Scikit-learn como librería de machine learning

Scikit-learn es una librería de Python de aprendizaje automático que contiene una serie de herramientas para realizar análisis y minería de datos. Para instalar dicho módulo es necesario Python 2.6 o superior y la librería NumPy(3.5.5) versión 1.6.1 o superior.

Librería construída sobre Scipy que contiene implementados muchos de los algoritmos de machine learning.

Scikit-learn es la principal librería que existe para trabajar con Machine Learning , incluye la implementación de un gran número de algoritmos de aprendizaje. La podemos utilizar para clasificaciones , extracción de características , regresiones , agrupaciones , reducción de dimensiones , selección de modelos , o preprocesamiento . Posee una API que es consistente en todos los modelos y se integra muy bien con el resto de los paquetes científicos que ofrece Python . Esta librería también nos facilita las tareas de evaluación, diagnóstico y validaciones cruzadas ya que nos proporciona varios métodos de fábrica para poder realizar estas tareas en forma muy simple.

Introducción sklearn: http://scikit-learn.org/stable/tutorial/basic/tutorial.html

Instalación: http://scikit-learn.org/stable/install.html

Introducción a scikit-learn

Librería que proporciona un amplio conjunto de algoritmos de aprendizaje supervisado y no supervisado a través de una consistente interfaz en Python.

Esta librería se ha construido sobre SciPy (Scientific Python), que debe ser instalada antes de utilizarse, así como:

- NumPy
- SciPy
- Matplotlib
- SymPy
- Pandas

Características

Esta librería se centra en el modelado de datos y no en cargar y manipular los datos. Para estos fines, es mejor utilizar NumPy y Pandas. Algunas cosas que aporta Scikit-Learn son:

- Algortimo de Clustering.
- Validación cruzada.
- Datasets de prueba.
- Selección de características
- Tunning de parámetros de la mayoría de los algoritmos de machine learning

Ventajas vs Desventajas respecto a otras librerías

Ventajas

- Interfaz consistente ante modelos de machine learning.
- Proporciona muchos parámetros de configuración.
- Documentación excepcional.
- Desarrollo muy activo.
- Comunidad.

Desventajas

- Más difícil que R.
- R proporciona más facilidades para entender los modelos que se están construyendo

Instalación de scikit-learn

Scikit-learn depende de otros dos paquetes como NumPy y SciPy. Para el trazado y el desarrollo interactivo, también debe instalar matplotlib, IPython y el cuaderno Jupyter.

Recomendamos utilizar uno de las siguientes distribuciones, que proporcionará los paquetes necesarios.

Si ya tiene instalada una instalación de Python, puede usar pip para instalar todos estos paquetes:

$ pip install numpy scipy matplotlib ipython scikit-learn pandas

Scikit-learn depende de NumPy y SciPy , de los cuales hemos hablado. Así que asegúrese de actualizar tanto a la última versión antes de instalar el paquete, lo que se hace, por supuesto, con el administrador de paquetes python.

$ pip install scikit-learn