Máster en Business Analytics y Big Data

Edición 2015

Asignatura: APRENDIZAJE AUTOMÁTICO
Módulo: DATA SCIENCE/TÉCNICAS DE ANÁLISIS
Profesores:
Daniel Rodríguez Pérez, daniel.rodriguez.perez@gmail.com

OBJETIVOS

El objetivo general del módulo es el de proporcionar una introducción al **aprendizaje automático** como herramienta analítica para construir **modelos** que permiten aprender a partir de los datos. El módulo parte de la base del conocimiento de entornos de data science como R o IPython/SciPy a nivel de uso de herramientas de análisis estadístico y técnicas de limpieza y transformación de datos. En la práctica, en muchos casos el uso de técnicas de aprendizaje automático tiene lugar como una de las actividades del ciclo de **minería de datos**, que se apoya habitualmente en data warehouses, tal y como se introduce en el módulo de inteligencia de negocio y minería de datos.

Es importante resaltar que muchos entornos en la nube incluyen ya bibliotecas de aprendizaje automático preparadas para su ejecución e incluso para la distribución paralela y escalable de las mismas en *clusters* de máquinas cuando es necesario por el volumen o dimensionalidad de los datos.

El énfasis de la asignatura está en adquirir las habilidades y competencias prácticas para saber **seleccionar** algoritmos de aprendizaje automático para una situación dada de negocio y **evaluar** los modelos resultantes para decidir si son aplicables a la práctica. No se pretende entrar en los detalles internos de los algoritmos ni en su implementación.

Se sigue un enfoque práctico, utilizando bibliotecas abiertas de aprendizaje automático y entornos de *data science* para utilizar y evaluar los algoritmos de manera interactiva.

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

1. Entender el papel del aprendizaje automático dentro de los procesos de minería de datos y cómo se utilizan para descubrir patrones en bases de datos y construir modelos predictivos.
2. Entender la diferencia entre aprendizaje supervisado y no supervisado y saber a qué situaciones se aplican.
3. Entender el proceso de entrenamiento y de evaluación de los algoritmos y el papel de las técnicas de validación cruzada.
4. Ser capaces de seleccionar familias de algoritmos y algoritmos

concretos de acuerdo a las características del problema.
5. Saber utilizar bibliotecas de algoritmos de aprendizaje automático en entornos de data science.
6. Saber medir la capacidad predictiva de un modelo e identificar los modelos que han de ser reentrenados o sustituidos por otros.

METODOLOGÍA

Las clases se desarrollarán de manera interactiva, utilizando ejemplos para practicar directamente en las clases utilizando datasets de ejemplo.

El elemento central de la metodología docente es el trabajo práctico con un entorno de análisis de datos, para entender el trabajo del *data scientist* y adquirir habilidades básicas para seleccionar y aplicar algoritmos de aprendizaje automático. Se proporcionará una máquina virtual VirtualBox con el software instalado. Se utilizará fundamentalmente el entorno IPython mediante Notebooks, y las bibliotecas scikit-learn salvo para los casos en los que no se encuentren algoritmos en ese framework.

Los estudiantes realizarán ejercicios propuestos fuera de clase, que servirán como preparación para las prácticas de evaluación.

PROGRAMA

Sesión 1: Introducción a aprendizaje automático y minería de datos
Actividades: Práctica básica con scikit-learn, reglas de asociación, modelos de regresión lineal, overfitting y validación cruza
Materiales: IPython Notebooks.

Sesión 2: Modelos de regresión y clasificación.
Actividades: Práctica con modelos de regresión, modelos de clasificación, selección de variables y medida de la bondad del ajuste.
Materiales: IPython Notebooks.

Sesión 3: Modelos de clustering y aprendizaje profundo.
Actividades: Práctica con modelos de clustering, PCA y Aplicaciones. Introducción a Aprendizaje profundo.
Materiales: IPython Notebooks.

MATERIALES

Materiales obligatorios

Manuales de referencia de scikit-learn disponibles en:
http://scikit-learn.org/stable/documentation.html

Se proporcionan otros recursos on-line en los diferentes temas.

Referencias recomendadas

Una introducción muy básica al aprendizaje automático, muy recomendable para quien no tenga ningún conocimiento previo es el siguiente libro

- Bootstrapping Machine Learning (Louis Dorard), disponible en http://www.louisdorard.com/machine-learning-book/ Cuenta con edición electrónica y recursos adicionales para profundizar si se desea.

Introducción al aprendizaje automático utilizando con las librerías scikit-learn

- Mastering Machine Learning With scikit-learn (Gavin Hackeling), Packt Publishing, Birmingham, 2014

EVALUACIÓN

Niveles de consecución de los objetivos

Objetivo específico	*Nivel alto*	*Nivel medio*	*Nivel bajo*
O1 – aprendizaje automático	Identificar que método que se ha de utilizar para diferentes problemas.	Identificar el tipo de aprendizaje que se ha de utilizar en cada tipo de problema.	Conocer la diferencia entre aprendizaje supervisado y no supervisado.
O2 – modelos de regresión	Medida de la capacidad predictiva del modelo.	Selección de las variables y creación de modelos de regresión.	Conocimiento de los modelos de regresión y sus aplicaciones.
O3 – modelos de clustering	Utilización de los modelos de clustering para segmentar.	Realización de análisis de clustering utilizando las librerías de scikit-learn.	Conocimiento de los modelos de clustering y sus aplicaciones.
O4 –	Aplicación de los	Utilización de scikit-	Conocimiento de

modelos de clasificación	modelos de clasificación	learn para la creación de modelos de clasificación.	los modelos de clasificación y sus aplicaciones.
O5 – componentes principales	Interpretación de los resultados de análisis de PCA	Utilización de las herramientas para la realización de un PCA.	Compresión del procedimiento de análisis de componentes principales (PCA).

Modelo de evaluación

La siguiente tabla detalla los pesos de cada una de las actividades de evaluación.

Elemento	*Peso*
Cuestionarios	40%
Prácticas	60%

PROFESORADO

Daniel Rodríguez tiene más de 10 años de experiencia en el desarrollo de soluciones analíticas para bancos, seguros y retail. Actualmente es VP Análisis de Datos en Analytika, desarrollando anteriormente su carrera en everis y PricewaterhouseCoopers y participado en proyectos en Inglaterra, España, América Latina y Sudáfrica. Daniel Rodríguez es licenciado en Física por la Universidad de Santiago de Compostela y doctor por la Universidad Politécnica de Madrid. Además es autor de numerosas publicaciones científicas y paquetes para R publicados en el CRAN.