Entornos de data science con Python
Bloque 0 – Guía de estudio
2015/16

Introducción

El uso de entornos de data science requiere el manejo de herramientas de computación interactiva y de amplias bibliotecas de métodos analíticos que constituyen la "caja de herramienta" del data scientist.

El primer concepto importante es el del **entorno de computación interactivo**, en nuestro caso, utilizando la interfaz de Notebooks de Jupyter. Hay que entender que la interfaz de Notebooks es una forma de hacer análisis de datos interactivo que soporta la exploración y la compartición de resultados. Un Notebook es una realización práctica de dos conceptos previos:

- "*Literate programming*", una idea de una forma de programar en la que se refleja el flujo de pensamiento del programador, enunciada originalmente por D. Knuth: https://en.wikipedia.org/wiki/Literate_programming
- "*Executable experiments*" (o "executable papers"), la idea de incluir los datos y el código de los análisis como resultado de los artículos científicos (o de estudios en general). Algunas editoriales ya lo implementan, por ejemplo, Elsevier: https://www.elsevier.com/books-and-journals/content-innovation/executable-papers

Es importante comprender que Jupyter es una interfaz, un protocolo de comunicación y un conjunto de kernels que permiten hacer computación interactiva desde un navegador. Aunque históricamente Jupyter proviene del proyecto IPython, ya desde sus orígenes era independiente del lenguaje de programación.

Objetivos específicos

- Entender el stack para data science basado en Python y saber identificar sus partes.
- Saber utilizar las utilidades adicionales de IPython como entorno de computación interactiva.
- Saber utilizar el entorno Jupyter para data science a través de su interfaz de Notebooks.
- Comprender los elementos básicos de una instalación de Jupyter.
- Entender cómo se puede utilizar IPython para las diferentes fases de un estudio predictivo.

Recursos

Además de la información que se puede encontrar en las páginas de Jupyter (http://jupyter.org/) y de IPython (https://es.wikipedia.org/wiki/IPython) los siguientes son recursos interesantes para reforzar el estudio:

- Una visión general de IPython se proporciona en el capítulo 3 de PFDA[1].
- Alternativamente, hay una introducción a IPython en los capítulos 1 y 2 de LIICDV[2].
- Dentro del "pandas cookbook" de Julia Evans hay un documento "A quick tour of IPython Notebook" muy recomendable como breve recordatorio.

Para una visión más avanzada a la que se puede volver al final del módulo, la sesión PyCon 2015 "IPython & Jupyter in depth: high productivity interactive and parallel python" es muy recomendable. La primera hora aproximadamente es la parte básica de uso de Jupyter:

- https://www.youtube.com/watch?v=05fA_DXgW-Y

Actividades

- Trabajar alguno de los recursos mencionados u otros similares.
- Trabajar con Jupyter para acostumbrarse al entorno.

[1] McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.

[2] Rossant, C. (2013). *Learning IPython for interactive computing and data visualization*. Packt Publishing Ltd.