Trabajo Práctico 1: Uso de R en Ciencia de Datos

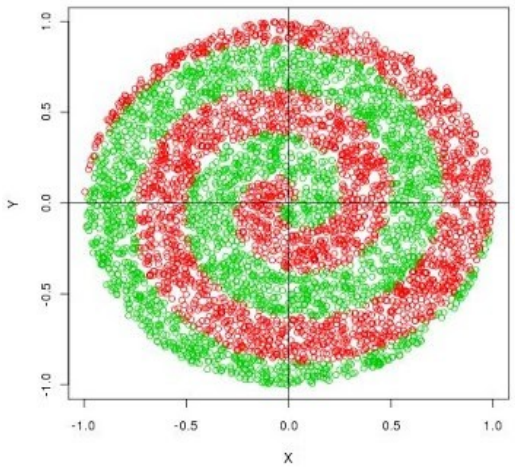
Crear un notebook en Colab: https://colab.to/r

(también hay formas de correr R directamente en Jupiter, instalando el kernel IRkernel en https://irkernel.github.io/installation/)

1. Prepare código en R que genere conjuntos de datos (de longitud dada n) de acuerdo a las siguientes descripciones:

a) Diagonal: Los datos tienen d inputs, todos valores reales, correspondientes a la posición del punto en un espacio d-dimensional. El output es un factor (binario), y corresponde a la clase a la que pertenece el ejemplo. La clase 1 corresponde a puntos generados al azar, provenientes de una distribución normal, con centro en el (1, 1, 1, ... , 1) y matriz de covarianza diagonal (todas las variables son independiantes entre si), con desviación estándar igual a C * SQRT(d). La clase 0 tiene la misma distribución, pero centrada en el (-1, -1, -1, ... , -1). Se puede encontrar información sobre Gaussianas multidimensionales y el caso especial de una matriz diagonal en http://cs229.stanford.edu/section/gaussians.pdf (secciones 1 y 3). Los parámetros que se deben ingresar son d y n (enteros) y C (real). De los n puntos generados, n/2 deben pertenecer a cada clase. Argumentos de la función: n-d-C. Salida: un dataframe con d+1 columnas y n filas.

b) Espirales anidadas: Los datos tienen 2 inputs, x e y, que corresponden a puntos generados al azar con una distribución UNIFORME (en dicho sistema de referencia x-y) dentro de un circulo de radio 1. El output es binario, correspondiendo la clase 0 a los puntos que se encuentran entre las curvas ro = theta/4pi y ro = (theta + pi)/4pi (en polares) y la clase 1 al resto. De los n puntos generados, n/2 deben pertenecer a cada clase.



2. Use R para generar gráficas de los datos en a y b.

3. Genere un conjunto de entrenamiento chico (300 puntos) y uno de test grande (10000 puntos) para cada problema. Ajuste los clasificadores de árboles y de k-vecinos. Mida el error de test. Ahora repita el procedimiento usando solo el conjunto de entramiento y una estimación en 5-fold cross-validation. Compare los resultados entre el metodo "directo" y la estimación en K-folds.

Entrega el notebook (si corre, mucho mejor ;))