

Análisis Exploratorio de Datos, 2009-2010

Ejercicios de Programación en R

Programas

1. Realizar un programa en un fichero denominado *miprograma.r* que:
 - (a) pida por consola el valor de dos vectores x e y mediante la función *scan()*.
 - (b) sume los dos vectores.
 - (c) produzca mediante la función *plot* un gráfico de línea del vector x contra el vector y^2 .

Ejecutar posteriormente el programa mediante la función *source*.

Funciones

1. Construir una función para calcular la moda de una tabla de datos.
2. Construir una función denominada *posicion* para calcular la media, mediana y moda de un conjunto de datos.
3. Construir una función para tipificar las columnas o las filas de una matriz, incluyendo un argumento que determine si se desea lo primero o lo segundo.
4. Construir una función para, dada una matriz A , calcular la potencia A^p de una matriz con $p \in \mathbb{R}$.

Generación de datos

1. Cargar el paquete **MASS** (instalarlo antes si no estuviese incluido en la instalación).
2. Crear dos clases C_1, C_2 cada una con 200 puntos $\mathbf{x} \in \mathbb{R}^2$ aleatorios de tal forma que las densidades condicionadas $p(\mathbf{x}|C_j)$, $j = 1, 2$ sean una mixtura de gaussianas, es decir:

$$\mathbf{x}|C_j \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_{j,1}, \boldsymbol{\Sigma}_{j,1}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_{j,2}, \boldsymbol{\Sigma}_{j,2}), j = 1, 2$$

siendo:

$$\boldsymbol{\mu}_{1,1} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \boldsymbol{\mu}_{1,2} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \boldsymbol{\mu}_{2,1} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \boldsymbol{\mu}_{2,2} = \begin{bmatrix} 5.5 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{j,i} = \sigma_{j,i} \cdot \mathbf{I}_{2 \times 2} \text{ con: } \begin{cases} \sigma_{1,1} = \sigma_{2,2} = 0.75 \\ \sigma_{2,1} = \sigma_{1,2} = 0.25 \end{cases}$$

Para ello, utilizar la función **mvrnorm** del paquete **MASS** para generar vectores aleatorios gaussianos.

3. Utilizar la codificación $Y \in \{-1, 1\}$ para las clases. Es decir, las clases son: $\mathcal{C}_1 = \{\mathbf{x} : Y(\mathbf{x}) = -1\}$ y $\mathcal{C}_2 = \{\mathbf{x} : Y(\mathbf{x}) = +1\}$.
Otra notación equivalente para las clases será su etiqueta: $Y = -1$ (para la clase \mathcal{C}_1) e $Y = +1$ (para la clase \mathcal{C}_2).
4. Dividir aleatoriamente el conjunto total de datos en dos muestras de tamaño 200: una para entrenamiento y otra para test.
5. Graficar en \mathbb{R}^2 el conjunto de datos de entrenamiento utilizando colores diferentes para las clases.

Evaluación de clasificadores mediante Validación Cruzada

1. Utilizando los datos anteriores, para un discriminante lineal:
 - Evaluar su eficacia (tasa de error de clasificación), mediante validación cruzada con 10 grupos (10-fold cv) sobre la muestra de entrenamiento.
 - Evaluar la eficacia en la muestra de test del modelo entrenado con la muestra de entrenamiento, comparándola con la obtenida en el punto anterior.
 - Graficar sobre el gráfico de dispersión de los puntos, la frontera de clasificación que produce el modelo entrenado con la muestra de entrenamiento.
2. Lo mismo para una regresión logística.
3. Lo mismo con un árbol de clasificación.
4. Lo mismo con una SVM.
5. Lo mismo con una red neuronal MLP.