Dendrograma

Sisi Guevara García

2/6/2022

Dendrograma

Introducción

Un dendrograma es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado.

Paqueterías nesesarías

```
library(cluster.datasets)
```

Base de datos

```
data("all.mammals.milk.1956")
AMM=all.mammals.milk.1956
```

Usaremos el data set de **"all.mammals.milk.1956"**, el cual contiene datos sobre la leche de diferentes especies de animales.

Revisión de la base de datos

Dimensión

```
dim(AMM)
```

```
## [1] 25  6
```

Esta base cuenta con 25 observaciones y 6 variables.

Datos faltantes

```
anyNA(AMM)
```

```
## [1] FALSE
```

La búsqueda sale negativa así que continuamos con el dendrograma.

Tipo de variables

```
str(AMM)
```

```
## 'data.frame':    25 obs. of  6 variables:
##  $ name   : chr  "Horse" "Orangutan" "Monkey" "Donkey" ...
##  $ water  : num  90.1 88.5 88.4 90.3 90.4 87.7 86.9 82.1 81.9 81.6 ...
##  $ protein: num  2.6 1.4 2.2 1.7 0.6 3.5 4.8 5.9 7.4 10.1 ...
##  $ fat    : num  1 3.5 2.7 1.4 4.5 3.4 1.7 7.9 7.2 6.3 ...
##  $ lactose: num  6.9 6 6.4 6.2 4.4 4.8 5.7 4.7 2.7 4.4 ...
##  $ ash    : num  0.35 0.24 0.18 0.4 0.1 0.71 0.9 0.78 0.85 0.75 ...
```

Encontramos que la base esta conformada por 5 variables numéricas y una carácter donde se encuentra registrado el nombre de los animales, en las numéricas teneos la cantidad de proteína, nivel de agua, grasa, lactosa, los minerales de la leche.

Cálculo de la matriz de distancias de Mahalonobis

```
dist.AMM<-dist(AMM[,2:6])
```

Calculamos la distancia de Mahalanobis para las variables que comprende de la dos a la seis, variables numéricas.

Con la distancia de Mahalanobis podemos calcular la similitud que existe entre las variables teniendo en cuenta la correlación que hay entre ellas.

Redondeo

```
round(as.matrix(dist.AMM)[1:6, 1:6],3)
```

```
##       1     2     3     4     5     6
## 1 0.000 3.327 2.494 1.226 4.759 4.107
## 2 3.327 0.000 1.206 2.794 2.798 2.592
## 3 2.494 1.206 0.000 2.375 3.716 2.348
## 4 1.226 2.794 2.375 0.000 3.763 4.007
## 5 4.759 2.798 3.716 3.763 0.000 4.176
## 6 4.107 2.592 2.348 4.007 4.176 0.000
```

Realizamos un redondeo de los cálculos de la distancia de Mahalanobis y los convertimos a una matriz, proyectamos e indicamos que solo usaremos a los primeros 6 individuos así que especificamos la selección de las 6 filas y 6 columnas pertenecientes a dichos individuos.

Calculo del dendrograma

```
dend.AMM<-as.dendrogram(hclust(dist.AMM))
```

Se calcula el dendrograma para nuestras observaciones elegidas donde usaremos el método de agrupación por Clústers **"hclust"**, el cual nos ofrece una agrupación jerárquica.

Graficación del dendrograma

Creamos un vector para las etiquetas que le asignaremos al Dendograma para el cual necesitaremos la librería "dendextend".

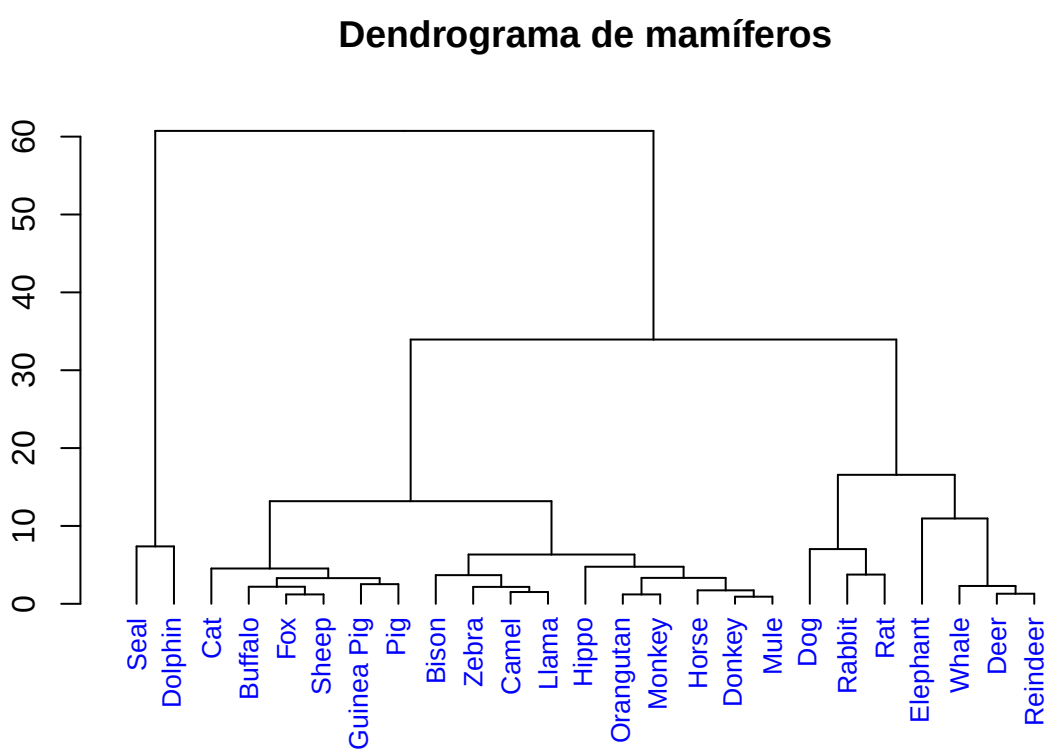
```
library(dendextend)
```

```
L=labels(dend.AMM)

labels(dend.AMM)=AMM$name[L]
```

Posteriormente graficamos el dendrograma cambiando el tamaño de las etiquetas y aplicando color a las etiquetas para que sobresalgan.

```
dend.AMM %>%
  set(what="labels_col", "blue") %>% #Colores etiqueta
  set(what="labels_cex", 0.8) %>%
  plot(main="Dendrograma de mamíferos")
```



Obtenemos el dendrograma agrupado y podemos ver que esta dividido en dos grupos en el primer grupo se encuentran la leche de las especies de foca y delfín son deferentes de el segundo grupo el cual esta sub-dividido en dos grupos más y dos mas para estos grupos que a su vez contienen más.

Pero lo interesante es ver que el gráfico nos muestra dos grupos en los que se pueden separara la leche de estos animales.