Clusters: Análisis de Conglomerados

Análisis de Conglomerados

A. M. Bianco & P. M. Spano

IntroAE

2021-05-19

Haciendo un poco de orden. . .

Haciendo un poco de orden. . .

Idea Básica

La idea de clustering es la de hallar grupos o conglomerados en un conjunto de datos.

Veamos un ejemplo.

Old Faithful Geyser

Datos de Old Faithful Geyser

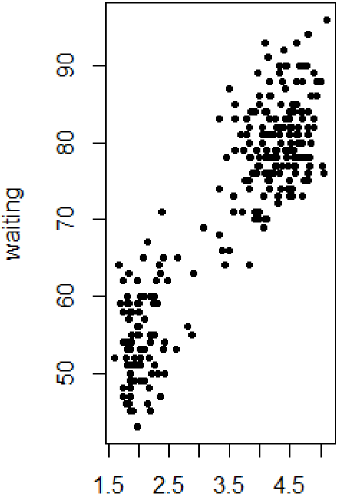
faithful {datasets}

Old Faithful geyser en Yellowstone National Park: uno de los geysers mejor estudiados.

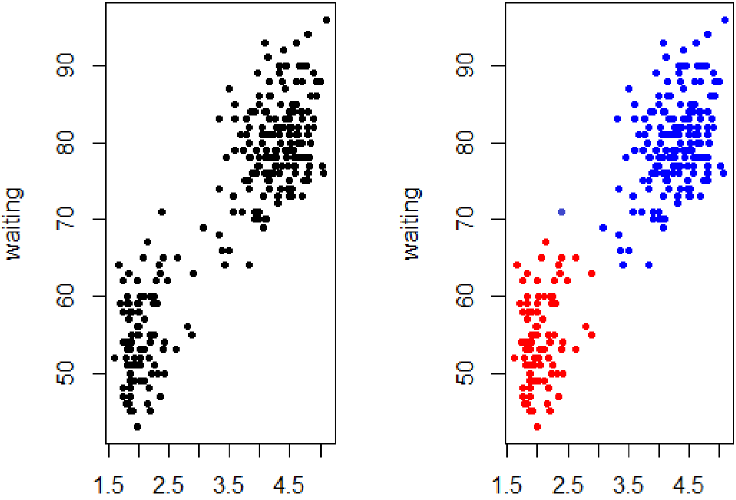
- duration: duración de una erupción (en minutos)
- waiting: tiempo de espera hasta la siguiente erupción (en minutos).

$n = 272$.

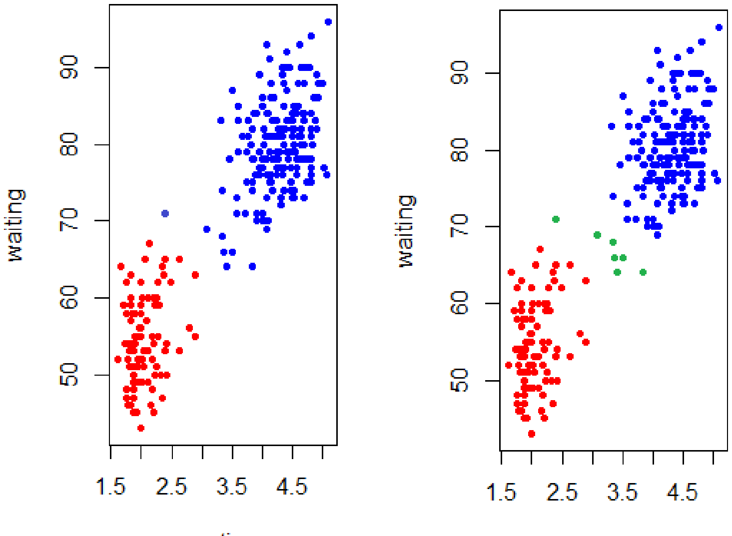
Datos de Old Faithful Geyser



Datos de Old Faithful Geyser



Datos de Old Faithful Geyser



Volviendo a las Ideas Básicas

La idea de clustering es la de hallar grupos o conglomerados en un conjunto de datos.

Tengamos en cuenta que el agrupamiento y el ordenamiento son tareas básicas en el desarrollo del pensamiento humano.

Son técnicas usadas en ámbitos muy diversos: ciencias sociales, biología, psicología, etc.

Volviendo a las Ideas Básicas

La idea de clustering es la de hallar grupos o conglomerados en un conjunto de datos.

Técnicas de Aprendizaje No Supervisado (*unsupervised classification*).

- Clasificación: tenemos clases predefinidas, queremos predecir la clase de un nuevo sujeto (aprendizaje supervisado).
- Clustering: determinamos las clases que son desconocidas (aprendizaje no supervisado)

¿Por qué agrupar?

1. exploración de los datos y búsqueda de patrones interesantes que puedan dar lugar a nuevas interpretaciones, preguntas o hipótesis de trabajo.
2. reducción de la información y/o de la complejidad para un análisis posterior a la investigación.
3. investigación de una relación entre el clustering y otros agrupamientos o características de los datos.

Insumos

Medimos p características en n puntos: $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $1 \leq i \leq n$.

$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

$\mathbf{X}$: matriz de $n \times p$

Cada fila es un punto $p-$dimensional.

Cada columna corresponde a una variable.

Objetivo

Dividir los datos en clusters de puntos de modo que los que están dentro de un mismo cluster sean similares (*within-cluster homogeneity*) y a la vez distintos a los de cualquier otro conglomerado (*between-cluster separation*).

Preguntas

- Similitud / Disimilitud
- ¿Cuántos grupos?
- Evaluación de los clusters resultantes

Un poco más técnicamente

Nuestra tarea es agrupar los puntos $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ en K conjuntos disjuntos $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$, siendo $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ *la clusterización*.

$\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ produce una partición de $\mathcal{D}$: $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_K = \mathcal{D}$ y además $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, i \neq j$.

El mecanismo de *cluster* asigna cada punto a un conglomerado: $\mathbf{x}_i \Longleftrightarrow \mathcal{C}_k$, luego la etiqueta de $\mathbf{x}_i$ es k.

Disimilitudes: ¿cuán disímiles son $\mathbf{x}$ e $\mathbf{y}$?

Dados dos puntos $\mathbf{x}$ e $\mathbf{y} \in \mathbb{R}^p$ (variables cuantitativas)

Distancia Euclídea

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{p} (x_j - y_j)^2 \right]^{1/2}$$

Disimilitudes: ¿cuán disímiles son $\mathbf{x}$ e $\mathbf{y}$?

Dados dos puntos $\mathbf{x}$ e $\mathbf{y} \in \mathbb{R}^p$ (variables cuantitativas)

Distancia Euclídea

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{p} (x_j - y_j)^2 \right]^{1/2}$$

En general, $d(x, y)$ es una disimilitud si:

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
3. $d(\mathbf{x}, \mathbf{y}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{y}$

Si además cumple

4. **Desigualdad triangular :** $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$

es una métrica.

Matriz de Disimilitud

- D matrix de $n \times n$ donde $D_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$
- Cuando d es una disimilitud, entonces la matriz D es simétrica y con diagonal nula.
- D depende fuertemente de la disimilitud elegida y determina fuertemente el análisis.

Disimilitud Total

$$T = \frac{1}{2} \sum_{i,j} d_{ij}$$

Disimilitud Total

$$\begin{aligned} T &= \frac{1}{2}\sum_{i,j} d_{ij} \\ &= \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\sum_{j:\mathbf{x}_j\in\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right) + \sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\sum_{j:\mathbf{x}_j\notin\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right)\right] \qquad (1) \end{aligned}$$

Disimilitud Total

$$
\begin{aligned}
T &= \frac{1}{2}\sum_{i,j} d_{ij} \\
&= \frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\ \sum_{j:\mathbf{x}_j\in\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right) + \sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\ \sum_{j:\mathbf{x}_j\notin\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right)\right] \qquad (1)\\
&= \frac{1}{2}\sum_{k=1}^{K}\sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\ \sum_{j:\mathbf{x}_j\in\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right) + \frac{1}{2}\sum_{k=1}^{K}\sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\ \sum_{j:\mathbf{x}_j\notin\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&= \mathsf{Within}(\mathcal{C}) + \mathsf{Between}(\mathcal{C})
\end{aligned}
$$

Notemos que T, disimilitud total entre todas las observaciones, no depende de la clusterizacion.

Similitud intra y entre clusters

$$
\begin{aligned}
T &= \frac{1}{2}\sum_{k=1}^{K}\sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\sum_{j:\mathbf{x}_j\in\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right) + \frac{1}{2}\sum_{k=1}^{K}\sum_{i:\mathbf{x}_i\in\mathcal{C}_k}\sum_{j:\mathbf{x}_j\notin\mathcal{C}_k} d\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&= \mathsf{Within}(\mathcal{C}) + \mathsf{Between}(\mathcal{C})
\end{aligned}
$$

$$
\mathsf{Within}(\mathcal{C}) \quad \longleftrightarrow \quad \mathsf{Between}(\mathcal{C})
$$

Si hacemos decrecer a Within($\mathcal{C}$), aumentamos a Between($\mathcal{C}$) y viceversa.

Función Objetivo

Simplificando un poco la notación:

$$W(\mathcal{C}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{C}_k} d\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

¿Cómo minimizamos?

Tenemos $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Supongamos que queremos formar K grupos.

Buscamos $\mathcal{C}_1, \ldots, \mathcal{C}_K$ tales que

- $\#\{\mathcal{C}_j\} > 0$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$
- $\cup_{i=1}^{K} \mathcal{C}_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

¿Cómo minimizamos?

Tenemos $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Supongamos que queremos formar K grupos.

Buscamos $\mathcal{C}_1, \ldots, \mathcal{C}_K$ tales que

- $\#\{\mathcal{C}_j\} > 0$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$
- $\cup_{i=1}^K \mathcal{C}_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

El número de posibles particiones (Jain & Dubes, 1988) es

$$S(n, K) = \frac{1}{K!} \sum_{j=1}^{K} (-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) (K-j)^n \approx \frac{K^n}{K!}$$

¿Cómo minimizamos?

Tenemos $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Supongamos que queremos formar K grupos.

Buscamos $\mathcal{C}_1, \ldots, \mathcal{C}_K$ tales que

- $\#\{\mathcal{C}_j\} > 0$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$
- $\cup_{i=1}^{K} \mathcal{C}_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

El número de posibles particiones (Jain & Dubes, 1988) es

$$S(n, K) = \frac{1}{K!} \sum_{j=1}^{K} (-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) (K - j)^n \approx \frac{K^n}{K!}$$

Por ejemplo, $S(19, 3) = 1.9 \times 10^8$.

Si K no se especifica, tenemos $T = \sum_{K=1}^{n} S(n, K)$ configuraciones. Para $n = 25$, $T > 4 \times 10^{18}$.

Resumiendo. . .

- Minimizar $W(\mathcal{C})$
- Evaluar en todas las posibles clusterizaciones conduciría a un mínimo global.
- Con datos reales puede ser impracticable.
- En la práctica solo es factible examinar una pequeña fracción de clusterizaciones.
- El objetivo es identificar una pequeña fracción que tenga posibilidades de contener al óptimo, o al menos una buena partición subóptima.

Tomando un atajo: $K-$medias

Estrategia no exhaustiva, muy popular y muy intuitiva. MacQueen (1967).

Supongamos que tenemos K grupos $\mathcal{C}_1, \ldots, \mathcal{C}_K$.

- Usamos el cuadrado de la distancia euclídea:

$$\|\mathbf{x}-\mathbf{y}\|^2 = \sum_{j=1}^{p} (x_j - y_j)^2$$

- Llamamos $n_k = \#\mathcal{C}_k$ (cantidad de puntos en el cluster).
- Consideramos el promedio de todos los puntos del cluster: $\overline{\mathbf{x}}_k = (\overline{x}_{1k}, \ldots, \overline{x}_{pk})$, es

$$\overline{\mathbf{x}}_k = \frac{1}{n_k} \sum_{\ell : \mathbf{x}_\ell \in \mathcal{C}_k} \mathbf{x}_\ell$$

$K-$medias

Función Objetivo:

$$W(\mathcal{C}) = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2$$

Es fácil comprobar que

$$W(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i:\mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \overline{\mathbf{x}}_k\|^2 \tag{2}$$

donde $\overline{\mathbf{x}}_k = (\overline{x}_{1k}, \ldots, \overline{x}_{pk})$ es el vector de p medias de cada cluster.

Ya que

para el $k-$ésimo cluster:

$$\frac{1}{n_k} \sum_{i,j:\mathbf{x}_i,\mathbf{x}_j \in \mathcal{C}_k} \sum_{s=1}^{p} (x_{is} - x_{js})^2 = 2 \sum_{i:\mathbf{x}_i \in \mathcal{C}_k} \sum_{s=1}^{p} (x_{is} - \overline{x}_{sk})^2 \qquad (3)$$

$K-$medias

Dado un cluster, digamos k,tenemos que

$$\overline{\mathbf{x}}_k = \arg\min_{\mathbf{m}_k} \sum_{i:\mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2$$

Luego, esto sugiere

$$\min_{\mathbf{m}_1,\ldots,\mathbf{m}_k} \sum_{k=1}^{K} \sum_{i:\mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{m}_K\|^2$$

Algoritmo de $K-$medias

- Comenzar con una asignación de K centros iniciales: $\mathbf{m}_1, \ldots, \mathbf{m}_K$
- Para cada $\mathbf{x}_i$ encontrar el centro $\mathbf{m}_k$ más cercano y asignar a ese cluster.
- Para cada cluster computar los K vectores de medias.
- Reasignar las observaciones al cluster más cercano en base a las medias computadas.
- Iterar hasta que no haya más reasignaciones.

Asegura convergencia, pero puede hacerlo a un mínimo local.

Algunas cuestiones prácticas

Para aplicar el método de $K-$medias necesitamos

- una inicialización
- seleccionar el número de clusters, K (acá esta el desafío)

Sobre la inicialización

Posibilidades:

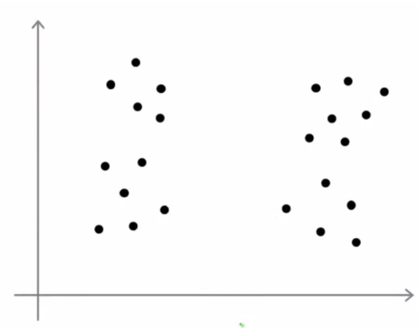
- Dividir aleatoriamente las observaciones en K grupos y tomar sus promedios como el centro de cada grupo.
- Tomar como centros los puntos más alejados entre sí.
- Construir grupos iniciales con información a priori y calcular sus centros.
- Seleccionar centros iniciales con información a priori.

Sobre K

El número de clusters depende del objetivo.

- En un problema de segmentación K puede estar definido de antemano.
  Ej: delivery– repartidores
- En un análisis descriptivo de los datos, donde se intenta explorar en cuántos conglomerados naturalmente se agrupan los datos, K^* es desconocido.

¿ K ?

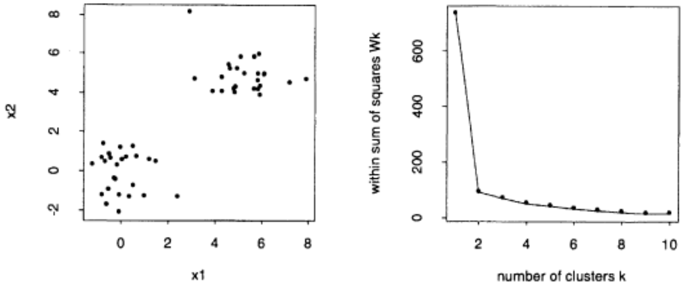


Eligiendo K

- Típicamente los métodos determinan K^* estudiando a la variación intra-clusters W como función de K: W_K.
- W_K tiende a disminuir con el número de clusters.
- **Elbow**: El K óptimo se corresponde con un quiebre en el gráfico de $W(\mathcal{C})$ aumentando la cantidad de clusters.

Método Elbow

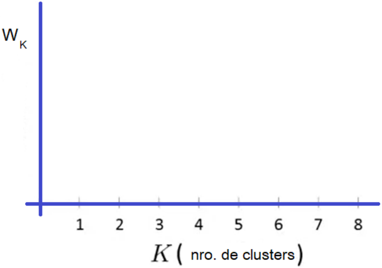
Graficamos K vs. W_K buscando un quiebre o un codo.

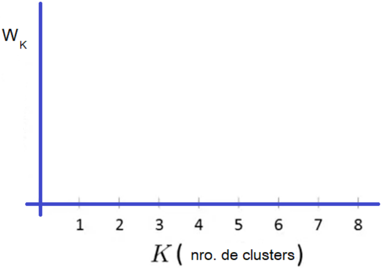


$W_1 >> W_2$ pues los grupos que vemos fueron asignados a clusters separados.

Un menor decrecimiento se observa al pasar de W_2 a W_3.

Método Elbow





Validación de Clusters

Se quiere evaluar la calidad del clustering: informativo? confiable?

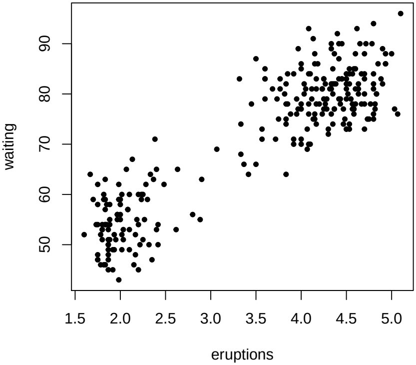
- Exploración visual
- Validación interna de los clusters: buscan que los clusters sean lo más homogéneos posibles y los más diferentes a los otros los clusters. Se calculan índices, por ejemplo estabilidad, silhouette y Dunn.
- Uso de información externa

Bibliografía específica

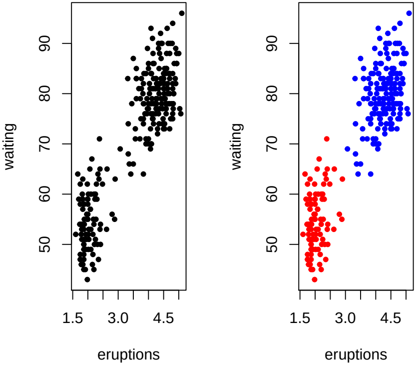
- Hastie, T., Tibshirani, R. & Friedman, J. H. (2001). The elements of statistical learning: data mining, inference, and prediction. New York: Springer.
- Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2015). Handbook of Cluster Analysis (1st ed.). Chapman and Hall/CRC.
- https://www.andrewng.org/courses/

Ejemplo

```
geyser0<- datasets::faithful
plot(geyser0$eruptions,geyser0$waiting,pch=20,xlab="eruptions",ylab="waiting")
```

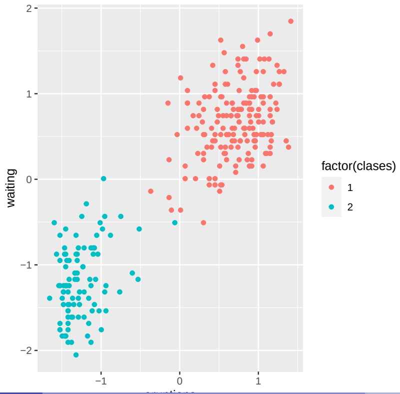


Ejemplo



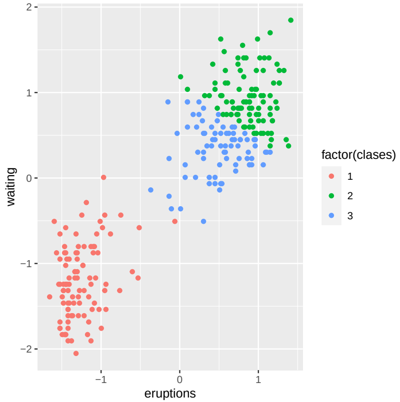
Ejemplo

```
library(ggplot2)
library(factoextra)
datos.sc<- scale (geyser0)
set.seed(123)
cluster2 <- kmeans(datos.sc, centers = 2, nstart = 30)
# nstart: particiones iniciales. Se queda con la mejor
# nstart >= 25

dd<- as.vector(cluster2$cluster)
datos.sc<- as.data.frame(datos.sc)
datos.sc$clases<- dd
ggplot(data=datos.sc,aes(eruptions, waiting, color = factor(clases)))+geom_point()
```



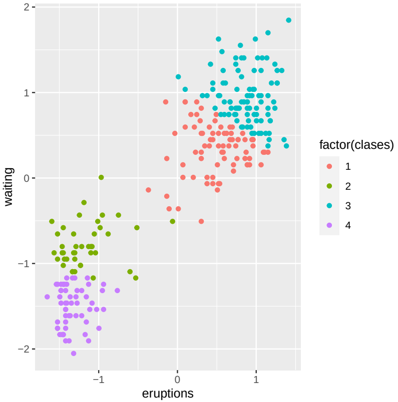
Ejemplo

```
set.seed(123)
cluster3 <- kmeans(datos.sc, centers = 3, nstart = 30)

dd<- as.vector(cluster3$cluster)
datos.sc<- as.data.frame(datos.sc)
datos.sc$clases<- dd
ggplot(data=datos.sc,aes(eruptions, waiting, color = factor(clases)))+geom_point()
```



Ejemplo

```
set.seed(123)
cluster4 <- kmeans(datos.sc, centers = 4, nstart = 30)

dd<- as.vector(cluster4$cluster)
datos.sc<- as.data.frame(datos.sc)
datos.sc$clases<- dd
ggplot(data=datos.sc,aes(eruptions, waiting, color = factor(clases)))+geom_point()
```



Ejemplo