Minería de datos y big data analysis: fundamentos, tecnologías y aplicaciones

TEMARIO

Tema 1
Introducción a la minería de datos:

- Principales escenarios de análisis: clasificación supervisada, clustering, sistemas de recomendación, clasificación semi-supervisada, reglas de asociación, clasificación multi-label y multi-dimensional, "weak-supervision" (label proportions, partial labels, multiple-instance learning, partial labels, crowd learning, etc.).
- Ilustración de las principales y actuales aplicaciones para cada uno de los escenarios de clasificación anteriores: marketing, bioinformática, industry 4.0, imágenes

Tema 2
Minería de datos: desde la teoría a la práctica. Ilustración de los anteriores escenarios de análisis mediante el software WEKA.

Tema 3
Visualización de datos. Práctica con el software R.

Tema 4
Preprocesado de datos para su posterior análisis. Principales técnicas y filtros.

Tema 5
Introducción a la selección de variables. Tipos de técnicas de selección de variables.

Tema 6
Estimación del porcentaje de bien clasificador y tests estadísticos para la comparación de clasificadores: evaluación y credibilidad de los modelos aprendidos

Tema 7
Estudio de distintos casos de uso y recursos:

- Sistemas de recomendación – "Recommender systems": músicas, películas
- Bioinformática: selección de genes diferencialmente expresados para el diagnóstico y pronóstico de enfermedades
- Informe del "World Economic Forum" sobre las posibilidades económicas que brinda el análisis masivo de los flujos de datos modernos
- Problemas de la plataforma kaggle.com
- Portal de referencia sobre el uso de la minería de datos en el mundo empresarial, industrial: www.kdnuggets.com
- Visitar la siguiente página web para consultar el abanico de las

aplicaciones y casos de uso que manejará el profesor durante el curso y estudiará junto con los alumnos: http://www.sc.ehu.es/ccwbayes/members/inaki/DM-applications.htm
- Aplicaciones es: marketing y publicidad dirigida, transporte y logística, sentiment-analysis…

Tema 8
Tutorial con el software R (paquete "caret"). Creación de todo el "pipeline"-flujo de análisis

Tema 9
Tutorial con el software R (paquete "h2o"): mining big data

BIBLIOGRAFÍA

BIBLIOGRAFÍA BÁSICA:

- M. Kuhn, K. Johnson (2013). Applied Predictive Modeling. Springer.
- I.H. Witten, E. Frank (2011). Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, 3rd edition.
- B. Sierra (2006). Aprendizaje Automático: Conceptos Básicos y Avanzados. Pearson – Prentice Hall.

BIBLIOGRAFÍA DE PROFUNDIZACIÓN:

- J. Albert, M. Rizzo (2012). R by Example. Springer.
- G. Williams (2011). Data Mining with Rattle and R. Springer.
- F. Hahne, W. Huber, R. Gentleman, S. Falcon (2008). Bioconductor Case Studies. Springer.
- D. Sarkar (2008). Lattice; Multivariate Data Visualization with R. Springer.
- S. Aiello, E. Eckstrand, A. Fu et al. (2016). Machine Learning with R and h2o. H2O.ai Inc.
- I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, J.A. Lozano (2010). "Machine learning: an indispensable tool in bioinformatics". Methods in Molecular Biology. R. Matthiesen (ed.). Humana Press.
- Y. Saeys, I. Inza, P. Larrañaga (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517

DIRECCIONES DE INTERNET:

- Kdnuggets: data mining, web mining, text mining, and knowledge discovery: http://www.kdnuggets.com
- A compilation of data mining applications: http://www.sc.ehu.es/ccwbayes/members/inaki/DM-applications.htm
- National Center for Biotechnology Information: http://www.ncbi.nlm.gov/
- Competiciones de minería de datos: http://www.kaggle.com
- Fast and scalable machine learning: http://www.h2o.ai