

Entrega 4

Agustin Muñoz Gonzalez

2/6/2020

Preparamos el entorno

```
rm(list=ls())
```

En primer lugar definimos el clasificador ClasificoVecinos para lo cual vamos a usar una función auxiliar `k_posiciones_cercanas(X,xCentro,k)` que devuelve las `k` posiciones del vector de datos `X` que están mas cerca del dato `xCentro`.

```
k_posiciones_cercanas_X=function(X,xCentro,k){
  distancias=abs(X-xCentro)
  posiciones=c()
  for(i in (1:k)){
    menor_distancia=which.min(distancias)
    posiciones=c(posiciones,menor_distancia)
    distancias[menor_distancia]=NA
    # Le pongo NA al lugar de la menor distancia, asi en la
    # proxima iteracion la menor distancia cambia.
  }
  posiciones
}
ClasificoVecinos=function(X, Y, xNuevo, k){
  proporcion_F=mean(Y[k_posiciones_cercanas_X(X,xNuevo,k)=='F'])
  if(proporcion_F>=0.5)
  {'F'}else{'M'}
}
```

Definimos a continuación el clasificador ClasificoMovil.

```
ClasificoMovil=function(X, Y, xNuevo, h){
  if(sum(xNuevo-h<=X & X <= xNuevo+h)==0){
    NA}else{
      proporcion_F=mean(Y[xNuevo-h<=X
        & X <= xNuevo+h]=='F'])
      if(proporcion_F>=0.5)
      {'F'}else{'M'}
    }
}
```

Ahora definimos el clasificador ClasificoGenerativo junto con las estimaciones de las funciones de densidad `f_0`, `f_1` y las probabilidades $P(Y=0) \rightarrow$ `proporcion_F` y $P(Y=1) \rightarrow$ `proporcion_M`. Para estimar tales funciones y probabilidades necesitamos leer el archivo `alturas.testeo.csv`.

```
alturas=read.csv('alturas_n_500.csv')
attach(alturas)
```

```

ClasificoGenerativo=function(X, Y, xNuevo){
  f_0=function(x){
    media_M=mean(altura[genero=='M'])
    desvio_M=sd(altura[genero=='M'])
    dnorm(x,media_M,desvio_M)
  }
  f_1=function(x){
    media_F=mean(altura[genero=='F'])
    desvio_F=sd(altura[genero=='F'])
    dnorm(x,media_F,desvio_F)
  }
  proporcion_F=mean(Y=='F')
  proporcion_M=1-proporcion_F
  if(f_1(xNuevo)*proporcion_F>f_0(xNuevo)*proporcion_M)
  {'F'}else{'M'}
}

```

Por último mostramos los errores empíricos de cada uno de los clasificadores respecto de los datos de testeo `alturas.testeo.csv`. Lo haremos mediante el comando `lapply`, por lo que definimos una función auxiliar para cada clasificador donde fijamos `X=alturas_testeo$altura`, `Y=alturas_testeo$genero`, `k=10`, `h=1` y `xNuevo` lo dejamos como variable de la nueva función.

```

alturas_testeo=read.csv('alturas.testeo.csv')
ClasificoVecinos_test=function(xNuevo){
  ClasificoVecinos(alturas_testeo$altura,alturas_testeo$genero,xNuevo,10)
}
ClasificoMovil_test=function(xNuevo){
  ClasificoMovil(alturas_testeo$altura,alturas_testeo$genero,xNuevo,1)
}
ClasificoGenerativo_test=function(xNuevo){
  ClasificoGenerativo(alturas_testeo$altura,alturas_testeo$genero,xNuevo)
}
error_empirico_vecinos=mean(lapply(alturas_testeo$altura,ClasificoVecinos_test)
                             !=alturas_testeo$genero)
error_empirico_vecinos

```

```
## [1] 0.03225806
```

```

error_empirico_movil=mean(lapply(alturas_testeo$altura,ClasificoMovil_test)
                           !=alturas_testeo$genero)
error_empirico_movil

```

```
## [1] 0.03225806
```

```

error_empirico_generativo=mean(lapply(alturas_testeo$altura,ClasificoGenerativo_test)
                                !=alturas_testeo$genero)
error_empirico_generativo

```

```
## [1] 0.03225806
```

Dado que los 3 errores empíricos coinciden, podemos decir que cualquiera de los 3 clasificadores es un buen clasificador ya que solo tienen un 3% de error empírico.