DSC550-T301 Data Mining (2235-1)   Week 9: Hyperparameter Tuning

Week 9: Hyperparameter Tuning

Introduction

Contents of the Week

- Introduction
- Readings
- Supplemental Materials
- 9.1 Discussion/Participation
- 9.2 Exercise: Best Model Selection and Hyperparameter Tuning

Topics

- A Model selection
- B Hyperparameter tuning
- C Pipelines
- D Grid search

Readings

Read the following:
- Reread/review Chapters 11, 12, and 15 of *Machine Learning with Python Cookbook*

Supplemental Materials

All of the materials below are from external sources. Authorship and ownership are indicated within the sources themselves.

Readings 📖	**Videos** 🎞

🌐 [How To Fine Tune Your Machine Learning Models To Improve Forecasting Accuracy](#)

🌐 [Hyperparameter tuning for machine learning models](#)

🌐 [Hyperparameter Tuning](#)

🌐 [sklearn.pipeline.Pipeline](#)

🌐 [sklearn.model_selection.GridSearchCV](#)

🌐 [A Simple Guide to Scikit-learn Pipelines](#)

🌐 [Hyperparameter Tuning with GridSearchCV](#)

🌐 [What is the Difference Between a Parameter and a Hyperparameter?](#)

🌐 [Improve Your Model Performance using Cross Validation (in Python and R)](#)

🌐 [5 Reasons why you should use Cross-Validation in your Data Science Projects](#)

9.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is the difference between a parameter and a hyperparameter of a model?

2. What Python libraries are useful for hyperparameter tuning?

3. What is cross-validation? Why is this useful?

4. Is a validation set the same as a test set?

5. What are some key hyperparameters in various machine learning models?

9.2 Exercise: Best Model Selection and Hyperparameter Tuning

In this exercise, you will work with the Loan_Train.csv dataset which can be downloaded from this link: [Loan Approval Data Set](#).

1. Import the dataset and ensure that it loaded properly.
2. Prepare the data for modeling by performing the following steps:

    - Drop the column "Load_ID."
    - Drop any rows with missing data.
    - Convert the categorical features into dummy variables.

3. Split the data into a training and test set, where the "Loan_Status" column is the target.
4. Create a pipeline with a min-max scaler and a KNN classifier (see section 15.3 in the *Machine Learning with Python Cookbook*).
5. Fit a default KNN classifier to the data with this pipeline. Report the model accuracy on the test set. **Note**: Fitting a pipeline model works just like fitting a regular model.
6. Create a search space for your KNN classifier where your "n_neighbors" parameter varies from 1 to 10. (see section 15.3 in the *Machine Learning with Python Cookbook*).
7. Fit a grid search with your pipeline, search space, and 5-fold cross-

Submission Instructions

Click the title above to submit your assignment.

This exercise is due by Sunday 11:59 PM.

Submit your code, output, and answers at the link above. Comment all your code and answer any questions that are asked in the instructions. It is perfectly fine to answer a question by displaying output from your code, but make sure you are displaying the appropriate output to answer the question. I would recommend using and submitting a Jupyter Notebook, but this is not required.

View the rubric for this Assignment by clicking on the link below:

Exercise Rubric

validation to find the best value for the "n_neighbors" parameter.

8. Find the accuracy of the grid search best model on the test set. **Note**: It is possible that this will not be an improvement over the default model, but likely it will be.

9. Now, repeat steps 6 and 7 with the same pipeline, but expand your search space to include logistic regression and random forest models with the hyperparameter values in section 12.3 of the *Machine Learning with Python Cookbook*.

10. What are the best model and hyperparameters found in the grid search? Find the accuracy of this model on the test set.

11. Summarize your results.