UDACITY

PROJECT

Predicting Boston Housing Prices

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

Dear student,

well done with your good submission. There are only a few issues to be addressed in order to meet requirements, please refer to my comments in the appropriate section for some hints. I've left some Pro Tips as well in case you might be interested in learning more about some specific topics.

Keep up your good work!

Data Exploration

✓

All requested statistics for the Boston Housing dataset are accurately calculated. Student correctly leverages NumPy functionality to obtain these results.

✓

Student correctly justifies how each feature correlates with an increase or decrease in the target variable.

Very good!

Developing a Model

Student correctly identifies whether the hypothetical model successfully captures the variation of the target variable based on the model's R^2 score.
The performance metric is correctly implemented in code.

Student provides a valid reason for why a dataset is split into training and testing subsets for a model.
Training and testing split is correctly implemented in code.

Analyzing Model Performance

Student correctly identifies significant qualities of the training and testing errors as the training set size increases.
Alternatively, student correctly identifies the trend of both the training and testing curves from the graph as more training points are added. Discussion is made as to whether additional training points would benefit the model.

Please discuss in more detail the behavior of both training and testing curves, at the chosen max depth, when more data is added. For each curve please describe if they are initially increasing or decreasing (and if that increase/decrease is sharp way or not) and if they eventually plateau and when.

Student correctly identifies whether the model at a max depth of 1 and a max depth of 10 suffer from either high bias or high variance, with justification using the complexity curves graph.

Student picks a best-guess optimal model with reasonable justification using the model complexity graph.

The answer and the rationale that are not correct: The best model that generalises the dataset is to be found at the max_depth where the validation R squared is maximised: Increasing complexity does not improve the R squared, most of the times it actually reduces it. We always look for the least complex model that maximises the R squared.

Evaluating Model Performance

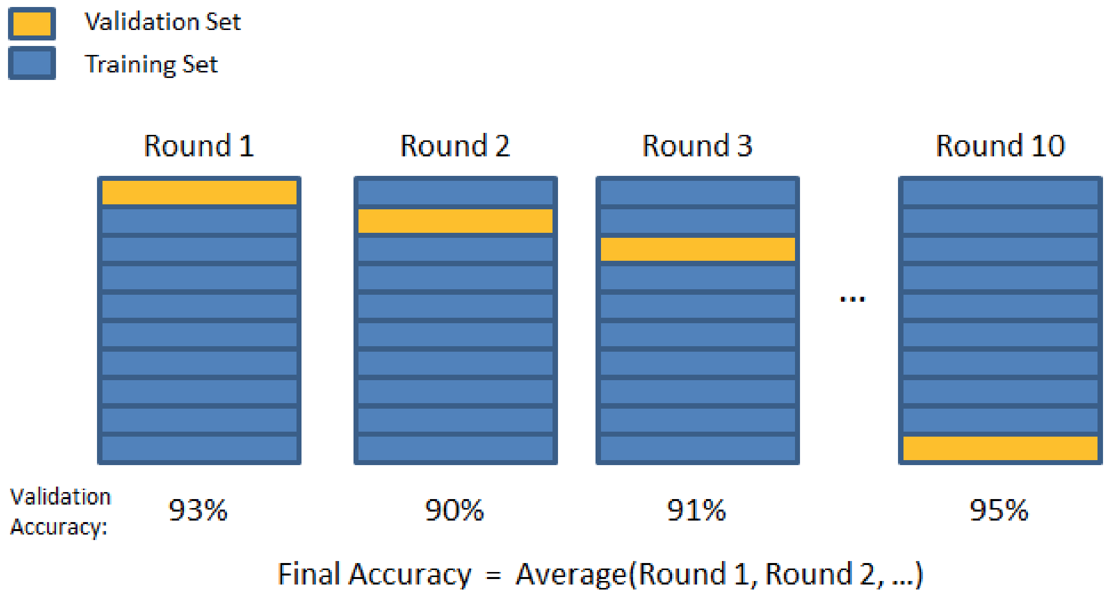
Student correctly describes the grid search technique and how it can be applied to a learning algorithm.

Pro tip: There are other techniques that could be used for hyperparameter optimization in order to save time like RandomizedSearchCV, in this case instead of exploring the whole parameter space just a fixed number of parameter settings is sampled from the specified distributions. This proves useful when we need to save time but is not necessary in cases in cases like ours where the data set is relatively small.

Student correctly describes the k-fold cross-validation technique and discusses the benefits of its application when used with grid search when optimizing a model.

Please discuss more in detail how cross validation works,some hints:

Cross validation is an iterative process where train/test sets are randomly generated multiples times in order to evaluate the algorithm at each split, the results are then averaged over the splits. When a dataset is limited in size cross validation becomes extremely useful as it allows for an extensive exploitation of available data allowing assessing the real potential of our algorithm in terms of performance metrics.
Over the benefits of cross validation:
http://www.anc.ed.ac.uk/rbf/intro/node16.html
There are several cross validation techniques, you can find some interesting info here:
http://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-iterators



Student correctly implements the `fit_model` function in code.

Student reports the optimal model and compares this model to the one they chose earlier.

Student reports a valid predicted selling price for the client's data and adequately justifies the prediction using the earlier calculated statistics.

Alternatively, if three clients are listed, discussion is made for each client as to whether these prices are reasonable given the data and the earlier calculated statistics.

Pro tip: To assess if your prediction is reasonable, besides from comparing it with the median, the mean and checking if it is included in one standard deviation range, you could use SKlearn to find the nearest neighbours of the feature vector. You can then contrast your results with the closest neighbours, the ones that have similar characteristics.

```
from sklearn.neighbors import NearestNeighbors
num_neighbors=5
def nearest_neighbor_price(x):
    def find_nearest_neighbor_indexes(x, X):  # x is your vector and X is
 the data set.
        neigh = NearestNeighbors( num_neighbors )
        neigh.fit(X)
        distance, indexes = neigh.kneighbors( x )
        return indexes
    indexes = find_nearest_neighbor_indexes(x, features)
    sum_prices = []
    for i in indexes:
        sum_prices.append(prices[i])
    neighbor_avg = np.mean(sum_prices)
    return neighbor_avg
print nearest_neighbor_price( [4, 55, 22])
index = 0
for i in client_data:
    val=nearest_neighbor_price(i)
    index += 1
    print "The predicted {} nearest neighbors price for home {} is: ${:,.2
f}".format(num_neighbors,index, val)
```

http://scikit-learn.org/stable/modules/neighbors.html#finding-the-nearest-neighbors

✓

Student thoroughly discusses whether the model should or should not be used in a real-world setting.

RESUBMIT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review

Student FAQ