CHAPTER 18

Introduction to Scikit-learn

Scikit-learn is a Python library that provides a standard interface for implementing machine learning algorithms. It includes other ancillary functions that are integral to the machine learning pipeline such as data preprocessing steps, data resampling techniques, evaluation parameters, and search interfaces for tuning/optimizing an algorithm's performance.

This section will go through the functions for implementing a typical machine learning pipeline with Scikit-learn. Since, Scikit-learn has a variety of packages and modules that are called depending on the use case, we'll import a module directly from a package if and when needed using the **from** keyword. Again the goal of this material is to provide the foundation to be able to comb through the exhaustive Scikit-learn library and be able to use the right tool or function to get the job done.

Loading Sample Datasets from Scikit-learn

Scikit-learn comes with a set of small standard datasets for quickly testing and prototyping machine learning models. These datasets are ideal for learning purposes when starting off working with machine learning or even trying out the performance of some new model. They save a bit of the time required to identify, download, and clean up a dataset obtained from the wild. However, these datasets are small and well curated, they do not represent real-world scenarios.

Five popular sample datasets are

- Boston house-prices dataset
- Diabetes dataset

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_18