Máster en Business Analytics y Big Data

Edición 2015 - intensiva

Asignatura: ENTORNOS DE DATA SCIENCE en PYTHON
Módulo: DATA SCIENCE/HERRAMIENTAS DE ANÁLISIS
Coordinador: Miguel-Angel Sicilia, msicilia@uah.es

OBJETIVOS

El objetivo general del módulo es el de adquirir las habilidades para utilizar un entorno de data science interactivo. En este módulo se utiliza el lenguaje Python y el entorno IPython (ahora Jupyter).

El énfasis del módulo está en proporcionar habilidades de adquisición, preparación, transformación y manejo de datos que son esenciales para todas las asignaturas del bloque III (data science).

Los objetivos concretos del módulo son los siguientes resultados del aprendizaje:

1. Usar IPython (notebooks) como herramienta de trabajo interactiva en data science.
2. Cargar datasets y adquirir datos externos en diferentes formatos.
3. Dibujar (plotting) con propósitos de análisis
4. Ser capaz de transformar, mezclar y limpiar datasets.
5. Ser capaz de agrupar y resumir datos.
6. Ser capaz de aplicar bibliotecas estadísticas
7. Comprender cómo IPython puede utilizarse para computación paralela y trabajo en grupo

METODOLOGÍA

La metodología es completamente práctica, y se utilizará el entorno IPython/Jupyter como herramienta de trabajo. Los contenidos se expondrán mediante ejemplos en sesiones interactivas. Los estudiantes después utilizarán el mismo entorno para realizar ejercicios adicionales y adquirir práctica con el entorno y con las tareas básicas de data science.

Para ello, se facilitará una máquina virtual con una instalación del entorno IPython/Jupyter preparada para su uso.

Es importante resaltar que el módulo no está pensado para formar programadores Python, sino para formar data scientists que pueden utilizar Python de manera eficaz para sus propósitos analíticos.

PROGRAMA

Sesión 1: Computación interactiva con IPython/Jupyter
Actividades: Uso de IPython en consola. Familización con el entorno Notebook. Primer caso integrado simple de data science.
Materiales: IPython/Jupyter

Sesión 2: Manejo de arrays y matrices
Actividades: Uso de las bibliotecas NumPy y Pandas.
Materiales: IPython/Jupyter

Sesión 3: Adquisición de datos externos
Actividades: Uso de diferentes bibliotecas para obtener datos en diferentes formatos y incorporarlos al uso con SciPy.
Materiales: IPython/Jupyter

Sesión 4: Uso de bibliotecas estadísticas
Actividades: Uso de bibliotecas de statmodels, comparación con otras bibliotecas en SciPy.
Materiales: IPython/Jupyter

Sesión 5: Paralelización del manejo de matrices.
Actividades: Uso del entorno para paralelizar computaciones básicas, para comprender la arquitectura distribuida de IPython/Jupyter.
Materiales: IPython/Jupyter

MATERIALES

Texto básico:

- McKinney, W. (2012). Python for Data Analysis. Data Wrangling with Pandas, NumPy, and IPython O'Reilly Media

"Errata list" del libro:
http://www.oreilly.com/catalog/errata.csp?isbn=0636920023784

IPython project:
http://ipython.org/

ScyPy:
http://www.scipy.org/
La distribución ScyPy utilizada es esta:
https://store.continuum.io/cshop/anaconda/

Para saber más y estar al día:
http://pyvideo.org/

EVALUACIÓN

Niveles de consecución de los objetivos

Objetivo específico	*Nivel alto*	*Nivel medio*	*Nivel bajo*
1 – Usar IPython (notebooks) como herramienta de trabajo en data science.	Desarrollar análisis con el entorno más allá de los utilizados en clase.	Utilizar eficazmente el entorno IPython y saber localizar bibliotecas y funciones nuevas.	Entender y saber manejar de forma básica IPython para los ejemplos de clase.
2 – Cargar y adquirir datos en varios formatos	Utilizar partes de bibliotecas de adquisición de datos adicionales.	Realizar tareas de adquisición y transformación de datos adicionales	Ser capaz de reproducir y modificar los ejemplos de clase
3 – Dibujar (plotting) con propósitos de análisis	Uso avanzado de las bibliotecas de plotting	Utilizar representaciones que incluyan elementos no vistos en clase	Saber hacer representaciones básicas de datos como las vistas en clase.
4 - Ser capaz de transformar, mezclar y limpiar datasets.	Aplica técnicas de transformación y mezcla de manera sistemática y ordenada.	Aplica la fusión de datos con eficacia y razona las tareas de limpieza	Es capaz de reproducir tareas básicas de manejo de datasets
5 - Ser capaz de agrupar y resumir datos.	Es capaz de agrupar datos en varios niveles, manteniendo índices de diferentes tipos.	Utiliza de manera eficaz la agrupación y resumen en problemas no triviales.	Es capaz de reproducir tareas básicas de agrupación y resumen.
6 - Ser capaz de aplicar bibliotecas estadísticas	Es capaz de diferenciar y aplicar de forma avanzada diferentes bibliotecas de análisis estadístico.	Es capaz de encontrar en las bibliotecas las funciones estadísticas necesarias.	Es capaz de utilizar funciones de bibliotecas estadísticas.
7. Comprender cómo IPython puede utilizarse para computación paralela y trabajo en grupo	Es capaz de razonar y diseñar paralelizaciones para problemas reales.	Es capaz de hacer ejemplos básicos de paralelizaciones con matrices	Comprende la arquitectura paralela de IPython.

Modelo de evaluación

La siguiente tabla detalla los pesos de cada una de las actividades de evaluación. Todas las pruebas de evaluación son prácticas, consistentes en desarrollar un análisis con Notebooks.

Elemento	*Peso*
Prueba de evaluación 1 – Uso de NumPy y Pandas	35%
Prueba de evaluación 2 – Adquisición y uso de datos	35%
Prueba de evaluación 3 – Análisis de datos	30%

La tercera Prueba incluye una parte abierta en la que los estudiantes que lo deseen pueden ir más allá de lo visto en clase.

PROFESORADO

Miguel-Angel Sicilia es Catedrático de Lenguajes y Sistemas Informáticos en la Universidad de Alcalá y co-fundador y socio de Jaratech Social Technologies. Antes de incorporarse a la Universidad, desarrolló su trabajo como arquitecto software para el comercio electrónico, y participó en el diseño de soluciones de Inteligencia Artificial en iSOCO, empresa spinoff del IIA del CSIC. Miguel-Angel ha desarrollado su investigación en diferentes aplicaciones de semántica computacional y aprendizaje automático.