UDACITY

Back to Machine Learning Engineer Nanodegree

Predicting Boston Housing Prices

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Hello Udacian,
It was a great pleasure to me to review such great job. I can see you did read the suggestions and comments from the previous review and followed the required guidelines to adjust your answer carefully. I added some comments and pro tips and I hope they will be helpful.

The work now meets specifications. Congratulation. You made it. ☺

Carry on your Great Work!

Data Exploration

✓

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

Great job briefly exploring the Housing dataset and calculating it's key statistics! 😎

Pro Tips

- Checking your dataset statistics is an very useful routine in applying a predictive model. This is because:
    - It helps us to check if the key assumptions of our algorithms hold (thereby helping us choose which model to apply).

    - These statistics tend to be very handy when you obtain a prediction, to check whether the predictions are reasonable, and not off-chart, compared to central values of the dataset.
- NumPy as a library might have been new for you, and not that easy to learn. In this tips section I'll give you some tips you can use when learning and picking up a new library:
    - Two functions are very useful when investigating a module (library) or a simple Python Object: the **doc** functionality and the dir() functionality.
    - If you wish to rapidly explore documentation of a library/module/function/object, you can just type print obj.**doc**, and the documentation of the function will be printed.
    - If you wish to rapidly explore what attributes and functions are available for an object, you can just type dir(obj), and you'll get a Python `list` of the object's attributes and functions.
    - Remember to always read documentation and try examples in your interpreter if you feel confused about a new library.
    - Hopes these help in your future Machine Learning Endeavors!
- Utilizing numPy is quite common when we are doing some statistical analysis no matter we are doing machine learning or data analysis. Thus, it is always useful for us to learn more the function inside numPy. Here is a course in udacity actually teaching us how to use it : intro to data analysis watch it if you want to learn more about it.
- Here is a website which provide lots of example work about most common used numpy function

✓

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

You clearly understand the relationship between each parameter with the house price. Great job!

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.

Great, you successfully executed the `R^2` score function and your discussion here is great! 😄

✓

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.

You gave great reasons for splitting a dataset, and nice implementation using sklearn's `train_test_split` !

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.

Amazing job describing how the training and testing score change as the training set size increases. You're right! 😎

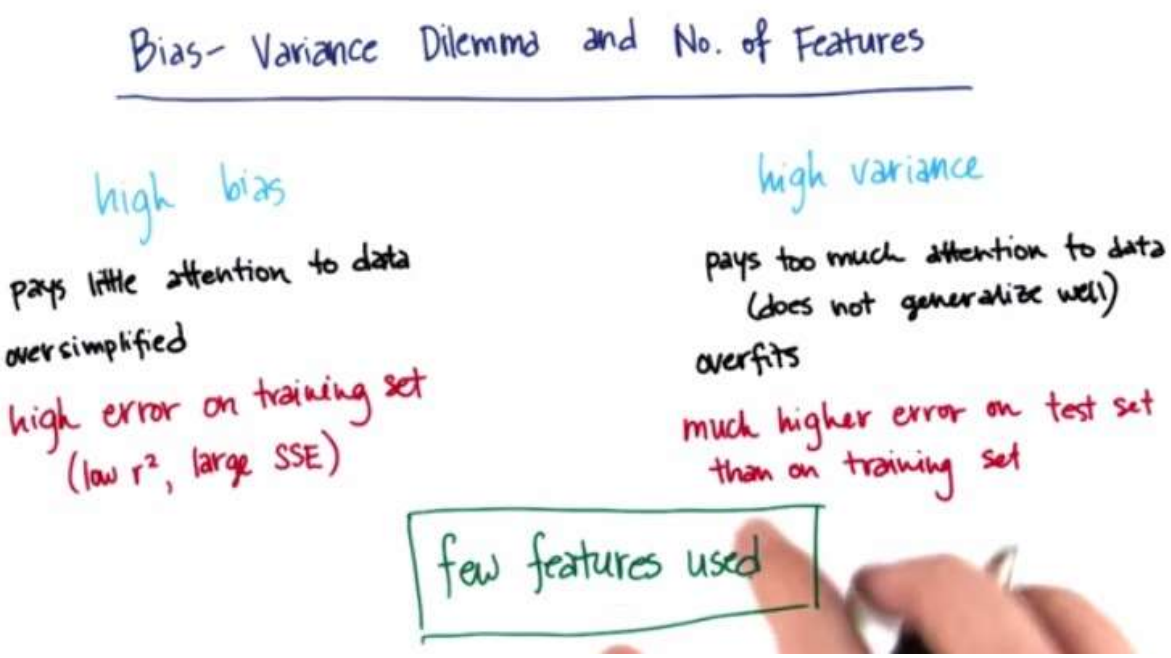
✓

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.

Great job identifying that the model suffers from high bias when `max_dept = 1` and that the model suffers from overfitting (high variance) when the `max_dept = 10`.

Suggestions and Comments:

- Please check out this link if you want to understand more about model bias-variance tradeoff.



✓

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

Nice rationale in guessing the best-guess optimal max depth for your work!

Evaluating Model Performance

✓

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

Awesome explanation of the grid search algorithm! 😄

Pro Tips:

- Another very powerful parameter tuning algorithm is RandomizedSearchCV. In contrast with GridSearchCV, not all parameters are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions.
- One particular advantage of RandomizedSearchCV is that it is much faster than GridSearchCV, and it is theoretically proven to find models that are as good; or even better than grid search.

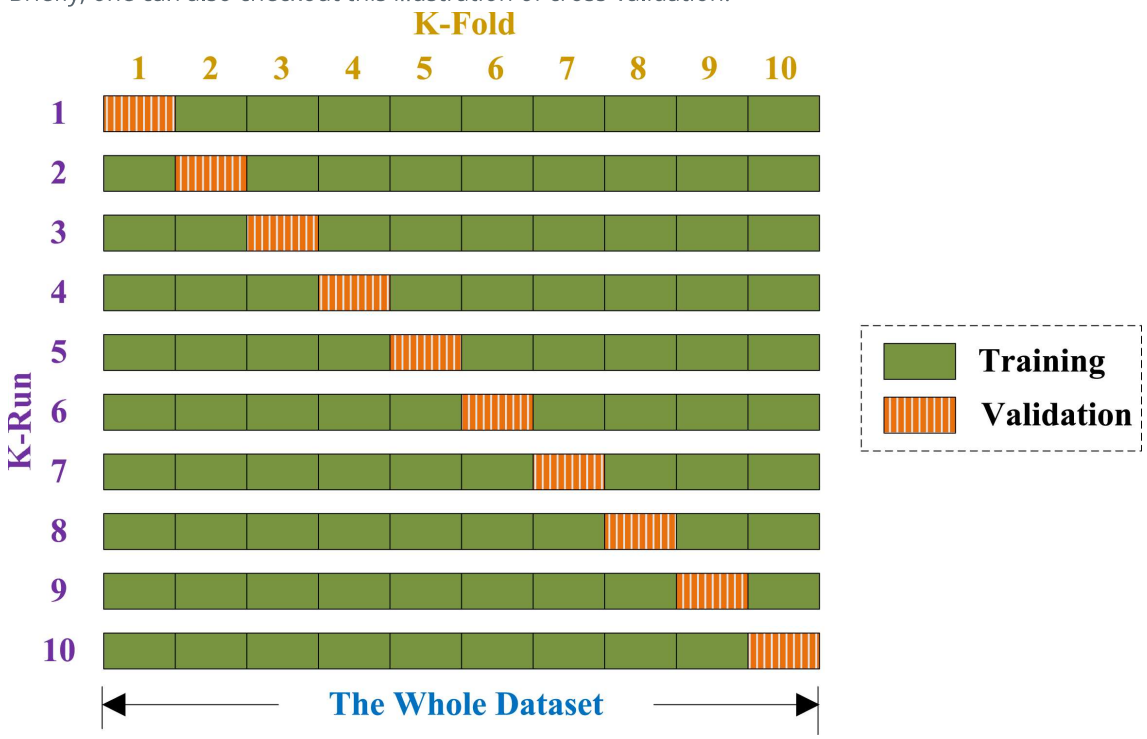
✓

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.

- Amazing description of k-fold cross validation and how it is performed on a model!
- I have greatly appreciated your illustration of cross-validation, which in my opinion very well captures how it works. 👌

Suggestions and Comments:

- Briefly, one can also checkout this illustration of cross-validation.



- Below is an explanation of how cross-validation works, with guidance from the diagram shown to the above:
  - As you can see from the diagram, the training set is divided into K-folds, or k subsets. For this particular diagram, it is 10 subsets.
  - The model is then trained and validated k times, or 10 times for this particular example.
  - At each run, one subset or fold is held out at validation set, and the other k-1 folds are used for training
  - At the end, the validation scores are collected and averaged out, to get the final score of the model being tested.
- One particular advantage of validating a model in this way is that it makes particular good use of the data available, especially if the dataset is small. So it can help mitigate overfitting.
- Also, as stated on the sklearn page for cross-validation, if a single set is used for testing, and parameter tuning, then information can leak away from the only test set into the model being tuned. So, using multiple test/validation sets can help mitigate this.

- Hope this helps you to understand what is cross-validation and how it works, as this is a quite important concept in Machine Learning.

Please note that, we are not restricted to use 10 fold Cross Validation.

✓

Student correctly implements the `fit_model` function in code.

Stellar implementation! 😎

✓

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

Great job here! Your `max_depth` matched exactly the best-guess optimal model `max_depth` you gave earlier.

✓

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.

Well done predicting, these are valid prices for your clients' houses. And also, your discussions on whether these prices are reasonable or not, sound great. 😄

✓

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

You made a very thorough and logical discussion here!

DOWNLOAD PROJECT

RETURN TO PATH