A primer on scikit-learn

By Ted Petrou

Pandas vs Scikit-Learn

- Pandas and scikit-learn are two related, but different libraries.
- Pandas is likely where the vast majority of your workflow will reside when performing a data analysis.
- Once you have thoroughly explored and investigated your data and have prepared it for machine learning, you then turn to scikit-learn to train and test your models.

Scikit-Learn

- Scikit-Learn is the most popular Python library to build basic machine learning models
- It is easy to use and can train a model in 3 lines of code
- Does not focus on Deep Learning (Use TensorFlow or Keras instead)
- Built on top of NumPy
- In addition to training machine learning models Scikit-Learn provides a host of other tools for data preprocessing and model evaluation

Scikit-Learn Vocabulary

- Input data (**samples)** - usually given variable name **X**
- Each column of this array is a **feature**
- Each row is a **sample** (or observation)
- Scikit-Learn uses the term **estimator** for the Python object that learns from the data. This is our **model**.
- In supervised learning, each sample (row) is **labeled** - a.k.a **target**.

Scikit-Learn Gotchas

- No missing data
- Input data must be in a numeric 2-d array. Even if there is one feature it must be a 2d array
- No string data - must encode as numeric

Segregate the Scikit-Learn API

- There are two separate types of objects in the scikit-learn API, **estimators** and **helper functions**
- Estimator is simply the term that scikit-learn uses for all of the objects that do the machine learning.
- The helper functions do not do machine learning, but provide support for things such as scoring, evaluation, and random data generation.
- Let’s take a look at the API now

Importing Scikit-Learn into the Workspace

- It is composed of a couple dozen modules which we import our estimators and functions from.
- By convention you will see imports like this:

```
from sklearn.some_module import SomeEstimator

from sklearn.some_module import some_function
```

Basic steps for building an ML model with Scikit-Learn

- Import the estimator
- Instantiate the estimator
- Train the model with the fit method
- Predict and score with the trained model
- The API is consistent throughout. Easy to use