uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

GUÍA PARA LAS PRÁCTICAS

El objetivo de esta guía es describir las prácticas disponibles y los tutoriales que se pueden usar como ayuda para realizarlas.

Aprendizaje Automático con R :

En esta sección se hace una breve introducción a como hacer aprendizaje automático directamente con librerías de R. Se contrapone esto al aprendizaje automático que se hará posteriormente con la librería MLR. Más adelante se verá que hacer aprendizaje automático con MLR es más sencillo porque uniformiza y automatiza muchas tareas de aprendizaje automático. Pero de momento, se va a hacer una breve introducción a cómo hacerlo directamente con R, porque es conveniente conocerlo, y para comprender lo que aporta MLR.

Se comienza con un breve repaso de R, centrándonos en los tipos de datos que se pueden manejar en R, especialmente los dataframes. Se explica que los dataframes, además de poderse ver como matrices, son realmente listas de vectores, y por tanto se pueden usar funciones de alto nivel como lapply para analizar los dataframes por columna, o para transformalos (por ejemplo, la normalización de cada atributo (o columna). Lapply puede servir por tanto para, por ejemplo, preprocesar las tablas de datos, aunque más adelante se verá que MLR automatiza la mayor parte de las tareas de preproceso. Además, se ofrece un ejemplo de como entrenar modelos basados en árboles y en vecino más cercano con las librerías de R adecuadas. Se puede observar como cada librería se utiliza de una manera ligeramente diferente, y esto es justamente lo que uniformiza MLR, como se verá más adelante. Se estudia como entrenar modelos y evaluarlos con un conjunto de test.

Tutoriales de MLR :

A continuación se ofrecen una serie de tutoriales bastante completos para aprender a utilizar MLR. Los propios tutoriales proponen ejercicios que se pueden resolver (y más adelante se dan las respuestas). Para usar los tutoriales es necesario cargar los ficheros adecuados como ficheros markdown (.Rmd) en Rstudio e ir ejecutando celda a celda. Se ofrecen tutoriales que cubren las varias tareas, tales como entrenar modelos, evaluarlos tanto con Holdout como con

crossvalidation y el ajuste de hiper-parámetros. Se dan tutoriales donde se ajustan los hiper-parámetos para métodos concretos (Gradient Boosting, Random Forests, SVMs).

MLR para muestras desbalanceadas :

Un tipo de problemas muy frecuente en clasificación son los conjuntos de datos desbalanceados (aquellos en los que una clase tiene muchos menos datos que otra). Se ofrece un tutorial para tratar con este tipo de muestras con MLR, cubriendo las dos técnicas básicas : rebalanceo mediante SMOTE y thresholding.

Tutorial de Spark con R :

En este tutorial se explica cómo usar Spark desde R. Una de las maneras más sencillas es la librería Sparklyr. Dado que está basado en otra librería llamada dplyr, también se explica esta última.

Prácticas :

Se ofrecen tres prácticas para que los alumnos practiquen los conceptos anteriores :

- Práctica 1 : conceptos básicos de aprendizaje automático en un problema de regresión, incluyendo el ajuste de hiper-parámetros.
- Práctica 2 : muestras desbalanceadas (y sus datos asociados « mammography ».
- Práctica 3 : sparklyr

Otros recursos :

Para terminar, se ofrecen dos referencias a libros que tratan los temas vistos (**PR-B-001** y **PR-B-002**) y varios enlaces a tutoriales y cursos sobre los temas de las prácticas y de las herramientas usadas (PR-E-00X).