Data mining

Jordi Gironès Roig

PID_00203551

Material docente de la UOC

UOC
Universitat Oberta de Catalunya

Jordi Gironès Roig

El encargo y la creación de este material docente han sido coordinados por los profesores: Jordi Conesa Caralt, David Masip Rodó (2013)

Primera edición: octubre 2013

Av. Tibidabo, 39-43, 08035 Barcelona
Diseño: Manel Andreu
Realización editorial: Eureca Media, SL
Depósito legal: B-14.754-2013

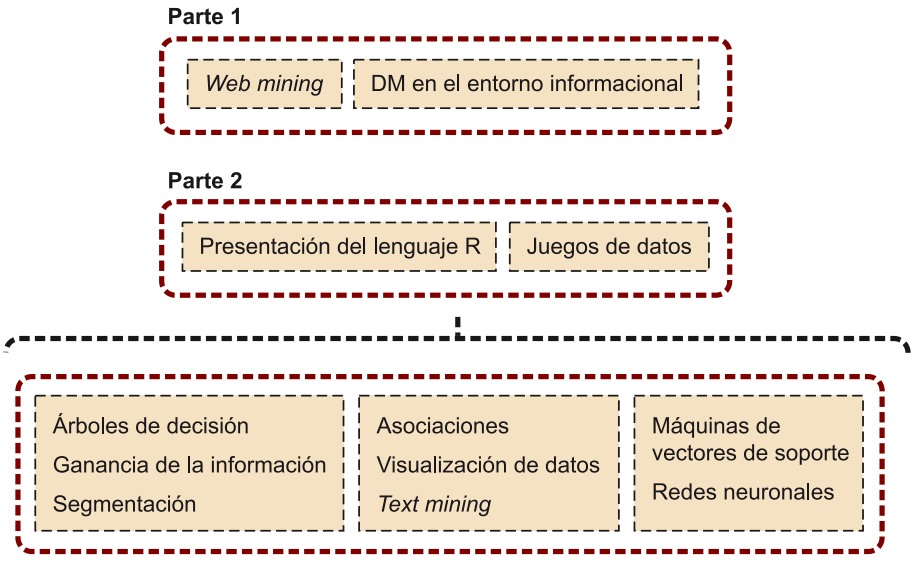
Introducción

Las técnicas de minería de datos nos permiten extraer patrones y relaciones, que nos llevan en definitiva a la obtención de un nuevo conocimiento latente en los datos.

Para conseguir este objetivo, la minería de datos se apoya en disciplinas como las matemáticas, la estadística y la inteligencia artificial, pero también en los lenguajes de programación. Tanto en ámbitos académicos como en ámbitos corporativos, R se ha convertido *de facto* en un entorno idóneo para la implementación de algoritmos de minería de datos.

Mediante este material didáctico el estudiante tendrá la oportunidad de practicar con casos de estudio técnicas concretas sobre datos concretos.

Para situar de una forma visual al estudiante en la estructuración del material didáctico, introducimos en la figura 1 un esquema de contenidos.



En la primera parte estudiaremos las particularidades de la minería web por tener unas especificidades que la diferencian del resto de la minería de datos y por la importancia estratégica que la web tiene para cualquier organización.

Asimismo, enmarcaremos la minería de datos dentro del entorno informacional y estudiaremos especialmente los procesos de extracción de conocimiento por ser intrínsecos a la minería de datos.

El cierre del ciclo analítico, las estrategias de *scoring* de modelos y los servicios de minería de datos dentro de la infraestructura tecnológica nos ayudarán a entender mejor los distintos escenarios y estrategias de puesta en producción de modelos.

Todo ello constituirá una sólida base teórica que permitirá al estudiante aprovechar mejor los conocimientos técnicos que mediante el lenguaje de programación R adquirirá en los siguientes apartados del material didáctico.

En la segunda parte pondremos en práctica muchos de los conceptos estudiados hasta el momento. Una introducción al entorno de programación R, el proceso de instalación y el vocabulario e instrucciones básicas constituirán la primera aproximación.

El estudiante ganará agilidad y conocimiento en R al mismo tiempo que irá practicando y experimentando con cada uno de los casos prácticos que se presentan en el material didáctico, de modo que de una manera natural adquirirá habilidades tanto en R como en minería de datos.

Se estudiarán ocho algoritmos representativos de las técnicas supervisadas y no supervisadas. Exploraremos sus posibilidades y se facilitarán *scripts* para que el estudiante pueda familiarizarse mejor con el entorno de programación R.

Las implementaciones R de algoritmos de minería de datos que estudiaremos tienen en común una gran flexibilidad en su utilización, parámetros que permiten cambiar de fórmulas matemáticas o simplemente ajustarlas hasta conseguir mejores aproximaciones.

En definitiva se abre para el estudiante el mejor de los escenarios posibles para aprender. La posibilidad de experimentar, de cambiar, de probar, de equivocarse y a veces de acertar, pero sobre todo de llegar a sus propias conclusiones siempre desde una sólida base tanto teórica como práctica.

Objetivos

En la primera parte del material didáctico se adquirirán las siguientes habilidades:

1. Conocer las especificidades del *web mining* dentro de la disciplina *data mining*.

2. Saber encajar la minería de datos dentro del entorno informacional. De este modo el estudiante se familiarizará con conceptos como procesos del entorno informacional, procesos de extracción de conocimiento y cierre del ciclo analítico.

3. Entender los procesos del *scoring* de modelos, así como los distintos escenarios para la puesta en producción de los mismos.

En la segunda parte se utilizará el lenguaje de programación R para que el estudiante adquiera habilidades de desarrollo de técnicas de minería de datos soportadas por software. De modo que las competencias que se desarrollarán son:

1. Adquirir capacidades básicas en el entorno de programación R: proceso de instalación, proceso de ampliación por paquetes, acceso a la documentación en línea, dominio de la consola de ejecución de comandos, acceso a bases de datos PostgreSQL y manejo de *scripts*.

2. Mediante el lenguaje R, ser capaz de aplicar las siguientes técnicas de minería de datos sobre juegos de datos apropiados: árbol de decisión, ganancia de la información, algoritmo Kmeans, segmentación jerárquica, detección de valores *outliers*, obtención de reglas de asociaciones, máquinas de vectores de soporte, redes neuronales, visualización de datos y técnicas propias de la minería de textos.

Contenidos