uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE

APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS

GRADO EN ESTADÍSTICA Y EMPRESA

Ricardo Aler

PRÁCTICA DE sparklyr (1.0 puntos)

En este caso vamos a utilizar el conjunto de datos “BostonHousing corrected”, que tenéis en un fichero en vuestro directorio. De manera parecida a la primera práctica, BostonHousing contiene datos acerca de la mediana de precio (medv) de unas 500 propiedades en EEUU y 13 variables explicativas. Tenéis que hacer lo siguiente usando dplyr (usando pipes %>% allí donde se pueda) y sparklyr.

Parte I (0.7 puntos)

1) Instalar Sparklyr con:

```
install.packages("sparklyr")
library(sparklyr)
spark_install(version = "2.1.0")
```

2) Arrancar el cluster Spark con

```
sc <- spark_connect(master = "local”)
```

3) Vamos a simular la lectura de datos en el cluster Spark. Normalmente lo leeriamos de un fichero (con spark_read_csv), pero en este caso lo crearemos a partir de un dataframe local con copy_to.

```
bh_spark  <- copy_to(sc, BostonHousing2, repartition=2)
```

4) En estos datos hay dos variables de respuesta: medv y cmedv. La última es como la primera, pero con ciertos valores corregidos. Sólo podemos tener una variable de respuesta, con lo que se pide quitar la columna mdev.
5) Dividir el Spark dataframe que contiene los datos (bh_spark) en tres particiones: train (52%), validation (16%), test (33%).
6) Vamos a usar algún método de regresión disponible en Sparklyr. En concreto, tenéis que elegir uno de Gradient Boosting o Random Forest y ajustar un único hiper-parámetro (número de árboles). Probad 3 valores, a elegir entre 20 y 200. La medida de error que vamos a usar en toda la práctica es RMSE.
7) Una vez determinados el mejor hiper-parámetro, construid el modelo final y evaluadlo en el conjunto de test.

Parte II (0.3 puntos)

8) La variable crim indica el número de crímenes per cápita. Responded a la siguiente pregunta: el modelo que habéis construido, acierta más con los pisos con mayor número de crímenes o con los pisos con menor número de crímenes. Mayor y menor es relativo a la mediana de número de crímenes. Se pide responder a esta pregunta con un código lo más breve posible (si es posible, una única tubería dplyr/sparklyr (o sea, la pipe %>%)).