Universidad del Rosario, Facultad de Economía
Big Data: Machine Learning
Taller 2

February 22, 2017

1. Para predecir presencia de células cancerosas con base en la características obtenidas de biopsias realizadas a diferentes mujeres, trabaje con la base `wisc_bc_data.csv`.

   (a) Lea los datos por medio de un objeto llamado `wbcd`. Remueva de este objeto la información de identificación de las mujeres. Redefina su base de datos de acuerdo con este cambio.

   (b) Use las funciones `table()`, `summary()` y otras que considere pertinentes para entender la estructura de los datos.

   (c) Reescale las variables de la base por medio de una estandarización (Z-score) de los datos. Bautice a su nueva base de datos, tras la estandarización, `wbcd_z`. Verifique, por medio de estadística descriptiva, que los datos están en la misma escala y son comparables.

   (d) Divida la base `wbcd_z` en dos partes: una con el 70% de las observaciones para entrenar el modelo, el restante 30% para probar el modelo. Llame a estas dos bases `wbcd_z_train` y `wbcd_z_test`, respectivamente.

   (e) Usando $k = 21$ vecinos, haga clasificación usando vecino más cercano (kNN). Presente en una tabla el resumen de las predicciones hechas para la base de prueba.

   (f) ¿Qué tan precisos son los resultados? Considera usted que en este caso es preferible minimizar el número de falsos positivos o de falsos negativos? ¿Por qué?

   (g) Compare sus resultados con los que se obtienen cuando en lugar de estandarizar usando Z-scores, se hace una normalización min-max.

2. Con los mismos datos e instrucciones del ejercicio anterior, pruebe cómo cambian los resultados si varía el número de vecinos k. Haga normalización min-max para rescalar los datos Pruebe con los siguientes valores y en cada caso indique el porcentaje de falso negativos, falsos positivos y el porcentaje total clasificado incorrectamente:

   (a) $k = 1$

   (b) $k = 7$

   (c) $k = 13$

   (d) $k = 17$

   (e) $k = 21$

   (f) $k = 29$

   A partir de los resultados, ¿qué diría usted sobre la robustez de las predicciones?

3. Desarrolle un bloque de código en `R` que permita construir una base de datos de profesores a partir del siguiente link: http://econ.as.nyu.edu/page/people. Construya un *data frame* que contenga las siguientes variables: Nombre del profesor, cargo, teléfono, correo electrónico, intereses de investigación y el link de la página personal.

Fecha de entrega: miércoles 1 de marzo. Enviar a jdavidmartinezg@gmail.com