UDACITY

‹ Return to "Machine Learning Engineer Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Predicting Boston Housing Prices

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Udacity student,

your project shows how committed you are to the course. You probably spent hours or days on this project and really should be proud of your work here and I'm proud of being part of your journey!

Keep the great work on the next sections of this amazing Udacity course!!!

I share with you some extra links from Medium:

A guide to start your path in Data Science and Machine Learning:

Fundamental Python Data Science Libraries

This last one is not for only a job interview, but it contains a lot of useful information about some great topics for a machine learning professional:

Data Science and Machine Learning Interview Questions

😄

Data Exploration

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

Correct! 😄

You have answered all statistics questions using the `Numpy` library. Nice job! 👍

It is important here to know why Udacity ask students to use the `NumPy` library:

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Numpy Documentation

It's always important to be aware of tools you use. For example, the Pandas' `Series.std()` will by default give you different result than `numpy.std()`. It's because **Numpy** takes in count the whole population while **pandas** assumes that you are evaluating the standard deviation for a sample of your dataset.

This article has a very good explanation about it.

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

Correct! 😄

You have correctly justified how each feature correlates with an increase or decrease in the target variable.

In addition, you can also plot your data to confirm your intuition and practice some coding skills using a library called `matplotlib`

```
import matplotlib.pyplot as plt
plt.figure(figsize=(15, 5))
for i, col in enumerate(features.columns):
    plt.subplot(1, 3, i+1)
    plt.plot(data[col], prices, 'x')
    plt.title('%s x MEDV' % col)
    plt.xlabel(col)
    plt.ylabel('MEDV')
```

Matplotlib Documentation

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 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.

Perfect! 😄

As more data we add the training score decreases and the testing score increases.

Moreover, adding even more points will not benefit the model and only "make the computer work harder" in terms of processing the data. I recommend this reading about this topic: "How much data is enough?".

Getting more training points may be hard and require a lot of additional work. More data also requires more computing resources or making performance improvements (this is not a problem in this case since the training set is small). Getting a dataset with more features, choosing a more complex model or increasing the max_depth hyperparameter is likely to produce better improvements than getting more training data.

In a short medium article you may go deeper in your studies about learning curves.

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.

Awesome! You picked one of the max depth available and made your guess.

Evaluating Model Performance

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

Correct! 😄

You show really a very good understanding about the Grid Search.

I'd like just to point out that Grid Search will test EVERY combination of hyperparameters. So it can be very computationally expensive if you want a model with many hyperparameters or many sets of them.

Usually I have some tips and links about Grid Search and I'd like to share with you. It can be good and complete your studies:

1. Official sklearn page on gridSearch
2. How gridSearch works
3. Specifying multiple metrics for evaluation
4. The scoring parameter: defining model evaluation rules
5. Defining your scoring strategy from metric functions

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.

Correct! 😄

You summarized well the K-fold concepts. I really liked your explanation and it shows that you understood this important machine learning tool.

The k-fold is a way to validate the parameters you've chosen for your model. It divides the trainset in k subsets and train the model in k-1 of them and evaluate it in the left set. To make it a stronger measure, it'll repeat the process k times, using a different subset from the k ones as the evaluation one, finally, takes the mean of the score used to evaluate the hyperparameters used in the k-fold cross-validation.

The benefits of grid search using k-fold cross-validation:

- Validation itself (it would be an advantage even with a single validation set), that keeps the test set really unseen. If we optimize the hyperparameters using an evaluation measure in the test set, it would kind leak information to the model, we may overfit the parameters to the test set because we're choosing them looking for a measure in it. And that's not good! The final result, the model assessment will be compromised, it would overestimate the generalization power of our model;
- The k-fold makes a robust measure taking the mean of a measure in different validation sets. The greater the k, more robust the measure;
- It's a good technique to use when we don't have many data points and couldn't split the data into a single and big validation set;

I'd like to share some extra readings about K-fold:

- What is Cross Validation?
- K-fold and Cross Validation

Student correctly implements the `fit_model` function in code.

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

Correct! 😄

Student reports the predicted selling price for the three clients listed in the provided table. Discussion is made for each of the three predictions as to whether these prices are reasonable given the data and the earlier calculated descriptive statistics.

Correct! 😄

The selling prices you evaluated are correct and your discussion shows your very good understanding about this project. Also, it shows that you can extract insights from data using your intuition and abilities.

You can also plot the data and the predictions to have an intuition about the model:

```
prediction_data = np.transpose(client_data)
pred = reg.predict(client_data)
for i, f in enumerate(house_features):
    plt.scatter(features[f], prices,alpha=0.25, c='green')
    plt.scatter(prediction_data[i], pred, color='red', marker='D')
    plt.xlabel(f)
    plt.ylabel('MEDV')
    plt.show()
```

Good job!

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

DOWNLOAD PROJECT

RETURN TO PATH