kNN Penguins_punto extra

Medel Colorado Yoselin Merari

2022-06-05

#K-vecinos próximos

Instalar paquete

```
install.packages("MASS")
library(MASS)
install.packages("readxl")
library(readxl)
```

Cargar matriz PENGUINS

```
penguins_1_ <- read_excel("penguins (1).xlsx")

Z<-as.data.frame(penguins_1_)
colnames(Z)
```

```
## [1] "ID"              "especie"         "isla"            "largo_pico_mm"
## [5] "grosor_pico_mm"  "largo_aleta_mm"  "masa_corporal_g" "genero"
## [9] "año"
```

Definir la matriz de datos y la variable respuesta

Con las clasificaciones

```
x<-Z[,4:5]

y<-Z[,8]
```

Se definen las variables y observaciones

```
n<-nrow(x)
p<-ncol(x)
```

Gráfico scatter plot

Creación de un vector de colores

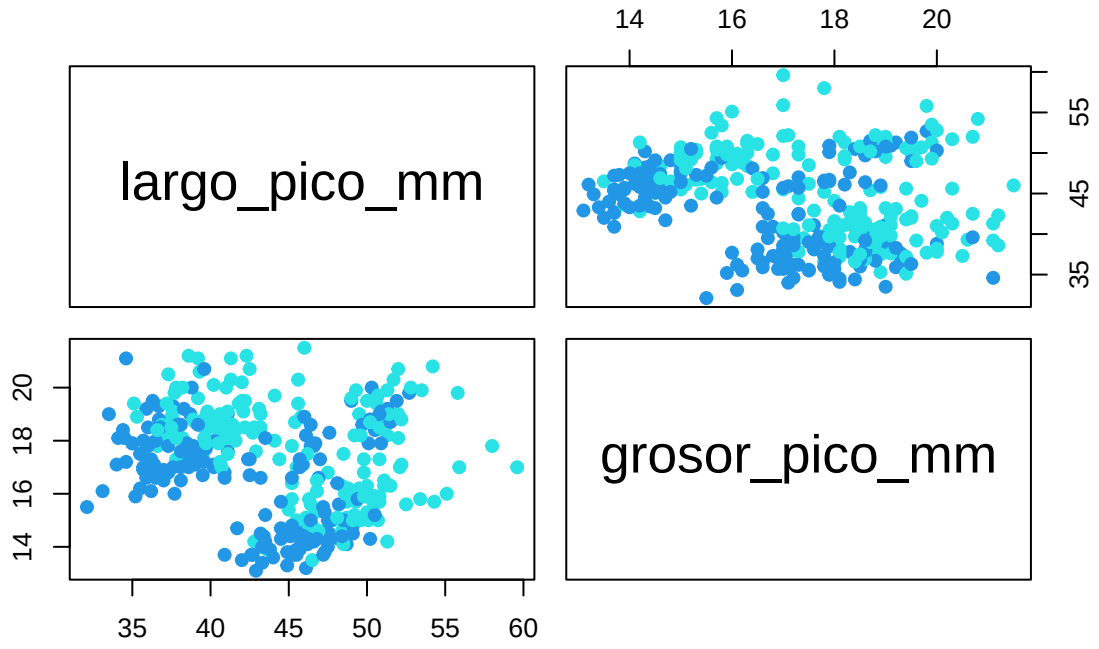
```
head(y)
```

```
## [1] "male"   "female" "female" "female" "female" "male"
```

```
col.iris<-c("blue","green","orange")[y]


pairs(x, main="Data set Penguins, largo_pico_mm (azul), grosor_pico_mm (verde)",
      pch=19, col = 4:5)
```

Data set Penguins, largo_pico_mm (azul), grosor_pico_mm (verde)



kNN

```
library(class)
```

Se fija una “semilla” para tener valores iguales

```
set.seed(1001)
```

Creación de los ciclos para k=1 hasta k=20. Selecciona el valor de k que tenga el error más bajo.

Inicialización de una lista vacia de tamaño 20

```
knn.class<-vector(mode="list",length=20)
knn.tables<-vector(mode="list", length=20)
```

Clasificaciones erroneas

```
knn.mis<-matrix(NA, nrow=20, ncol=1)
knn.mis
```

```
##       [,1]
##  [1,]   NA
##  [2,]   NA
##  [3,]   NA
##  [4,]   NA
##  [5,]   NA
##  [6,]   NA
##  [7,]   NA
##  [8,]   NA
##  [9,]   NA
## [10,]   NA
## [11,]   NA
## [12,]   NA
## [13,]   NA
## [14,]   NA
## [15,]   NA
## [16,]   NA
## [17,]   NA
## [18,]   NA
## [19,]   NA
## [20,]   NA
```

```
for(k in 1:20){
  knn.class[[k]]<-knn.cv(x,y,k=k)
  knn.tables[[k]]<-table(y,knn.class[[k]])
  # la suma de las clasificaciones menos las correctas
  knn.mis[k]<- n-sum(y==knn.class[[k]])
}


knn.mis
```

```
##       [,1]
##  [1,]   74
##  [2,]   71
##  [3,]   68
##  [4,]   64
##  [5,]   59
##  [6,]   59
##  [7,]   56
```

```
##  [8,]   58
##  [9,]   57
## [10,]   57
## [11,]   61
## [12,]   54
## [13,]   55
## [14,]   58
## [15,]   57
## [16,]   58
## [17,]   57
## [18,]   59
## [19,]   60
## [20,]   57
```

Número optimo de k-vecinos

```
which(knn.mis==min(knn.mis))
```

```
## [1] 12
```

```
knn.tables[[12]]
```

```
##
## y        female male
##   female    148   26
##   male       28  142
```

El más eficiente es k=14

Se señala el k mas eficiente

```
k.opt<-12

knn.cv.opt<-knn.class[[k.opt]]
head(knn.cv.opt)
```

```
## [1] male   female male   female female male
## Levels: female male
```

Tabla de contingencia con las clasificaciones buenas y malas

```
knn.tables[[k.opt]]
```

```
##
## y        female male
##   female    148   26
##   male       28  142
```

Cantidad de observaciones mal clasificadas

```
knn.mis[k.opt]
```

```
## [1] 54
```

Error de clasificación (MR)

```
knn.mis[k.opt]/n
```

```
## [1] 0.1569767
```

Gráfico de clasificaciones correctas y erroneas

```
col.knn.iris<-c("red","green")[1*(y==knn.cv.opt)+1]
pairs(x, main="Clasificación kNN de Penguins",
      pch=19, col=col.knn.iris)
```