Análisis Estadístico Multivariante

Problemas propuestos de Análisis Cluster

Francisco Javier Mercader Martínez

Problema 1

El conjunto de datos **iris** de **R** contiene las variables Sepal.Length(X1, longitud de los sépalos), Sepal.Width(X2, anchura de los sépalos), Petal.Length(X3, longitud de los pétalos),Petal.Width (X4, anchura de los pétalos) medidas en centímetros, de tres especies diferentes de flor de iris (Species: *setosa*, *versicolor* y *virginica*). Se desea realizar un análisis cluster para determinar las diferentes agrupaciones que pueden obtenerse de las flores de iris a partir de las magnitudes de sus pétalos y sépalos. Se pide:

1. Realizar un agrupamiento de los datos en 3 grupos considerando los datos estandarizados aplicando algoritmo **k-means** con semilla **set.seed(123456)** y 10 inicios aleatorios con todas las variables.

```
set.seed(123456)

d <- iris[,1:4]
```

```
CA <- kmeans(d, centers = 3, nstart = 10)
```

   a. Indicar cuál es el tamaño de cada grupo.

```
CA$size
```

```
## [1] 62 50 38
```

   b. ¿A qué grupos pertenecen las flores 15, 75 y 123?

```
CA$cluster[c(15, 75, 123)]
```

```
## [1] 2 1 3
```

   c. Obtener los centroides de cada grupo.

```
CA$centers
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.901613    2.748387     4.393548    1.433871
## 2     5.006000    3.428000     1.462000    0.246000
## 3     6.850000    3.073684     5.742105    2.071053
```

   d. ¿Qué porcentaje de variabilidad se reduce con estos 3 grupos con respecto de la variabilidad total correspondiente a 1 solo grupo?

```
1 - CA$tot.withinss / CA$totss
```

```
## [1] 0.8842753
```

2. Hacer un ACP con las 4 variables.

```r
PCA <- prcomp(d, scale. = TRUE)
```

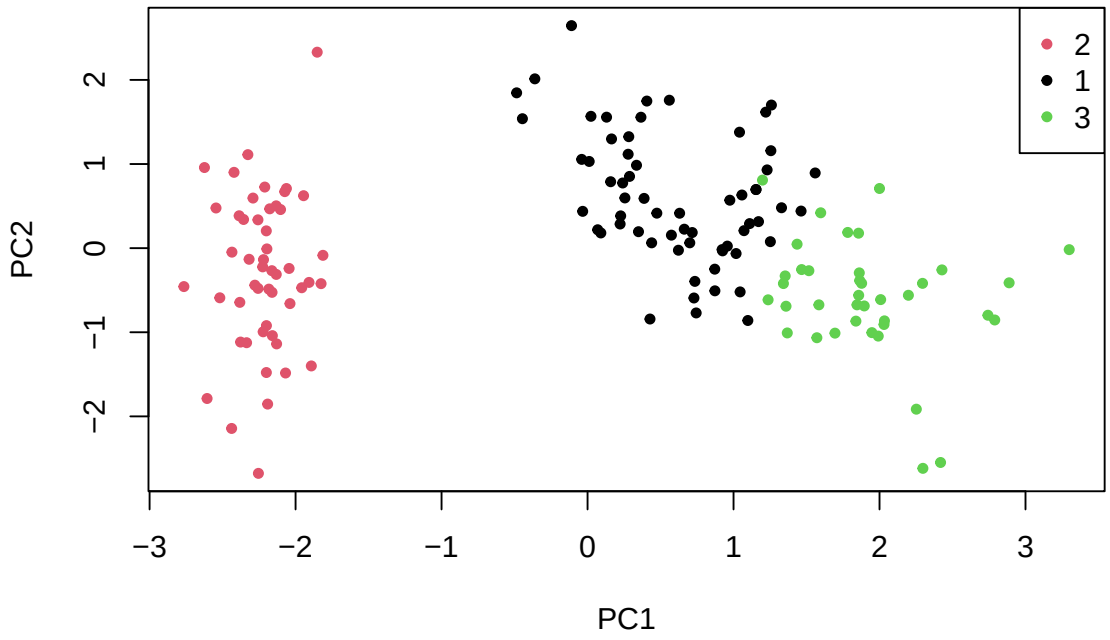
a. ¿Qué porcentaje de la variabilidad es explicada por las 2 primeras componentes?

```r
sum(PCA$sdev[1:2]^2) / sum(PCA$sdev^2)
```

```
## [1] 0.9581321
```

b. Representar gráficamente las puntuaciones no estandarizadas en el sistema de referencia definido por las 2 primeras componentes principales y etiquetar con símbolos los grupos obtenidos con el algoritmo de **k-means**.

```r
plot(PCA$x[, 1:2], col=CA$cluster, pch = 20)
legend("topright", legend = unique(CA$cluster), col=unique(CA$cluster), pch=20)
```



3. Realizar un análisis cluster jerarquizado de los datos usando la distancia euclídea con las variables estandarizadas y el método *complete*.

```r
ds <- scale(d)

# Calcular la matriz de distancias
matriz_distancias <- dist(ds, method = "euclidean")

# Agrupamiento jerárquico
hc <- hclust(matriz_distancias, method = "complete")
```

a. ¿Cuál es la distancia entre las flores 1 y 55?

```r
as.matrix(matriz_distancias)[1, 55]
```

```
## [1] 3.410625
```

   b. ¿Cuál es la distancia mínima? ¿A qué flores corresponden?

   c. ¿Cuál es el primer grupo que se forma? ¿Y el último?

   d. Si se formaran 2 grupos, ¿qué tamaño tendría cada grupo? Incluir los 2 grupos en el gráfico de las 2 primeras componentes principales y comentar cómo serían los grupos según este gráfico. ¿Dónde se incluirán las flores 1 y 55?

   e. Si se formaran 3 grupos, ¿qué tamaño tendría cada grupo? Incluir los 3 grupos en el gráfico de las 2 primeras componentes principales y comentar cómo serían los grupos según este gráfico. ¿Dónde se incluirán las flores 1 y 55?

   f. Representar el dendograma e indicar dónde se sitúan esas dos agrupaciones.

4. Comparar los 3 grupos formados con **`kmeans`** y con **`hclust`**.

   a. ¿Hay grupos coincidentes?

   b. ¿Qué grupos de **`kmeans`** se parece más al grupo 1 de **`hclust`**?

   c. ¿Cuántas flores tienen en común?

   d. ¿Cuáles son las flores que no tienen en común?

   e. ¿Cuáles son las principales diferencias entre todos los grupos de ambos métodos?

   f. Comparar los 3 grupos formados por cada método con los definidos por la especie de iris. ¿Cuáles es la especie más en cada grupo formado por cada método?

5. ¿Se podría obtener un número óptimo de grupos?

6. ¿Cómo cambiará la distribución de los grupos si se realizara un cluster jerárquico utilizando el método del enlace simple o del vecino más próximo (*single*)?