Scikit-Learn Python Package

What is Scikit-Learn?

- Scikit-learn (Sklearn) is a powerful and robust open-source machine learning library for Python.
- Sklearn provides tools for efficient implementation of classification, regression, clustering and dimensionality reduction techniques.
- Sklearn is written on top of NumPy, SciPy and Matplotlib packages.
- Basic knowledge of these packages plus Pandas is required to successfully implement machine learning models using Sklearn.

What is Scikit-Learn?

- Sklearn is a community project and anyone can contribute to it.
- Currently there are more than 2058 contributors on its github repository.
- Various organizations including Booking.com, JP Morgan, Evernote, Spotify use Sklearn
- Sklearn library is easy to use, offers tons of flexibility and has a very good documentation for both beginners and experts alike.

Origins of Sklearn

- 2007: Initially developed by David Cournapeau as Google summer of code project
- 2010: French Institute for Research in Computer Science and Automation took this to another level as they made the first public release of v0.1
- 2021: the latest sklearn version is 0.21.0 after 12 versions of iterations and improvement

Data Modelling

- Sklearn is focused on modelling data and offers plethora of tools for that:
  - Supervised machine learning algorithms
  - Unsupervised machine learning algorithms
  - Clustering
  - Cross validation
  - Dimensionality reduction
  - Ensemble methods
- Sklearn also offered preprocessing support:
  - Data encoding
  - Feature selection / extraction

Machine Learning

- Machine Learning (ML) is a study of algorithms that can learn to solve a specified task using data.
- ML models are trained using a sample of historical data called the training data and the model itself is evaluated based on its performance on an unseen data called the test data.
- ML has wide variety of application from research to health to finance to speech recognition and language translation.

Machine Learning

- There are two main types of ML models:
  1. Supervised:
     - Model learns to identify pattern in data using inputs and desired outputs called labels.
  2. Unsupervised:
     - Model learns to identify pattern and structure in the data without any labels

Install Sklearn using Anaconda

- conda install scikit-learn
- Prerequisite packages are installed

Additional resources

- Tutorials:
  - Quick Start Tutorial http://scikit-learn.org/stable/tutorial/basic/tutorial.html
  - User Guide http://scikit-learn.org/stable/user_guide.html
  - API Reference http://scikit-learn.org/stable/modules/classes.html
  - Example Gallery http://scikit-learn.org/stable/auto_examples/index.html
  - PyCon 2014 Scikit-learn Tutorial by Jake VanderPlas
  - Parallel Machine Learning with scikit-learn and IPython by Olivier Grisel (also offered at Strata 2014)

- Books:
  - Learning scikit-learn: Machine Learning in Python (2013)
  - Building Machine Learning Systems with Python (2013)
  - Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (2014).