Introduction to Scikit-Learn

Overview

- Scikit-learn (Sklearn) is the most popular and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

- It is built upon Numpy, SciPy and matplotlib frameworks.

History

- It was originally called **scikits.learn** and was initially developed by David Cournapeau as a Google summer of code project in 2007. Later, in 2010, Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and Vincent Michel, from FIRCA (French Institute for Research in Computer Science and Automation), took this project at another level and made the first public release (v0.1 beta) on 1st Feb. 2010.

Installation Requirements

- Before we start using scikit-learn latest release, we require the following −
- Python (>=3.5)
- NumPy (>= 1.11.0)
- Scipy (>= 0.17.0)
- Joblib (>= 0.11)
- Matplotlib (>= 1.5.1) is required for Sklearn plotting capabilities.
- Pandas (>= 0.18.0) is required for some of the scikit-learn examples using data structure and analysis.

Installation Procedure

- Using Pip (Preferred Installer Program)
    - pip install scikit-learn
- Using Anaconda or Miniconda (a data science toolkit)
    - conda install scikit-learn

Scikit-Learn API Features

- sklearn.base (Base classes and utility functions)
- sklearn.calibration (Probability Calibration)
- sklearn.cluster (Clustering)
- sklearn.compose (Composite Estimators)
- sklearn.covariance (Covariance Estimators)
- sklearn.cross_decomposition (Cross Decomposition)
- sklearn.datasets (Datasets)
- sklearn.decomposition (Matrix Decompostion)
- sklearn.discriminant_analysis (Discriminant Analysis)
- sklearn.dummy (Dummy estimators)
- sklearn.ensemble (Ensemble Methods)
- sklearn.exception (Exceptions and warning)
- skearn.experimental (Experimental)
- sklearn.feature_extraction (Feature Extraction)
- sklearn.feature_selection (Feature Selection)
- sklearn.gaussian_process (Gaussian Processes)
- sklearn.impute (Impute)
- sklearn.inspection (Inspection)
- sklearn.isotonic (Isotonic regression)
- sklearn.kernel_approximation (Kernel Approximation)

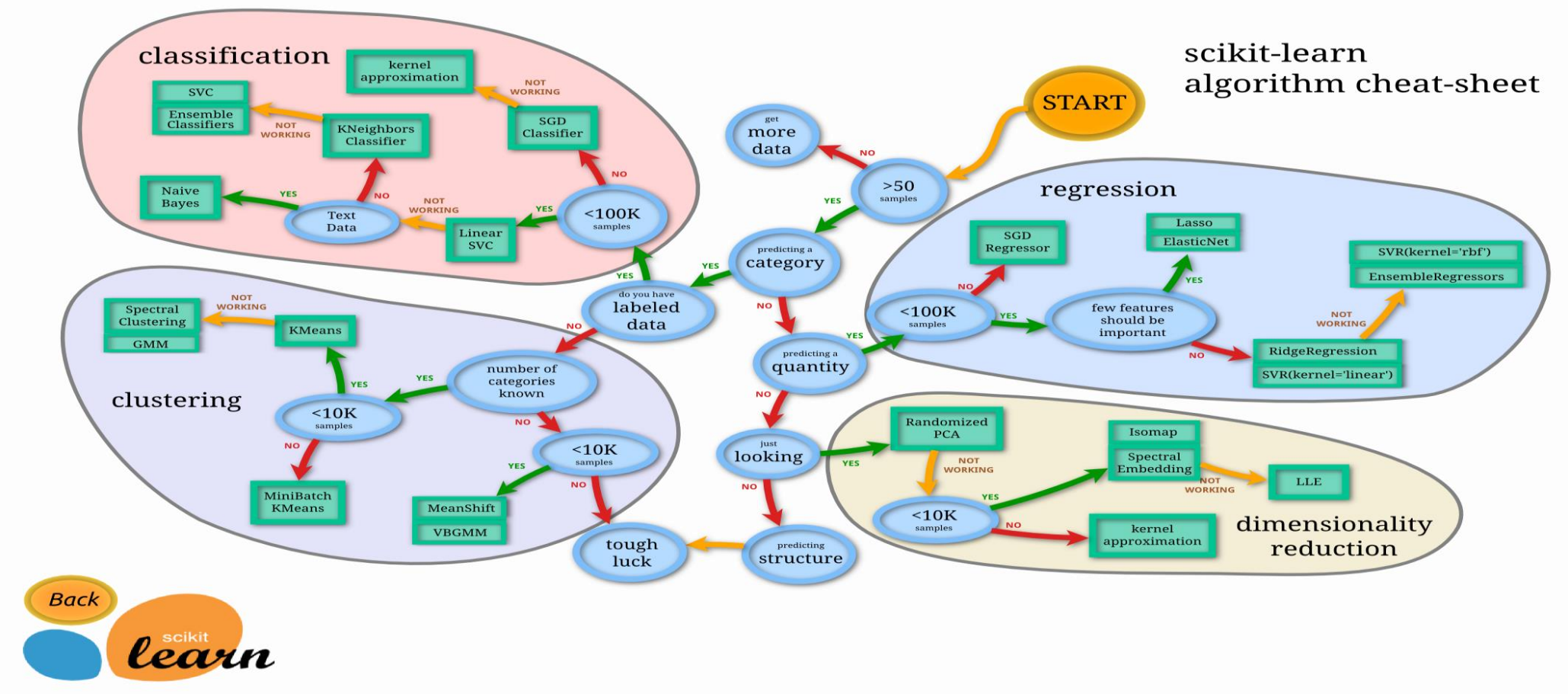
Scikit-Learn API Features Cont’d

- sklearn.kernel_ridge (Kernel Ridge Regression)
- sklearn.linear_model (Linear Models)
- sklearn.manifold (Manifold Learning)
- sklearn.metrics (Metrics)
- sklearn.mixture (Gaussian Mixture Models)
- sklearn.model_selection (Model Selection)
- sklearn.multiclass (Multiclass Classification)
- sklearn.multioutput (Multioutput Regression and Classification
- sklearn.naive_bayes (Naïve Bayes)
- sklearn.neighbors (Nearest Neighbors)
- sklearn.neural_network (Neural Network Models)
- sklearn.pipeline (Pipeline)
- sklearn.preprocessing (Preprocessing and Normalization)
- sklearn.random_projection (Random Projection)
- sklearn.semi_supervised (Semi-Supervised Learning)
- sklearn.svm (Support Vector Machines)
- sklearn.tree (Decision Trees)
- sklearn.utils (Utilities)

Common Features Explained

- Rather than focusing on loading, manipulating and summarizing data, Scikit-learn library is focused on modeling the data. Some of the most popular groups of models provided by Sklearn are as follows −
- **Supervised Learning algorithms** − Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.
- **Unsupervised Learning algorithms** − On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.
- **Clustering** − This model is used for grouping unlabeled data.
- **Cross Validation** − It is used to check the accuracy of supervised models on unseen data.
- **Dimensionality Reduction** − It is used for reducing the number of attributes in data which can be further used for summarization, visualization and feature selection.
- **Ensemble methods** − As name suggest, it is used for combining the predictions of multiple supervised models.
- **Feature extraction** − It is used to extract the features from data to define the attributes in image and text data.
- **Feature selection** − It is used to identify useful attributes to create supervised models.

Guide to choosing the right estimator in Scikit-Learn



Companies using Scikit-Learn

- JPMorgan
- Spotify
- Inria
- Betaworks
- Hugging Face
- Evernote
- Telecom ParisTech
- Booking.com
- Aweber
- Yhat
- Rangespan
- Birchbox
- Change.org
- DataRobot
- OkCupid
- Lovely etc.

Useful Links

- A bit of data science and scikit-learn (https://github.com/knathanieltucker/bit-of-data-science-and-scikit-learn/tree/master/notebooks)
- Machine Learning Mastery (https://machinelearningmastery.com/start-here)
- Global AI Hub GitHub repository for machine learning using scikit-learn (https://github.com/KutayAkalin/ML_Course_19-11-20)
- Top 10 Popular GitHub Repositories to learn about Data Science (https://towardsdatascience.com/top-10-popular-github-repositories-to-learn-about-data-science-4acc7b99c44)