Udacity MLND Project 1 Submission

John Williams

Contents

Question 1

Of the features available for each data point, choose three that you feel are significant and give a brief description for each of what they measure.

Answer:

{'DIS': 'weighted distances to five Boston employment centres', 'RAD': 'index of accessibility to radial highways', 'TAX': 'full-value property-tax rate per $10,000'}

Question 2

Using your client's feature set CLIENT_FEATURES, which values correspond with the features you've chosen above?

Answer:

{'DIS': 1.385, 'RAD': 24, 'TAX': 680.0}

Question 3

Why do we split the data into training and testing sets.

Answer:

We split the data into training and testing subsets to test the model on data it was not trained on. Model performance evaluated on training data yields an optimistic model performance result. Model evaluation on testing data is a more accurate indication of model performance. Running K-fold cross-validation and averaging K performance results is preferred over a single performance test.

Question 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:

Either MSE or MAE are valid performance metrics for regression on continuous response variable(s); I prefer MSE as it enhances larger errors relative to MAE.

Question 5

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

Answer:

A grid search algorithm iterates through a "grid" of tuning paramters applied to a learning model, data set & scoring criteria. The algorithm returns a model fitted on the data with "tuned" parameters optimizing the scoring criteria. The algorithm is applicable when a learning model has tunable parameters.

Question 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 (CV) uses a full data set to obtain the most accurate score of model performance. CV separates the data into K unique training/test sets, averaging the score of K model runs. If CV is not used in grid search, the score of each model run, and hence comparison of tuning parameters, is less than optimal.

Question 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:

I choose the learning curve graph of the model with max depth of 6. As the size of the training set increases, the training error increases, and flattens out after around 300 training points. The testing error decreases, on average, as the training set size increases.

Question 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:

A smaller difference between training error and testing error, as seen in the max depth of 1 learning curve graph, indicates higher model bias. Conversely, a larger difference in error rate indicates higher model variance, as in the max depth 10 curve.

Question 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:

At a depth of 1, training error exceeds testing error, and they are roughly similar at a depth of 2. Above a depth of 2, the testing error is roughly constant (with constant standard deviation), while the training error illustrates a smooth asymptotic decline to zero.

Question 10

Using grid search on the entire dataset, what is the optimal `max_depth` parameter for your model? How does this result compare to your intial intuition?

Answer:

A bias-variance tradoff exists in modeling, where lower complexity (lower max depth) increases bias and decreases variance, while higher complexity (higher max depth) decreases bias and increases variance. The optimal max depth is where the model performance criteria (mean squared error) in minimized; the optimal max depth of 5 occurs near the midpoint of the bias-variance tradoff. Both high bias (low variance) given by low max depth, and high variance (low bias) given by high max depth, decrease model performance from the optimal level.

Question 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 parameter-tuned model prediction of selling price for the client's home, $20,968, is very close to the median selling price of $21,200, and the mean selling price of $22,533 of the data set.

Question 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:

The obvious problem is the data set is old and out of date. If a current data set was available, I would analyze the data for relevance to client comparables and the local real estate environment. Domain analysis is crucial to any data analysis.