Scikit from the article

David Arroyo Menéndez

March 8, 2020

Presentation

Scikit-learn: Machine Learning in Python

- Fabian Pedregosa
- Gaël Varoquaux
- Alexandre Gramfort
- Vincent Michel
- Bertrand Thirion
- More

URL: `http://www.jmlr.org/papers/v12/pedregosa11a.html`
Citations: 15926

Abstract

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language. Emphasis is put on ease of use, performance, documentation, and API consistency. It has minimal dependencies and is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings. Source code, binaries, and documentation can be downloaded from `http://scikit-learn.sourceforge.net`.

Introducción

- Lenguaje de programación Python
- Qué hace. Necesidad de hacerlo
- No gap en "context + gap + technical solution"

Context

The Python programming language is establishing itself as one of the most popular languages for scientific computing. Thanks to its high-level interactive nature and its maturing ecosystem of scientific libraries, it is an appealing choice for algorithmic development and exploratory data analysis (Dubois, 2007; Milmann and Avaizis, 2011). Yet, as a general-purpose language, it is increasingly used not only in academic settings but also in industry.

Tecnical Solution

Scikit-learn harnesses this rich environment to provide state-of-the-art implementations of many well known machine learning algorithms, while maintaining an easy-to-use interface tightly inte- grated with the Python

Project Vision

En NLTK es Design Criteria, pero el enfoque es diferente

- Code Qualitiy
- BSD Licensing
- Bare-bone design and API
- Community-driven development
- Documentation

Underlying Technologies

Este apartado no lo tiene NLTK, pero me parece muy interesante

- Numpy
- Scipy
- Cython

Code Design

- Objects specified by interface, not by inheritance
- Benchmarking between another machine learning tools
- Model Selection: cross-validation (GridSearch and Lasso).

Pipeline can combine several transformers

High-level yet Efficient: Some Trade Offs

- SVM
- LARS
- Elastic Net
- kNN
- PCA

Conclusion

- Resumen de secciones
- Qué hace el software
- Future work