Machine Learning Engineer Nanodegree

Project 1: Predicting Boston Housing Prices

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:

1. RM: Average number of rooms per dwelling.
2. DIS: weighted distance to five Boston employment centres. The houses nearer to the employment centres are likely to cost more.
3. LSTAT: Lower status of population (percentage)

Question 2: Using your client's feature set CLIENT_FEATURES, which values correspond with the features you've chosen above?

Answer:

1. RM: 5.609
2. DIS: 1.385
3. LSTAT: 12.13

Question 3: Why do we split the data into training and testing subsets for our model?

Answer: Evaluating the model after splitting the data into training and testing subsets increases the accuracy of the model. If we do not split the data, all the data points are used to construct a model. Such a model may give poor prediction results for new data points as we have not evaluated the model on a different dataset. After splitting the data randomly into two subsets, we use one subset to build the model (train) and the other to evaluate it (test). This helps us in assessing the performance of model on new data points and in reducing overfitting.

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

Answer: Mean Squared Error (MSE):

Accuracy, Precision, Recall and F1 score are only suitable for classification problems. As this is a regression problem in which we have to predict a value, only MSE and MAE are suitable. MSE is more appropriate as large errors have more influence on it than smaller errors.

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

Answer: Grid search algorithm is used for tuning parameters of a machine learning algorithm. A grid search algorithm searches for best parameters by fitting the model on all the combinations of given sets of parameters. The parameters are assessed on the basis of a performance metric. In the above example, the optimum value of maximum depth of a decision tree regressor is estimated by finding the depth of the tree that gives minimum mean squared error. A grid search algorithm automates the task of manually searching the best parameters of a machine learning algorithm.

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 is a technique that assesses the generalizability of a model. Cross validation techniques give an idea about whether the model can be applied to unknown data or not. Cross validation is used to avoid overfitting the data. Splitting data into training and testing datasets and their evaluation is also known as holdout cross validation.

K-fold cross validation is the most used cross validation technique. Data is randomly divided into k subsets. One of the dataset forms a testing dataset and the rest k-1 datasets form training dataset. Training dataset is used to fit the model and testing dataset is used to evaluate it. This process is repeated k times till each data point gets to be in testing dataset exactly once. Overall error is estimated by taking the mean of errors obtained by fitting the model k times.

Grid search algorithm in scikit-learn uses 3-fold cross validation by default. Using cross validation, we can validate the model generated by different parameters to find the optimum parameters.

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: Model: Max depth of decision tree = 6

Training error is slightly increasing with increase in the number of data points.

Testing error is decreasing with increase in the number of data points.

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: Model with max depth = 1: Model suffers from underfitting, which also means that it has high bias.

Model with max depth = 10: Model suffers from overfitting and it has high variance.

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: As the maximum depth of the decision tree is increasing, the training error is decreasing. This is due to the fact that the model is overfitting the training data. The testing error first decreases and then changes very little after the max depth of 4. From the graph, it looks like that the minimum testing error is obtained at max depth of 4, so this model could generalize the dataset well.

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 initial intuition?

Answer: The optimal max depth by the model is 4, which was also my selection.

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 predicted value of client's home is 21.630, which is slightly less than the mean (22.533) a very close to the median (21.2).

Question 12 (Final Question): 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: This model can be used in predicting the selling price of future clients' homes as the error obtained by using the optimum parameter is not very high. However, the model can be improved by using better machine learning algorithms.