

# Minería de datos

Profesores: Itziar Irigoien, Javier Muguerza, Ibai Gurrutxaga, José Ignacio Martín, Olatz Arbelaitz, Txus Perez

## TEMARIO

### Tema 1

Introducción a la minería de datos

- Principales escenarios de análisis: clasificación supervisada, clustering, clasificación semi-supervisada, reglas de asociación, otros escenarios de clasificación (multilabel y multidimensional, label proportions, multiple-instance learning, partial labels, etc.).
- Introducción a las principales y más conocidas aplicaciones para cada uno de los escenarios de clasificación anteriores: especialmente en el ámbito empresarial y del marketing, así como científicos, bioinformáticos, etc.
- El concepto de "Big data". Ejemplos y aplicaciones.

### Tema 2

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

### Tema 3

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

### Tema 4

Introducción de la selección de variables. Selección de las variables relevantes para el análisis.  
Tipos de técnicas.

### Tema 5

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

### Tema 6

Discusión y estudio de casos de estudio concretos:

- 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
- Ver 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 en: marketing y publicidad dirigida, transporte y logística, sentiment-analysis...

## BIBLIOGRAFÍA

### BIBLIOGRAFÍA BÁSICA

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.

P. Baldi, S. Brunak (2001). Bioinformatics. The Machine Learning Approach. MIT

H.C. Causton, J. Quackenbush, A. Brazma (2004). A Beginner's Guide: Microarrays Gene Expression Data Analysis. Blackwell Publishing.

## BIBLIOGRAFÍA DE PROFUNDIZACIÓN

C.M. Bishop (2006). Pattern Recognition and Machine Learning. Springer.

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 INTERÉS EN 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>

A quick introduction to biology for computer scientists:  
[http://www.ebi.ac.uk/microarray/biology\\_intro.html](http://www.ebi.ac.uk/microarray/biology_intro.html)

National Center for Biotechnology Information:  
<http://www.ncbi.nlm.gov/>

Intelligent Systems Group (UPV-EHU):  
<http://www.sc.ehu.es/isg>