Práctica: Curvas ROC y k-vecinos más cercanos (kNN)

1 Conceptual

1. La tabla de abajo muestra los datos de entrenamiento de un conjunto de 6 datos, consistentes en tres predictores y una respuesta categórica de respuesta.

Obs	X_1	X_2	X_3	Y
1	0	3	0	Rojo
2	2	0	0	Rojo
3	0	1	3	Rojo
4	0	1	2	Verde
5	-1	0	1	Verde
6	1	1	1	Rojo

Queremos predecir una clasificación para el punto $X_1 = X_2 = X_3 = 0$ usando KNN.

   (a) Calcule las distancias de todos los puntos de entrenamiento con respecto al punto de prueba $X_1 = X_2 = X_3 = 0$.

   (b) Evalúe KNN con K=1.

   (c) Evalúe KNN con K=3.

2 Aplicado: Regresión Logística y Curvas ROC para selección de umbral

1. Tomemos el data set **abalone.txt**, consideraremos para el mismo el modelo de clasificación logística basado en la longitud, el peso total y la cantidad de anillos. En vez de considerar el umbral de clasificación en 0.5, probaremos distintos valores en una grilla entre 0 y 1. Una vez entrenado el clasificador logístico usando un 70% de los datos, consideraremos diversos valores de corte para el umbral para hacer el análisis sobre el 30% restante de los datos de evaluación. Para cada uno de esos umbrales θ, calcularemos

$$TPR(\theta) = \frac{TP(\theta)}{TP(\theta) + FN(\theta)},$$

$$FPR(\theta) = \frac{FP(\theta)}{TN(\theta) + FP(\theta)}.$$

   Graficamos entonces una curva de FPR (eje x) vs TPR (eje y), llamada la curva ROC. El punto (0,1) del espacio ROC corresponde a un clasificador óptimo, cualquier punto en la diagonal sería un clasificador perfectamente aleatorio, mientras que en (1,0) tendremos un clasificador perfectamente malo (que también es óptimo si uno invierte sus interpretaciones). Para cada umbral, calcular la distancia del punto (fpr, tpr) con respecto al vértice (0,1) o (1,0) más cercano. Graficar una curva de umbral vs distancia y determinar así el mejor umbral. Arme la matriz de confusión correspondiente a dicho umbral con el 30% de los datos separados.

   Nota: se supuso aquí previamente elegido el modelo, es posible hacer también selección de modelo utilizando métricas sobre la curva ROC que permitan reducir a un valor comparativo la información de la misma.

3 Aplicado: kNN

1. En el archivo **productos.txt** tenemos un registro de productos que fueron exitosos y otros que fracasaron, junto con información de su precio y de su presupuesto en marketing.

   (a) Levantar los datos del archivo y graficarlos, en negro los productos "exitosos" y en rojo los "fracasados".

   (b) Para determinar un valor "óptimo" de k se propone:

i. Estandarizar el total de los datos con alguno de los dos esquemas propuestos.

ii. Particionar el conjunto en un **training** y un **testing set**.

iii. Para un valor de K y un mecanismo de estandarización clasificar los elementos del **testing set** en función del **training set**, aplicar como función de costo la cantidad de datos incorrectamente clasificados.

iv. Repetir con diversos valores de K y elegir como modelo aquel que minimice la función de costo.

(c) Replique el ejercicio anterior con un mecanismo de K-Fold para elegir el k de vecinos más cercanos.

(d) Clasificar un producto con precio \$70 y un presupuesto de marketing de \$100000 utilizando el modelo óptimo encontrado antes.

2. (Simulación)

   (a) Genere 1000 datos z_i provenientes de una uniforme $U(-1, 1)$ (**runif**).

   (b) Si z_i es menor que 0, entonces asignelo al grupo 0 ($y_i = 0$). Caso contrario, asignarlo al grupo 1 ($y_i = 1$).

   (c) Defina $x_i = 0.5 + 4 * z_i + 0.2 * z_i^2 + \epsilon$ donde ϵ_i sigue una distribución normal estándar. Será el valor "obervado".

   (d) Separe entre 200 y 300 de esos datos para "evaluación final".

   (e) Tome 700-800 datos restantes (x_i, y_i), con un esquema de K-fold elija el mejor modelo de kNN.

   (f) Clasifique los primeros datos separados y arme una matriz de confusión.

3. (a) Similar al ejercicio anterior, genere z_i provenientes de una uniforme $U(-1, 1)$ y w_i de la misma distribución.

   (b) Considere $y_i = 1$ si ambos z_i y w_i tienen el mismo signo, y 0 si no.

   (c) Se consideran "observados" y_i, z_i y w_i.

   (d) Grafique los datos (z_i, w_i) en rojo si son del grupo cero y en azul si no.

   (e) Repita el esquema del ejercicio anterior, separando 200-300 datos para evaluación final y definiendo el modelo con los restantes. Arme la matriz de confusión resultante.

4. Considere el archivo **abalone.txt**. Separe enre un 20%-30% de los datos que serán para evaluación final de la calidad de predicción repita el mismo procedimiento del ejercicio anterior con los 70% datos restantes para determinar un modelo de clasificación kNN para decidir si un especimen es adulto o infante en base a la longitud, el peso total y la cantidad de anillos (puede ser por K-fold o training/test). Una vez elegido el mejor valor de k, obtenga una matriz de confusión sobre los 30% que se separaron inicialmente para validar la calidad de clasificación.

5. Considere el archivo **iris.data**, que tiene información sobre la longitud del sépalo, su ancho, la longitud del pétalo y su ancho, todo en centímetros. Finalmente, la última columna tiene información sobre la clase a la que pertenece. Repita el esquema del ejercicio anterior para determinar un valor óptimo de k y arme la matriz de confusión sobre los datos separados inicialmente.