P1: Predicting Boston Housing Prices

Dalong Li

Feb 2016

Statistical Analysis and Data Exploration

1)Of the available features for a given home, choose three you feel are significant and give a brief description for each of what they measure.
Answer: 1.CRIM: people wouldn't like to living in a place with high crime rate; 2.RM: the number of rooms definitely affect house price, the more rooms the higher price; 3.RAD: accessible highway can bring plenty convenience for travelling.

2)Using your client's feature set CLIENT_FEATURES, which values correspond with the features you've chosen above?
Answer: CRIM:11.95, RM:5.609, RAD:24

3)Why do we split the data into training and testing subsets for our model?
Answer: Training dataset is used for training predict model to get the optimal parameters, testing dataset mainly used for evaluating the model's performance. The benefit is it'd provide two independent dataset to give estimate of performance and serve as check on over-fitting.

4)Which performance metric below did you find was most appropriate for predicting housing prices and analyzing the total error. Why?

- Accuracy
- Precision
- Recall
- F1 score
- Mean Squared Error(MSE)
- Mean Absolute Error(MAE)

Answer: Mean Squared Error(MSE). First, this is a regression problem, so we can't use classification metrics, such as accuracy, precision, recall and F1 Score. MAE and MSE are commonly used metrics for regression, some benefit of MSE is that it automatically converts all the errors as positives, emphasizes larger errors rather than smaller errors, and from calculus is differentiable which allow us to find the minimum or maximum values.

5)What is the grid search algorithm and when is it applicable?

Answer: Grid search is algorithm for searching estimator parameters, it can exhaustively considers all parameter combinations. We pass specified algorithm and dictionary of parameters to GridSearchCV function then it will return a estimator with its parameters optimized by cross-validated grid-search over a parameter grid. It's applicable when the data we will training and testing has big number of features.

6)What is cross-validation, and how is it performed on a model? Why would cross-validation be helpful when using grid search?

Answer: Cross-validation is a statistical method of evaluating by dividing data into two segments: one used to train a model and the other used to validate the model. The basic form of cross-validation is k-fold cross-validation. In k-fold cross-validation the data is first partitioned into k equally sized folds. Subsequently k iterations of training and validation are performed, within each iteration a different fold of the data is held-out for validation while the remaining k-1 folds are used for training. Cross validation is useful because it maximize both the training and testing data so that the data we can use to provide besting learning result and best validation. Cross-Validation can avoid accidentally over fit due to random imbalanced splitting when using grid search.

Analyzing Model Performance

7)Choose one of the learning curve graphs that are created above. What is the max depth for the chosen model? As the size of the training set increases, what happens to the training error? What happens to the testing error?

Answer: The model with max_depth=3, the training error increased and testing error decreased as training set size increases. When the training set is small, the trained model can essentially "memorize" all of the training data. As the training set gets larger, the model won't be able to fit all of the training data exactly. The opposite is happening with the test set. When the training set is small, then it's more likely the model hasn't seen similar data before. As the training set gets larger, it becomes more likely that the model has seen similar data before.

8)Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high variance when the max depth is 1? What about when the max depth is 10?

Answer: When the model is using the full training set, it will suffer from high bias when max depth is 1, bias due to a model being unable to represent the complexity of the underlying data so it always has large training error; it will suffer from high variance when the max depth is 10, variance due to a model that is overly sensitive to the limited data it has been trained on so usually it has very small training error but larger testing error.

9)From the model complexity graph above, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?
Answer: As the model's complexity increases, the training error reduces since the model is becoming more and more fit based on the given training data. However, the testing error raised after reaching to the global minimum, since the model's complexity brings in the high variance/over fitting problem to make the testing error becomes higher. The best balance point is the global minimum of testing error curve, so I guess max_depth=4 will results best.

Model Prediction

10)Using grid search, what is the optimal max_depth parameter for your model? How does this result compare to your initial intuition?
Answer: Cool! The running result approximately fits my initial intuition.

11)With your parameter-tuned model, what is the best selling price for your client's home? How does this selling price compare to the basic statistics you calculated on the dataset?
Answer: The model predicted best selling price is 20.968. The predicted price is between one standard deviation from the mean.

12)In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients' homes in the Greater Boston area.
Answer: Yes, of course. We have build an optimal model based on a statistical analysis with tools available. We've trained a decision tree regressor model using GridSearchCV and found the optimal parameter gives the best performance.