Máster en Business Analytics y Big Data

Edición 2014 / 2015

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

OBJETIVOS

El objetivo principal de este módulo es aprender a procesar datos y realizar análisis descriptivos básicos mediante el uso del lenguaje de programación R y las librerías de funciones asociadas.

Esta asignatura está orientada a que el alumno sea capaz de realizar tanto la preparación como el análisis de un conjunto de datos, así como adquirir una base sólida de los fundamentos del entorno R. Estos conocimientos permitirán que el alumno realice análisis más completos y aplique técnicas avanzadas de aprendizaje automático que serán cubiertas en futuros módulos del máster. Estás técnicas avanzadas generalmente están programadas en paquetes hechos por terceras personas, por lo que otro objetivo importante consiste en que el alumno sea capaz de encontrar, instalar y leer la documentación asociada a dichos paquetes.

Los objetivos concretos del módulo son los siguientes:

1. **Objetivo 1**: Conocer las bases del lenguaje de programación R y manejar con soltura el entorno RStudio.
2. **Objetivo 2**: Ser capaces de cargar un conjunto de datos en R, procesarlo y realizar análisis descritivos de los datos.
3. **Objetivo 3**: Ser capaces de visualizar los datos y presentar gráficamente los resultados del análisis.
4. **Objetivo 4**: Instalar nuevos paquetes y leer la documentación asociada a los mismos.

METODOLOGÍA

Las sesiones de este módulo son principalmente prácticas. Durante las sesiones se combinarán los aspectos teóricos resumidos en una presentación con demostraciones prácticas en el entorno Rstudio. El alumno pasará la mayor parte del tiempo con el portátil siguiendo las demostraciones prácticas, para a continuación realizar una serie de ejercicios propuestos en el lenguaje de programación R y el entorno Rstudio. Para familiarizarse con la sintáxis de R es importante que el alumno se equivoque y sea capaz de interpretar los mensajes de error que se generan.

Dada la complejidad del lenguaje R, es importante que el alumno escriba por si mismo todas las demostraciones prácticas que realice

el profesor. También se
anima al alumno a probar distintas variaciones y versiones de los comandos presentados en las prácticas.
Durante la presentación también se explicarán las principales características del lenguaje R, que lo hacen diferente de otros lenguajes. Para ello a menudo se comparará con otros lenguajes para cálculo científico, principalmente Python, que es uno de los más populares. Esto ayudará al alumno a evitar los principales errores que se cometen en R cuando se conocen varios lenguajes, así como elegir en un futuro el lenguaje más adecuado para el problema que quiere resolver.

PROGRAMA

El programa del módulo se estructurará en en 2 sesiones presenciales de 5 horas de duración cada una. A su vez, estás sesiones se dividirán en dos partes cada una, con un descanso aproximadamente en mitad de la sesión. Todos los materiales son comunes a las sesiones y se publicarán en el Moodle de la asignatura (enlaces, documentación, diapositivas, etc).

Sesión 1: Lenguaje R y entorno RStudio.
Actividades: En esta primera sección se aprenderá a manejar el entorno Rstudio. También se presentará el lenguaje R, con sus ventajas y desventajas con respecto a otros lenguajes. A continuación se mostrará los tipos de datos y estructuras del lenguaje: vectores, listas, matrices y data frames. Por último se estudiará la sintaxis básica del lenguaje R y las estructuras de control como bucles, sentencias condicionales, etc. Se realizarán numerosos ejercicios durante la clase para familiarizarse con el entorno y el lenguaje.

Sesión 2: Carga, filtrado y análisis de datos. Presentación de resultados.
Actividades: En esta sesión se realizará el proceso completo del análisis de unos datos. Primero se cargarán los datos en la memoria para realizar a continuación un análisis descriptivo de los mismos (media, varianza, etc). También se mostrarán estos datos de forma gráfica, lo que nos permitirá introducir los gráficos en R. En la segunda parte de esta sesión se procederá a transformar esos datos para adaptarlos a nuestras necesidades (filtrado, limpieza, etc.) y se aplicará alguna técnica de aprendizaje automático para extraer conocimiento de los mismos.

MATERIALES

Para realizar los ejercicios prácticos en clase se requiere tener instalada una versión reciente del software R, que se puede descargar desde la página web http://cran.es.rproject.org. El entorno R funciona en los principales sistemas operativos: Windows, Linux y OSX.

Aunque no es imprescindible, se recomienda instalar el IDE (Entorno Integrado de Desarrollo) RStudio. Este entorno proporciona un entorno similar a Matlab, con funcionalidades muy útiles en el desarrollo de aplicaciones en R como pueden ser autocompletado, ayuda integrada, visor de objetos en memoria, etc. Se puede descargar de forma gratuita de la web http://www.rstudio.com/products/rstudio.

Como material de referencia durante las clases, aunque hay muchos que se pueden descargar de forma gratuita en Internet, se recomienda el uso del libro "Introducción a R". El motivo es que se publica dentro del marco del proyecto de R y tiene versiones en múltiples idiomas, entre ellos el español y el inglés. La versión en español se puede descargar de la dirección:

http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf

La versión en inglés se puede descargar de:

http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Otra buena referencia en castellano es el libro "R para principiantes", que se puede encontrar en el enlace:

http://cran.r-project.org/doc/contrib/rdebuts_es.pdf

Una vez finalizadas las clases el alumno puede seguir profundizando en aquellos aspectos de R que más le interesen. En la página web del proyecto de R se recopilan manuales realizados tanto por el equipo oficial de desarrolladores

http://cran.r-project.org/manuals.html

como por la comunidad

http://cran.r-project.org/other-docs.html

De todos los manuales avanzados de R destacar "The R Inferno", que se puede descargar en www.burns-stat.com/pages/Tutor/R_inferno.pdf.

Por último, en cuanto a libros publicados en papel destacan "The R Book" de Michael J. Crawley y la colección "Use R!", publicada por la editorial

Springer.

EVALUACIÓN

Niveles de consecución de los objetivos

En la siguiente tabla se muestran los objetivos concretos a alcanzar en el módulo. Para cada uno de los objetivos definidos se especifican tres niveles de consecución de objetivos. La calificación final dependerá del nivel de consecución del alumno.

Objetivo específico	Nivel alto	Nivel medio	Nivel bajo
Entorno de trabajo	Comandos Help Objetos Gráficos Paquetes	Comandos Help Objetos	Comandos Help
Tipos de datos básicos	Vectores (numéricos) Matrices Data frames Listas Vectores (lógicos) Objetos	Vectores (numéricos) Matrices Data frames Listas	Vectores (numéricos) Matrices
Gráficos	Histograma Boxplot Scatter plot 2D Leyendas Anotaciones Gráficos múltiples	Histograma Boxplot Scatter plot 2D Leyendas	Histograma Boxplot Scatter plot
Estadística descriptiva	Media Max-min Varianza Correlación Outliers	Media Max-min Varianza	Media Max-min
Trabajo con datos	Carga Visualización Limpieza Transformación	Carga Visualización Limpieza	Carga Visualización

Modelo de evaluación

Para la evaluación de la asignatura se plantearán a lo largo de las clases una serie de ejercicios que el alumno deberá realizar y se dejará un tiempo para ello. Los ejercicios están diseñados para que sea posible acabarlos durante las propias clases, aunque existe la

posibilidad de que el alumno los termine fuera de la clase si ello fuera necesario. Adicionalmente, el alumno deberá realizar un trabajo final de módulo en el que deberá poner en práctica los conceptos fundamentales tratados durante las clases.
La siguiente tabla detalla los pesos de cada una de las actividades de evaluación:

Elemento	Peso
Resolución de los ejercicios propuestos en clase	70%
Trabajo final del módulo	30%

PROFESORADO

Alberto Torres Barrán es Ingeniero en Informática y Máster en Inteligencia Computacional por la Universidad Autónoma de Madrid. En la actualidad es profesor ayudante de prácticas e investigador también en la Universidad Autónoma de Madrid, donde realiza los estudios de doctorado en modelos lineales dispersos y métodos estadísticos de aprendizaje automático. Entre las clases de prácticas que imparte se encuentran “Programación I” y “Programación Orientada a Objetos”.

Alberto es desarrollador y usuario habitual de R, combinándolo con otros lenguajes de cálculo científico para realizar análisis de datos y algoritmos de aprendizaje. Para más información:

https://plus.google.com/+AlbertoTorresBarran

http://es.linkedin.com/in/albertotb