Análisis Cluster

Estadística multivariante

Francisco Solano López Rodríguez

Índice

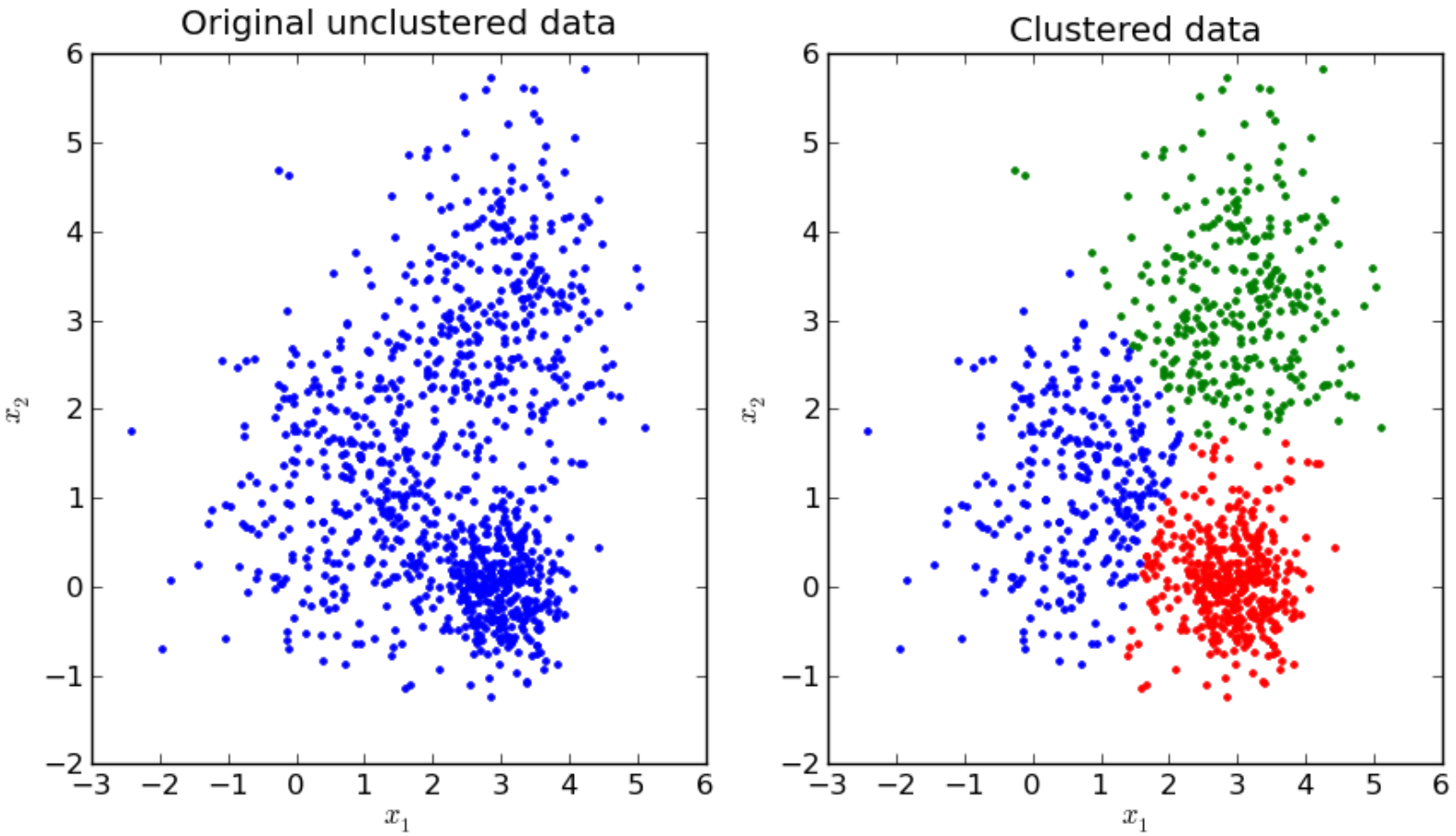
1. Introducción
2. Elección de las variables
3. Elección de las medidas de asociación
   - Distancias
   - Similaridades
4. Elección de la técnica cluster a emplear en el estudio
5. Caso práctico en R

Introducción

El Análisis Cluster, conocido como Análisis de Conglomerados, es un método estadístico multivariante de clasificación automática de datos, que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

Comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que llamaremos clusters.

Introducción



Aplicaciones del Análisis cluster

Algunas de las principales áreas de aplicación del Análisis Cluster son:

- Biología, biología computacional y bioinformática.
- Medicina.
- Empresarial y marketing.
- Análisis de red social.
- Agrupación de resultados de búsqueda.
- Sistemas de recomendación.
- Climatología.
- Etc, etc, ...

Etapas en Análisis Cluster

Las etapas a seguir en el empleo de una técnica cluster pueden ser resumidas en los siguientes puntos:

1. Elección de las variables.
2. Elección de la medida de asociación.
3. Elección de la técnica cluster a emplear en el estudio.
4. Validación de los resultados e interpretación de los mismos.

Índice

1. Introducción
2. Elección de las variables
3. Elección de las medidas de asociación
   - Distancias
   - Similaridades
4. Elección de la técnica cluster a emplear en el estudio
5. Caso práctico en R

Elección de las variables

Este paso es de gran importancia, en el se deberá elegir un conjunto concreto de características para describir a cada individuo.

Dependiendo del problema las variables pueden ser:

- Cualitativas:
  - Ordinales
  - Nominales
- Cuantitativas:
  - Discretas
  - Continuas

Índice

1. Introducción
2. Elección de las variables
3. Elección de las medidas de asociación
   - Distancias
   - Similaridades
4. Elección de la técnica cluster a emplear en el estudio
5. Caso práctico en R

Elección de las medidas de asociación

La mayoría de los métodos cluster requieren establecer una medida de asociación que permita medir la proximidad de los objetos en estudio.

En el Análisis Cluster de individuos la proximidad suele expresarse en términos de distancias.

En el Análisis Cluster de variables la proximidad suele expresarse en términos de unas funciones llamadas similaridades.

Medidas de asociación

Las medidas de asociación que vamos a considerar en un primer lugar son las siguientes:

- **Distancia:** cuando se elige una distancia como medida de asociación los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos debe ser pequeña.
- **Similaridad:** cuando se elige una similaridad los grupos formados contendrán individuos con una similaridad alta entre ellos.

Medidas de asociación: Distancias

Definición. Sea U un conjunto finito o infinito de elementos. Una función $d : U \times U \to \mathbb{R}$ se llama distancia métrica si $\forall x, y \in U$ se tiene:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z), \forall z \in U$

Medidas de asociación: similaridades

Definición. Sea U un conjunto finito o infinito de elementos. Una función $s : U \times U \to \mathbb{R}$ se llama similaridad si cumple las siguientes propiedades: $\forall x, y \in U$, s_0 número real finito arbitrario.

1. $s(x, y) \leq s_0$
2. $s(x, x) = s_0$
3. $s(x, y) = s(y, x)$

Definición. Una similaridad se llama similaridad métrica si verifica:

1. $s(x, y) = s_0 \to x = y$
2. $|s(x, y) + s(y, z)|s(x, z) \geq s(s, y)s(y, z), \forall z \in U$

Índice

1. Introducción
2. Elección de las variables
3. Elección de las medidas de asociación
   - Distancias
   - Similaridades
4. Elección de la técnica cluster a emplear en el estudio
5. Caso práctico en R

Elección de la técnica cluster a emplear en el estudio

Clasificación de las técnicas clusters

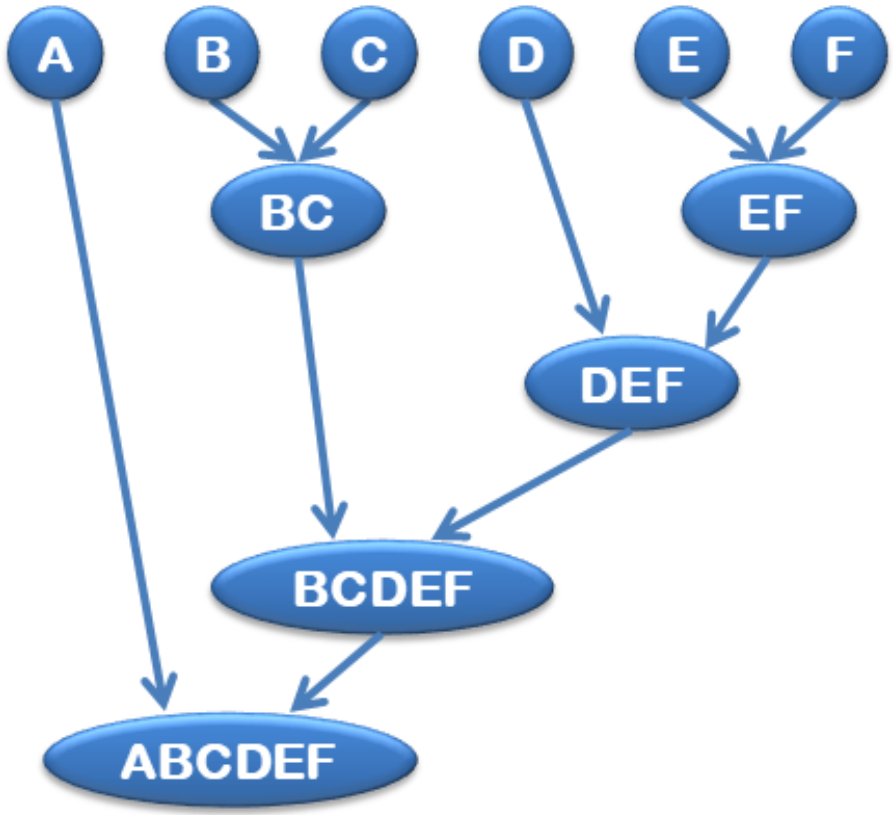
- **Métodos jerárquicos:** tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.
- **Métodos no jerárquicos:** también llamados de partición, tienen como objetivo dividir el conjunto de observaciones en K clusters, donde el valor de K ha sido definido previamente.

Métodos jerárquicos

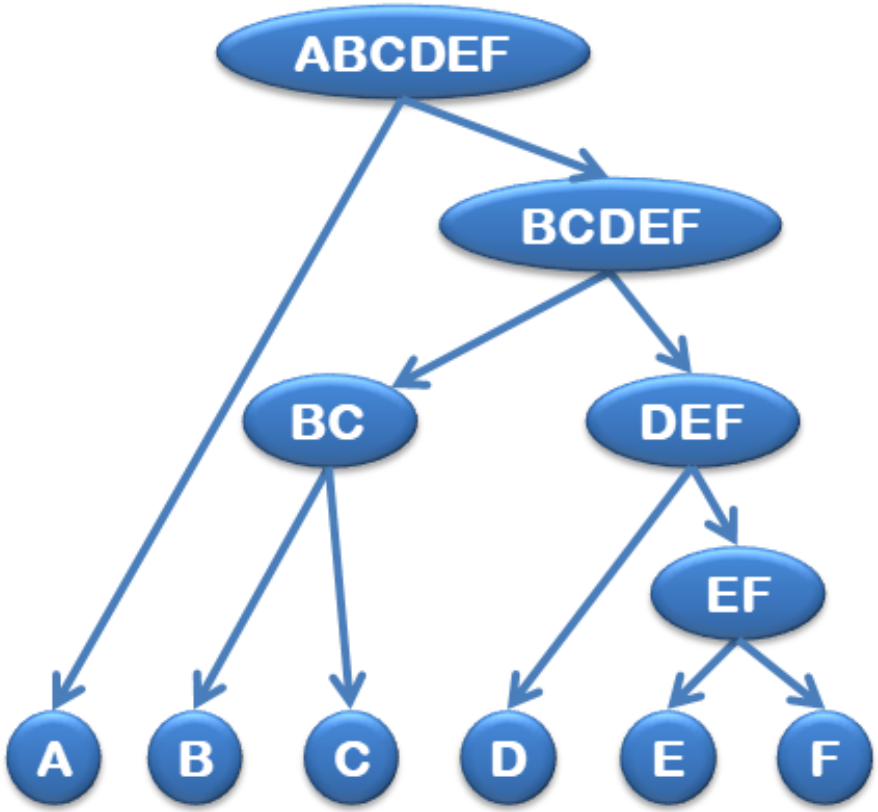
Métodos jerárquicos se subdividen en:

- **Métodos aglomerativos:** se parte de tantos grupos como individuos haya en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo.
- **Disociativos:** se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez más pequeños.

Métodos aglomerativos



Métodos disociativos



Métodos jerárquicos

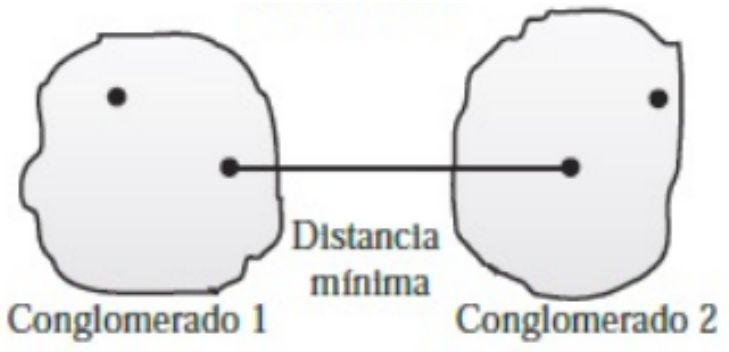
Los métodos jerárquicos para realizar la división o la aglomeración pueden utilizar diversas distancias realizar dicho proceso de división o aglomeración. Por ejemplo en el caso de aglomeración irán uniendo en cada nivel aquellos individuos o clusters que tengan una menor distancia entre ellos (o en el caso de la similitud se buscará una maximización).

En las siguientes diapositivas veremos aquellas distancias más comunes.

Métodos jerárquicos

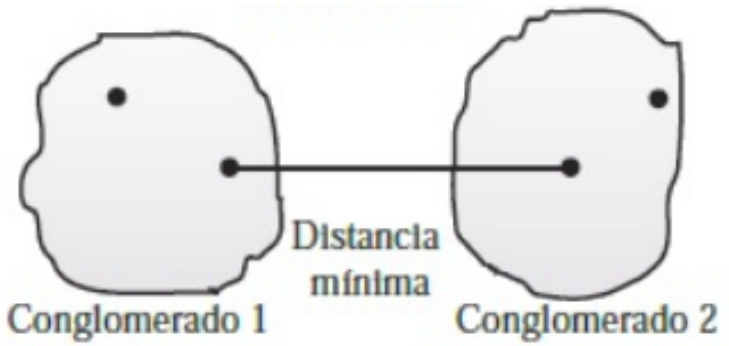
Distancia mínima (o similitud máxima)

$$d(C_i, C_j) = \min_{x_l \in C_i, x_m \in C_j} \{d(x_l, x_m)\}$$



Distancia máxima (o similitud mínima)

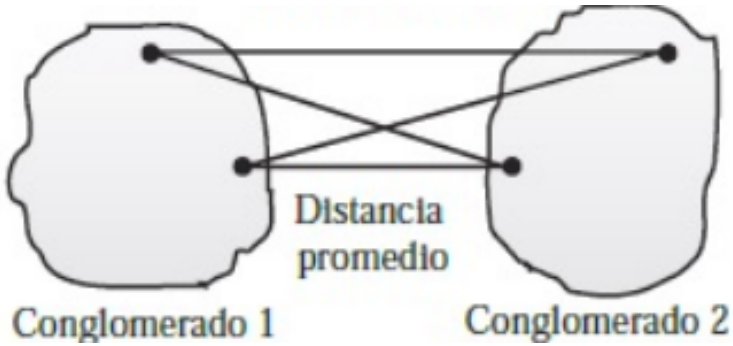
$$d(C_i, C_j) = \max_{x_l \in C_i, x_m \in C_j} \{d(x_l, x_m)\}$$



Métodos jerárquicos

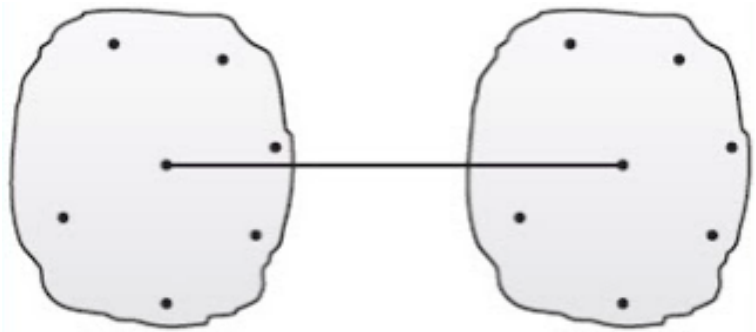
Distancia promedio

$$d(C_i, C_j) = \frac{1}{n_{C_i} n_{C_j}} \sum_{i \in C_i, j \in C_j} d(i,j)$$



Distancia entre centroides

$$d(C_i, C_j) = d(\bar{X}_{C_i}, \bar{X}_{C_j})$$



Métodos no jerárquicos

Métodos no jerárquicos se subdividen en:

- **Métodos de reasignación:** permiten que un individuo asignado a un grupo pueda ser reasignado en otro, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación optimice el resultado.
- **Métodos de búsqueda de la densidad:** se encuentran aquellos que proporcionan una aproximación tipológica y una aproximación probabilística.
- **Métodos directos:** Permiten clasificar simultáneamente a los individuos y a las variables.
- **Métodos de reducción de dimensiones:** consisten en la búsqueda de unos factores en el espacio de los individuos; cada factor corresponde a un grupo.

Elección de la técnica cluster a emplear en el estudio

- Jerárquicos
  - Aglomerativos
    - Simple Linkage (Vecino más próximo)
    - Complete Linkage (Vecino más lejano)
    - Promedio entre Grupos
    - Método del Centroide
    - Método de la Mediana
    - Método de Ward
  - Disociativos
    - Linkage Simple
    - Linkage Completo
    - Promedio entre Grupos
    - Método del Centroide
    - Método de la Mediana
    - Método de Ward
    - Análisis de Asociación
- No Jerárquicos
  - Reasignación
    - K – Medias
    - Nubes Dinámicas
  - Búsqueda de densidad
    - Análisis Modal
    - Métodos Taxap
    - Método de Fortin
    - Método de Wolf
  - Métodos Directos: Block – Clustering
  - Métodos Reductivos: Análisis Factorial tipo Q

Índice

1. Introducción
2. Elección de las variables
3. Elección de las medidas de asociación
   - Distancias
   - Similaridades
4. Elección de la técnica cluster a emplear en el estudio
5. Caso práctico en R

Caso práctico en R

Vamos a ver ahora un caso práctico en R, para ello vamos a utilizar el conjunto de datos que suele usarse para iniciarse en el clustering, debido a su sencillez. Este conjunto se trata del famoso Iris Dataset.
Para ello vamos a comenzar cargando el paquete 'datasets' y el conjunto de datos 'iris'.

```
1   require("datasets")
2   data("iris") # load Iris Dataset
3   str(iris) #view structure of dataset
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Caso práctico en R

Para entender mejor el conjunto de datos veamos un resumen estadístico del conjunto de datos con la función summary y veamos las primeras filas con head.

```
5   summary(iris)
6   head(iris)
```

```
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Caso práctico en R

Lo siguiente que vamos a hacer es preprocesar los datos. Para ello en primer lugar vamos a eliminar la etiqueta, ya que no la necesitamos para el clustering. Después vamos a definir una función para normalizar los datos y tras esto vamos a proceder a la normalización.

```
 8  # eliminamos la etiqueta
 9  iris.new<- iris[,c(1,2,3,4)]
10  iris.class<- iris[,"Species"]
11
12  # definimos la siguiente función para normalizar
13  normalize <- function(x){
14    return ((x-min(x))/(max(x)-min(x)))
15  }
16
17  # normalizamos los datos
18  iris.new$Sepal.Length<- normalize(iris.new$Sepal.Length)
19  iris.new$Sepal.Width<- normalize(iris.new$Sepal.Width)
20  iris.new$Petal.Length<- normalize(iris.new$Petal.Length)
21  iris.new$Petal.Width<- normalize(iris.new$Petal.Width)
```

Caso práctico en R

Apliquemos ahora el algoritmo kmeans sobre los datos para obtener los cluster. Vamos a definir k = 3. Tras realizar el clustering vamos a ver el número de individuos que hay en cada cluster y vamos a ver los centros de cada uno de los cluster

```
31  # aplicamos el algoritmo de clustering k-meas
32  result<- kmeans(iris.new,3) #aplly k-means algorithm with no. of centroids(k)=3
33
34  result$size # gives no. of records in each cluster
35  result$centers # gives value of cluster center datapoint value(3 centers for k=3)
```

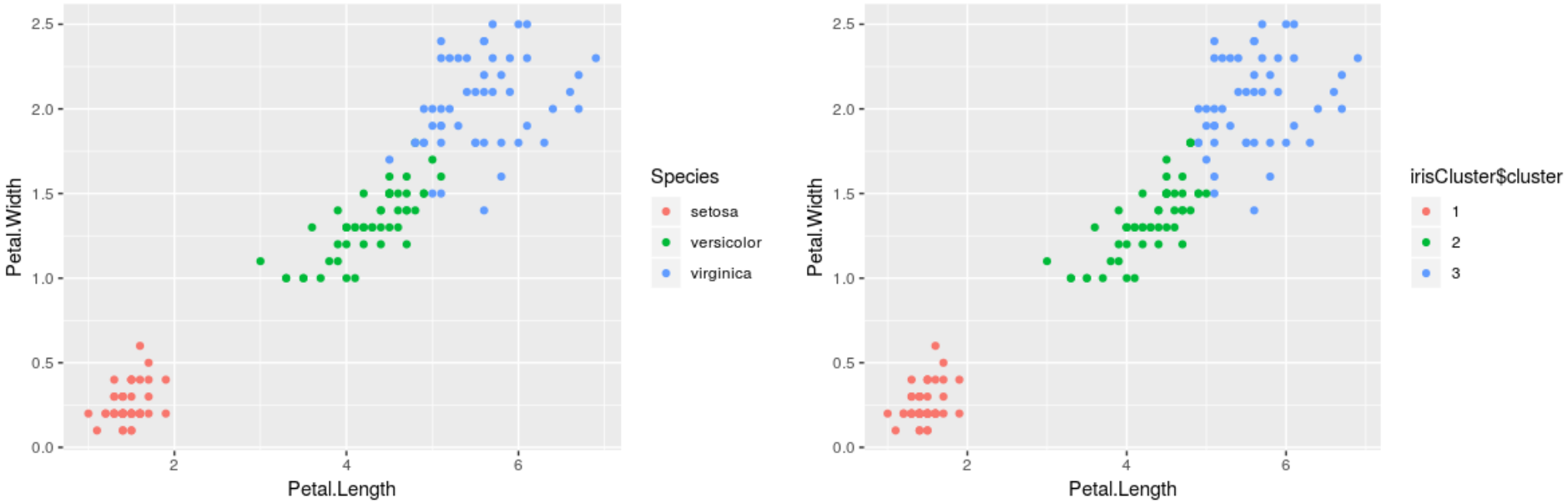
```
> result$size # gives no. of records in each cluster
[1] 39 50 61

> result$centers
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1    0.7072650   0.4508547   0.79704476  0.82478632
2    0.1961111   0.5950000   0.07830508  0.06083333
3    0.4412568   0.3073770   0.57571548  0.54918033
```

Caso práctico en R

Por últimos comparemos en un gráfico entre la anchura y longitud de pétalo para ver las diferencias entre las etiquetas originales y los cluster obtenidos. Para ello debemos cargar la biblioteca ggplot2, la cual deberemos de instalar en caso de no tenerla.

```
45  # cargamos ggplot2
46  library(ggplot2)
47  result$cluster <- as.factor(result$cluster)
48  ggplot(iris, aes(Petal.Length, Petal.Width, color = result$cluster)) + geom_point()
49  ggplot(iris, aes(Petal.Length, Petal.Width, color = Species)) + geom_point()
```



FIN DE LA PRESENTACIÓN