# Laboratorio de programación en R

**Minería de datos: Aspectos Avanzados**

**6 de Febrero, 2019**

Para esta sesión de laboratorio, se necesitará descargar los ficheros ordinal.pdf y esl-j48ordinal.R desde PRADO, además de los conjuntos de datos esl.arff, era.arff, lev.arff y swd.arff, para hacer pruebas.
Para esta actividad, tendréis que subir dos ficheros R generalordinal.R y monoxgboost.R a la actividad correspondiente de PRADO, antes del ***17 de Febrero de 2019 (23:55)***. Hay que asegurarse de que el código entregado esté suficientemente comentado. También, proporcione un análisis breve de los resultados obtenidos usando comentarios en los ficheros R.

## Clasificación ordinal

En este ejercicio, se implementará la técnica de múltiples modelos para clasificación ordinal utilizando el algoritmo J48 del paquete RWeka de R. Para ello, disponemos de una plantilla ordinal.pdf (ejemploOrdinal.pdf) que hace el proceso de múltiples modelos considerando el clásico data set iris y el clasificador J48 de RWeka. En este caso, se ha supuesto que las clases de iris tienen una disposición ordinal. Como iris tiene tres clases, se generan dos conjuntos de datos desde el original. Después, se contruyen dos árboles asociados a cada conjunto. Finalmente, éstos se utilizan para clasificar una instancia de test seleccionada y obtener las probabilidades.

### Modelos múltiples para el data set esl.arff

A partir del fichero esl-j48ordinal.R se puede adaptar el proceso de modelos múltiples de clasificación ordinal al conjunto de datos esl con el algoritmo J48 (C4.5). Para ello, se considerará una única partición entrenamiento test aleatoria con 100 ejemplos en el conjunto de test. Este conjunto tiene carácter ordinal y dispone de 9 clases [1, 9]. Una vez generados los modelos desde el conjunto de entrenamiento, habrá que clasificar los 100 ejemplos de test usando la cascada de probabilidades que define el modelo múltiple. Pero el objetivo es generalizar este modelo.

### Generalización del proceso de modelos múltiples

El objetivo de esta actividad es la de construir este proceso de forma genérica para cualquier data set ordinal y un clasificador diferente al J48. Se implementará una función que recibirá un data set como parámetro y será capaz de formar los data sets binarios derivados a partir del número de clases del data set original. Así mismo, lanzará un clasificador diferente al J48 (a elección por el estudiante) para cada uno y devolverá una colección de modelos construidos.
Una segunda función recibirá dicha colección y será capaz de predecir un ejemplo (o un conjunto de ejemplos) a partir de los mismos usando la cascada de probabilidades que define el modelo múltiple. El resultado se generará en un fichero llamado generalordinal.R.

## Clasificación monotónica

En esta segunda parte, se implementará un modelo similar de clasificación múltiple para problemas con restricciones de monotonía. Se utilizará el modelo OVA básico descrito en las diapositivas de teoría (diapositiva 107) con el algoritmo xgboost de R, usando el parámetro `monotone_constraints=1`. Este algoritmo obtiene monotonía global en problemas binarios, por lo que es necesario hacer una descomposición que sea monótona-consistente. El resultado se generará en un fichero llamado monoxgboost.R.