that generalizes best. We can't train using the test set, and to train with the whole training set would be prone to overfitting. So we scored with a 3-fold cross validation from the training set, which splits the data in three parts, trains with two of those parts and tests with the remaining one, then averages the three scores to get the overall score. This will be the strategy that we will use to find the best parameter.

3. Analyzing Model Performance

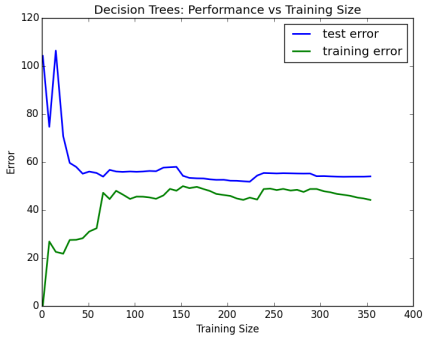
We can see from Figure 1, that across the different models, as the training size increases the testing error decreases until almost plateau. Similarly, the training error increases until a plateau, never surpassing the testing error.

Comparing the cases where the maximum depth is 1 and when the maximum depth is 10, we see that the former has a significant bigger error on both the training and the test sets than the latter. On the other hand, there's a closer relation between the training and test errors when the maximum depth is 1 than when it's 10, which has a training error of almost zero. This indicates that training the decision tree with a maximum depth of 1 underfits the data and it's left with high bias, while training with a maximum depth of 10 overfits the data and it's left with high variance.

We can see from Figure 2 that while increasing the maximum depth decreases to almost zero the training error, the testing error decreases until a maximum depth of around 6, then increases slightly. So lower maximum depths have high bias while high maxium depths have high variance. This suggest we can select a model that best generalizes with a maximum depth around 5.

4. Model Prediction

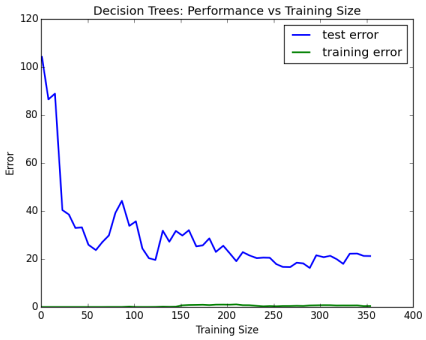
Running Grid Search with 3-fold cross validation gives us a model with maximum depth of 4. This is consistent with the model performance graph in Figure 2. This model predicts that a house representing the datapoint [11.95, 0.00, 18.100, 0, 0.6590, 5.6090, 90.00, 1.385, 24, 680.0, 20.20, 332.09, 12.13] would have a predicted price of 21.6297 thousands. This is close both to the mean price of 21.532 thousands and the median price of 21.2 thousands, so it makes sense given that the standard deviation is 9.118 thousands.



(A) Max Depth = 1

(B) Max Depth = 3

(C) Max Depth = 7



(D) Max Depth = 10

FIGURE 1. Performance vs Training Size

REFERENCES

1. M. Lichman, *UCI machine learning repository*, 2013.

E-mail address: droque@mit.edu

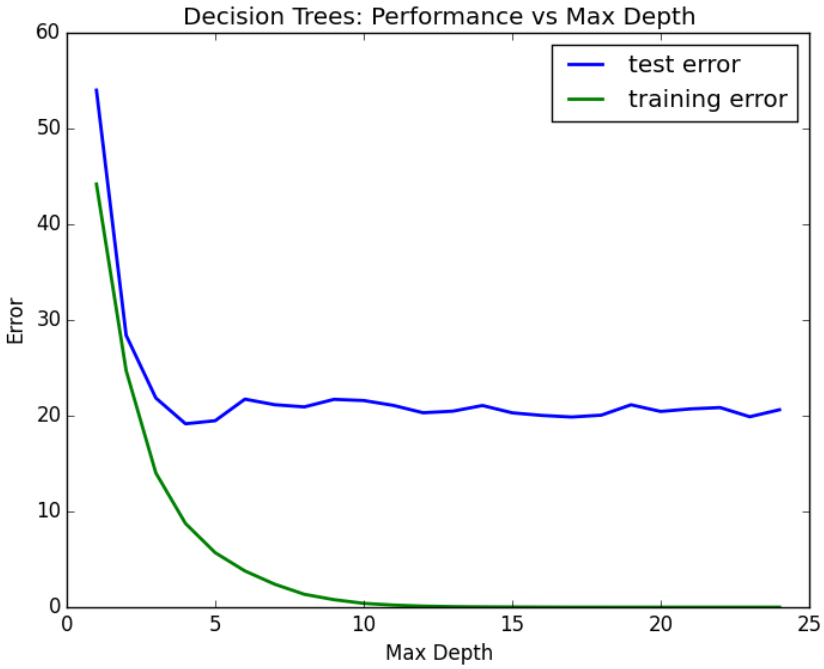


FIGURE 2. Performance vs Max Depth