



Taller: K-Means en R

Dra. Nelva Nely Almanza Ortega

Investigadora Catedrática, Conahcyt

Investigadoras e Investigadores por México



R-Ladies MX

Agenda

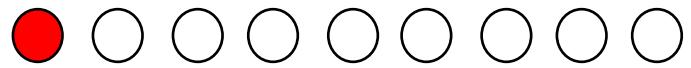
1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos

Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



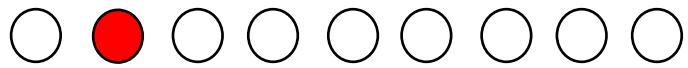
1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



Introducción



Disponemos de tanta información, que a veces es imposible procesarla con efectividad.



1. Introducción
(10 minutos)
 2. Algoritmo K-Means.
(5 minutos)
 3. K-Means en R.
(20 minutos)
 4. Variantes de K-Means en R.
(10 minutos)
 5. K-Means dinámico.
(5 minutos)
 6. Receso.
(15 minutos)
 7. Ejercicios con benckmark.
(45 minutos)
 8. Comentarios finales.
(10 minutos)

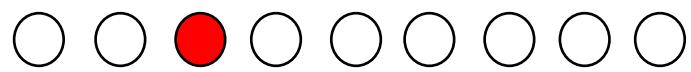
Introducción



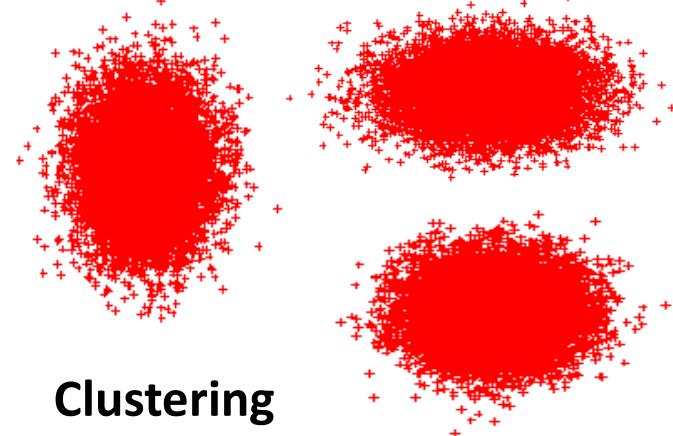
La clave está en descubrir patrones usando algoritmos para sacarle el máximo provecho.



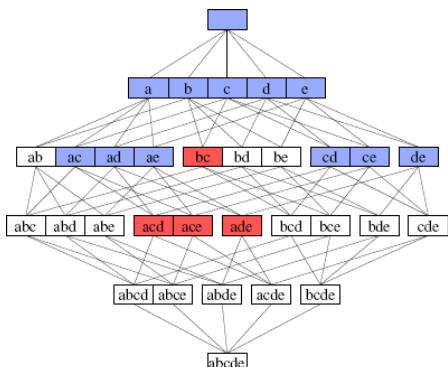
1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



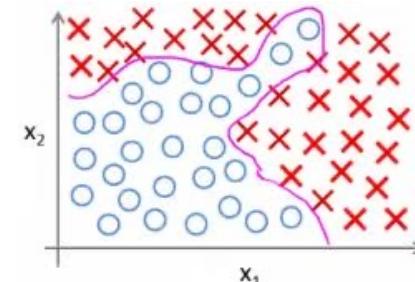
Introducción



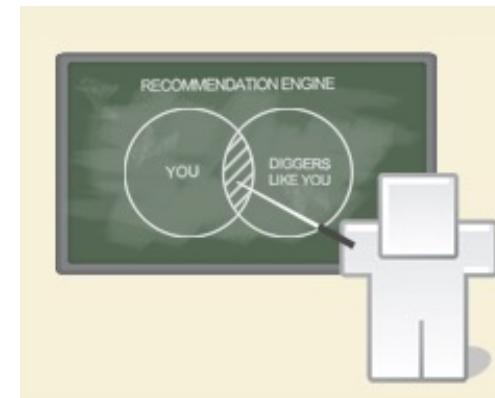
Clustering



Association



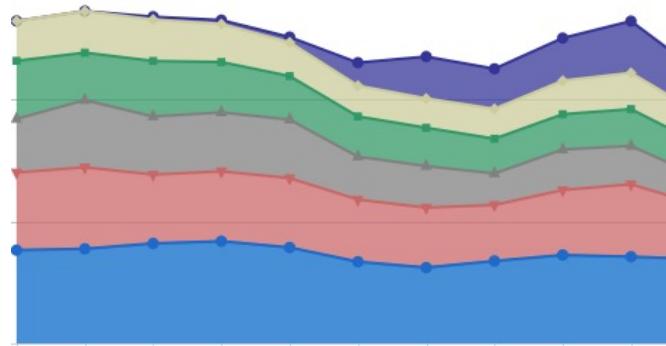
Classification



Recommendation Systems



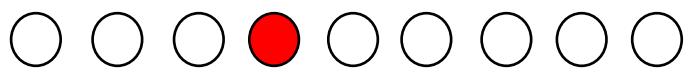
R-Ladies MX



**Real Time Analytics/
Big Data Streams**



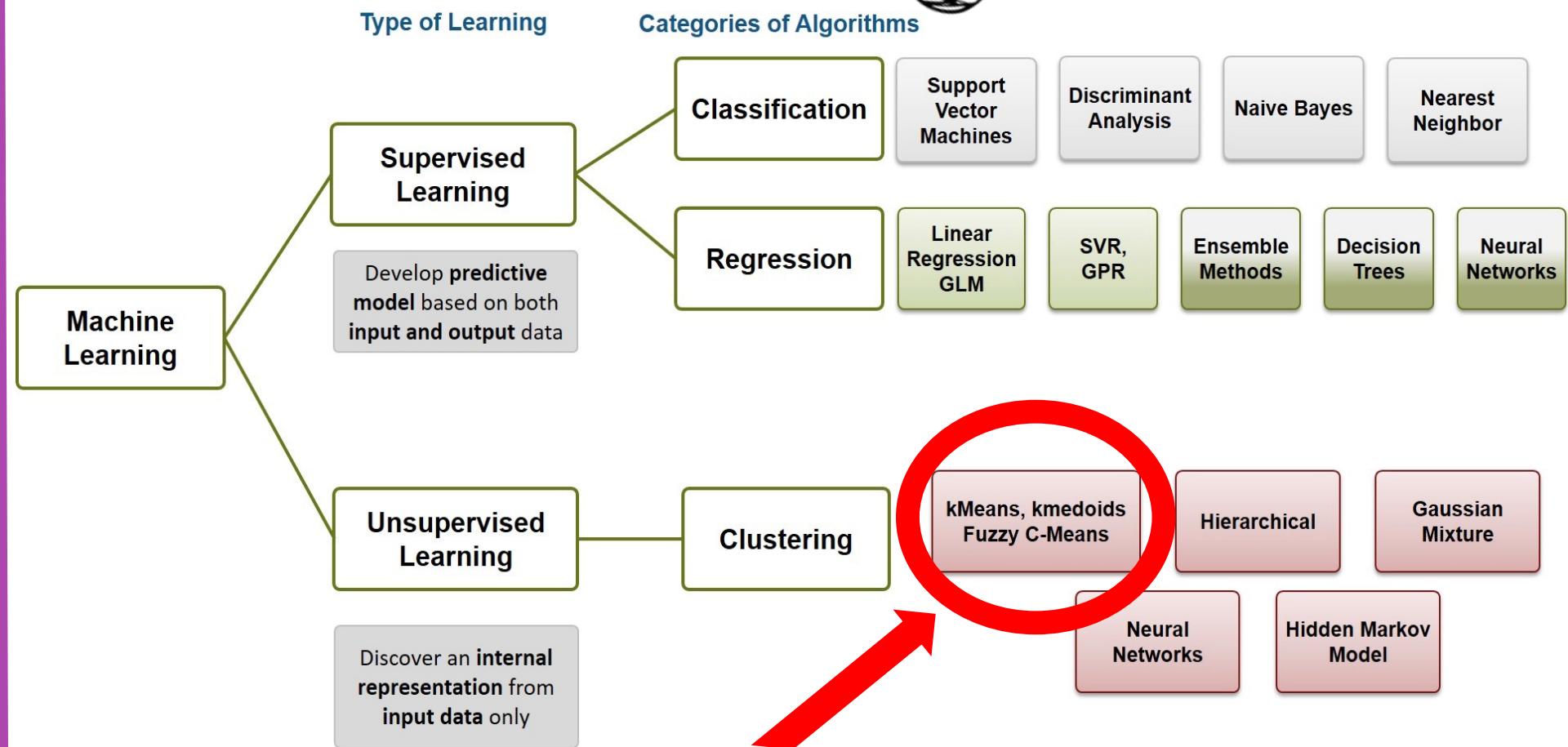
**Social Media Mining
Social Big Data**



Top 10

Introducción

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



Conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos.

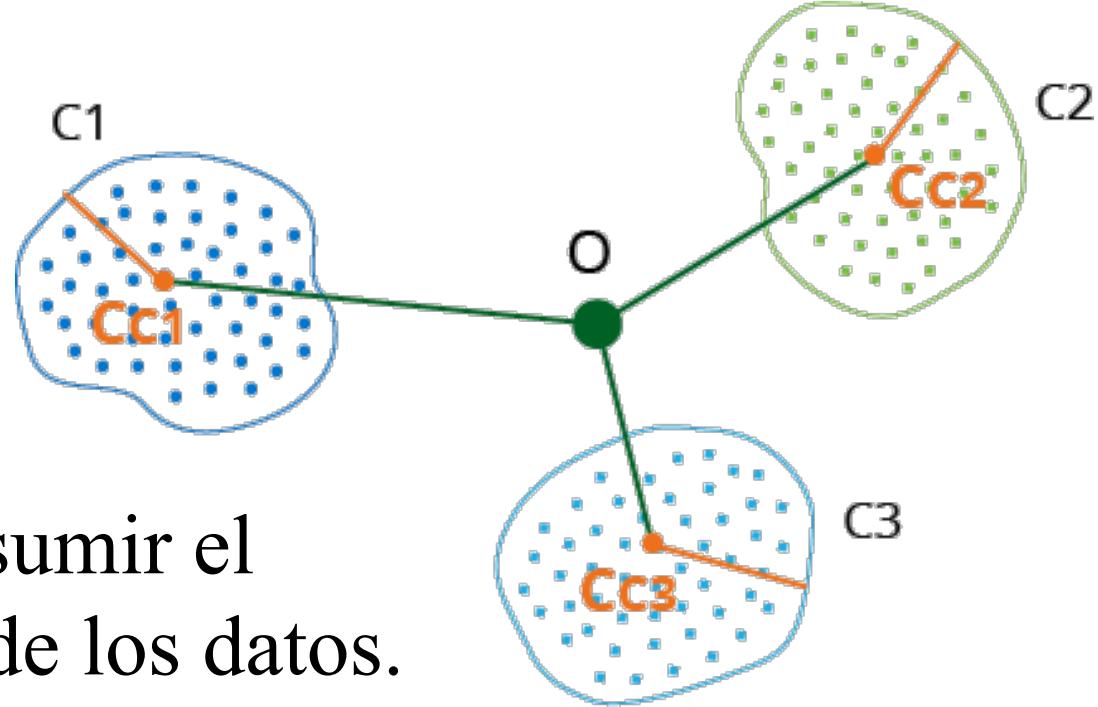


Introducción

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

Clustering

Comprender y resumir el comportamiento de los datos.



K-Means

Fácil de implementar y la interpretación de los resultados es sencilla.



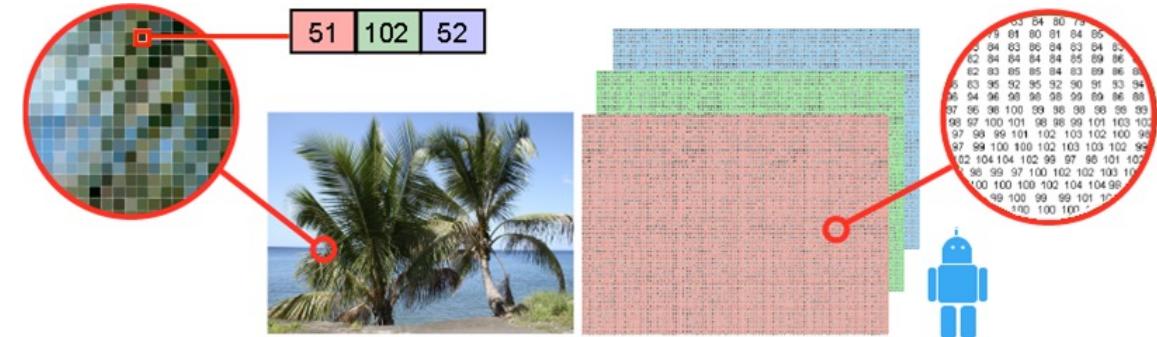
Introducción

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



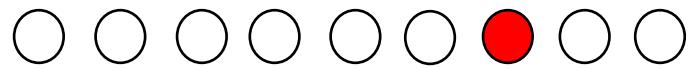
Segmentación de mercados

Aplicaciones del algoritmo



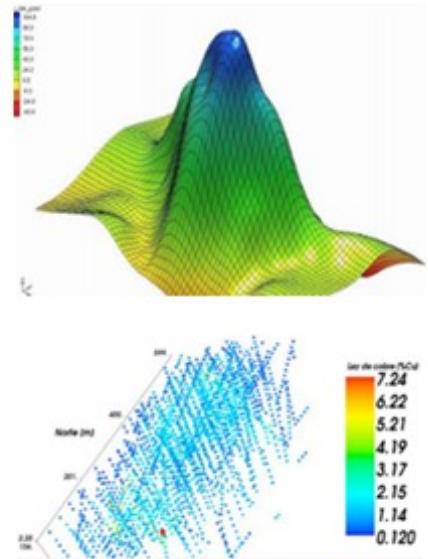
Visión por computadora

Contribuir en la mejora y
crecimiento de las empresas.

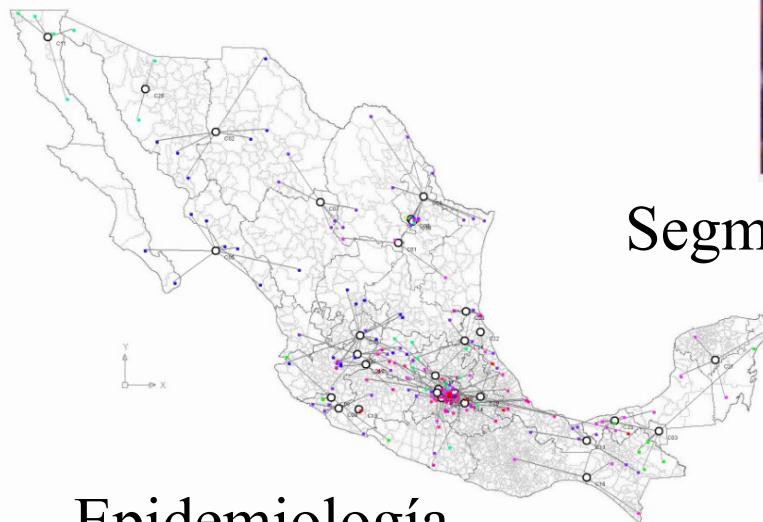


Introducción

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



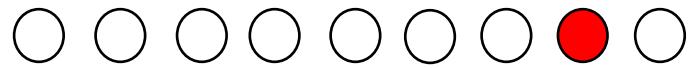
Geoestadística



Epidemiología



Segmentación de imágenes



Introducción

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

Complejidad del algoritmo K-Means

$$O(nkdi)$$

donde:

n = número de objetos

k = número de grupos

d = número de dimensiones

i = número de iteraciones

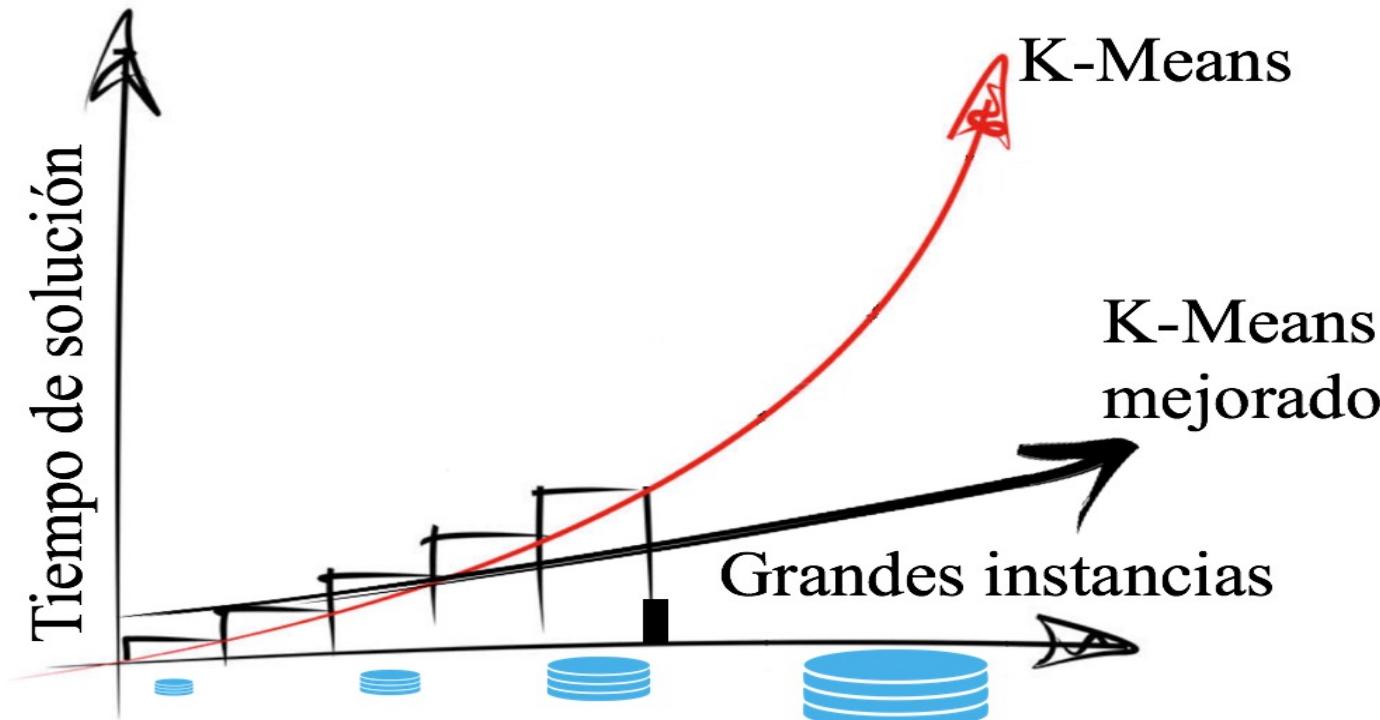




Introducción

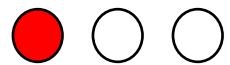
1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

Complejidad del algoritmo K-Means



Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



Algoritmo K-Means

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

El agrupamiento consiste en: Dado un conjunto $N = \{x_1, \dots, x_n\}$ denotado por n objetos con un número d de atributos, particionar el conjunto N en subconjuntos no vacíos del conjunto $K = \{1, \dots, k\}$ donde $k \geq 2$.

El objetivo de K-Means es minimizar la sumatoria del error al cuadrado de los K grupos formados mediante la expresión:

$$\sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$



1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



Algoritmo K-Means

Cantidad de grupos

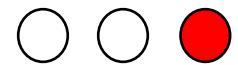
$$\sum_{k=1}^K \sum_{x_i \in C_k}$$

Centroide del grupo k

$$\|x_i - \mu_k\|^2$$

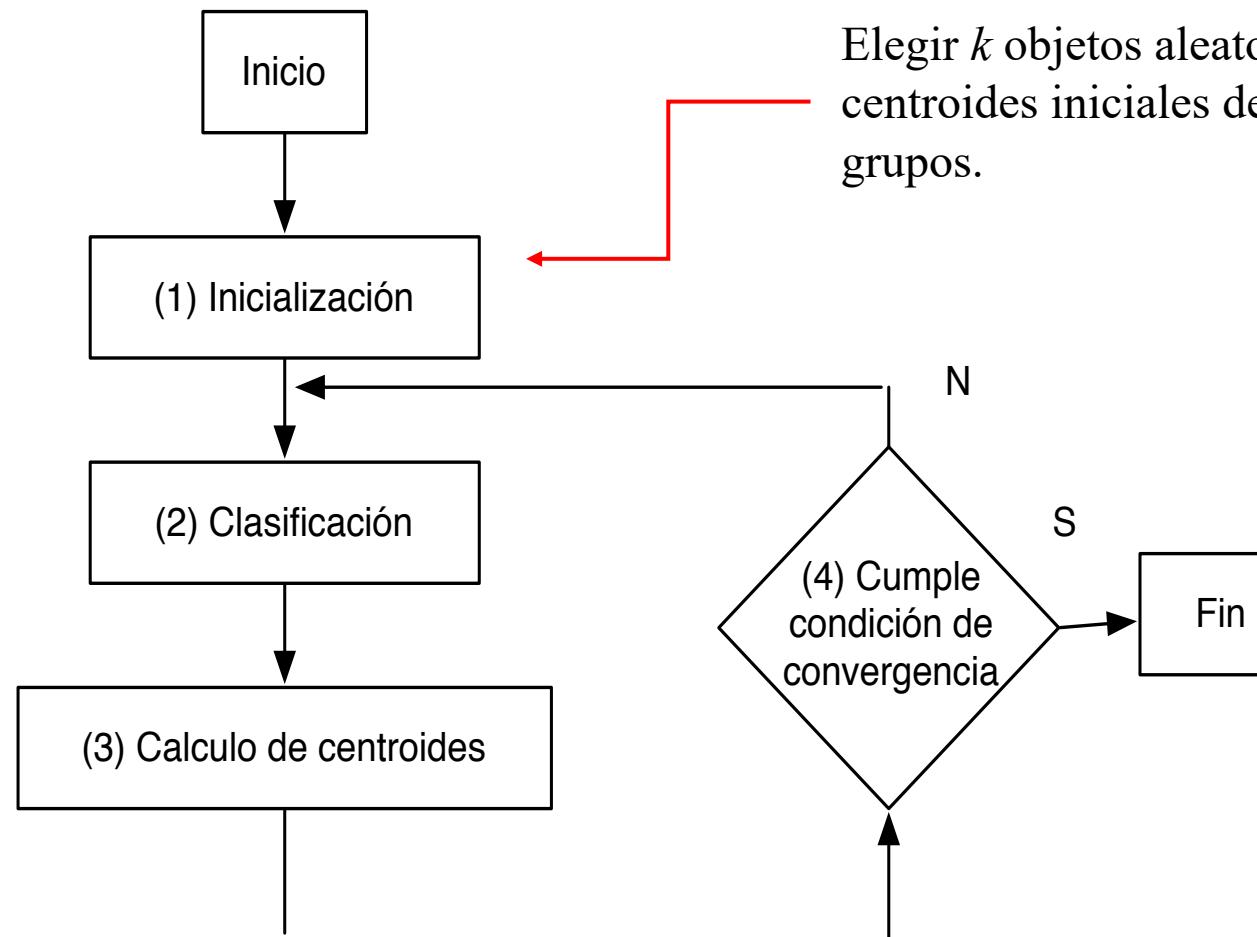
Para cada objeto i
asignado al grupo k

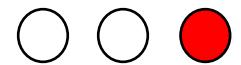
donde $\|\cdot\|$ denota
la distancia
Euclidiana



Algoritmo K-Means

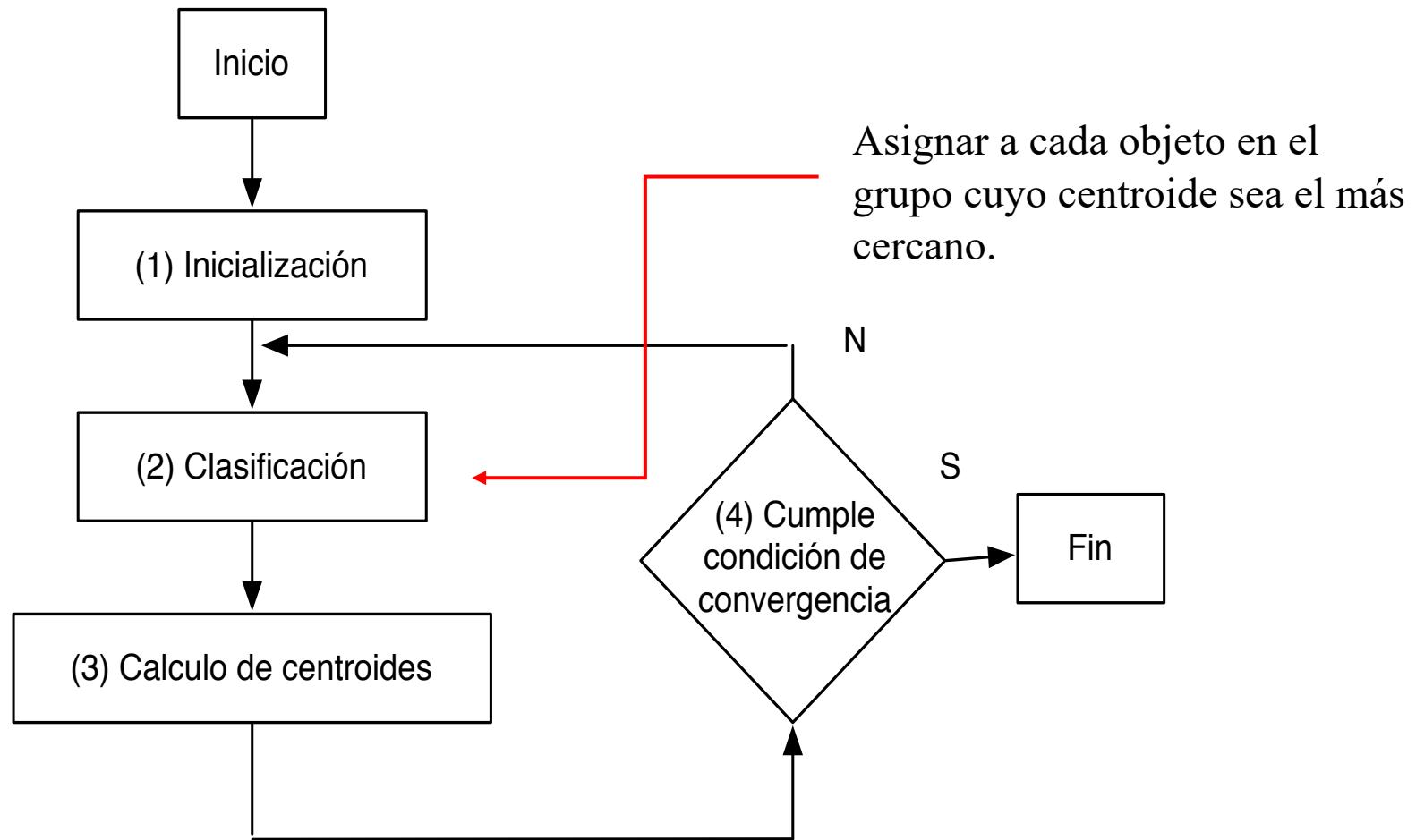
- 1. Introducción
(10 minutos)
- 2. Algoritmo K-Means.
(5 minutos)
- 3. K-Means en R.
(20 minutos)
- 4. Variantes de K-Means en R.
(10 minutos)
- 5. K-Means dinámico.
(5 minutos)
- 6. Receso.
(15 minutos)
- 7. Ejercicios con benckmark.
(45 minutos)
- 8. Comentarios finales.
(10 minutos)

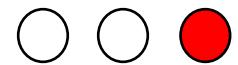




Algoritmo K-Means

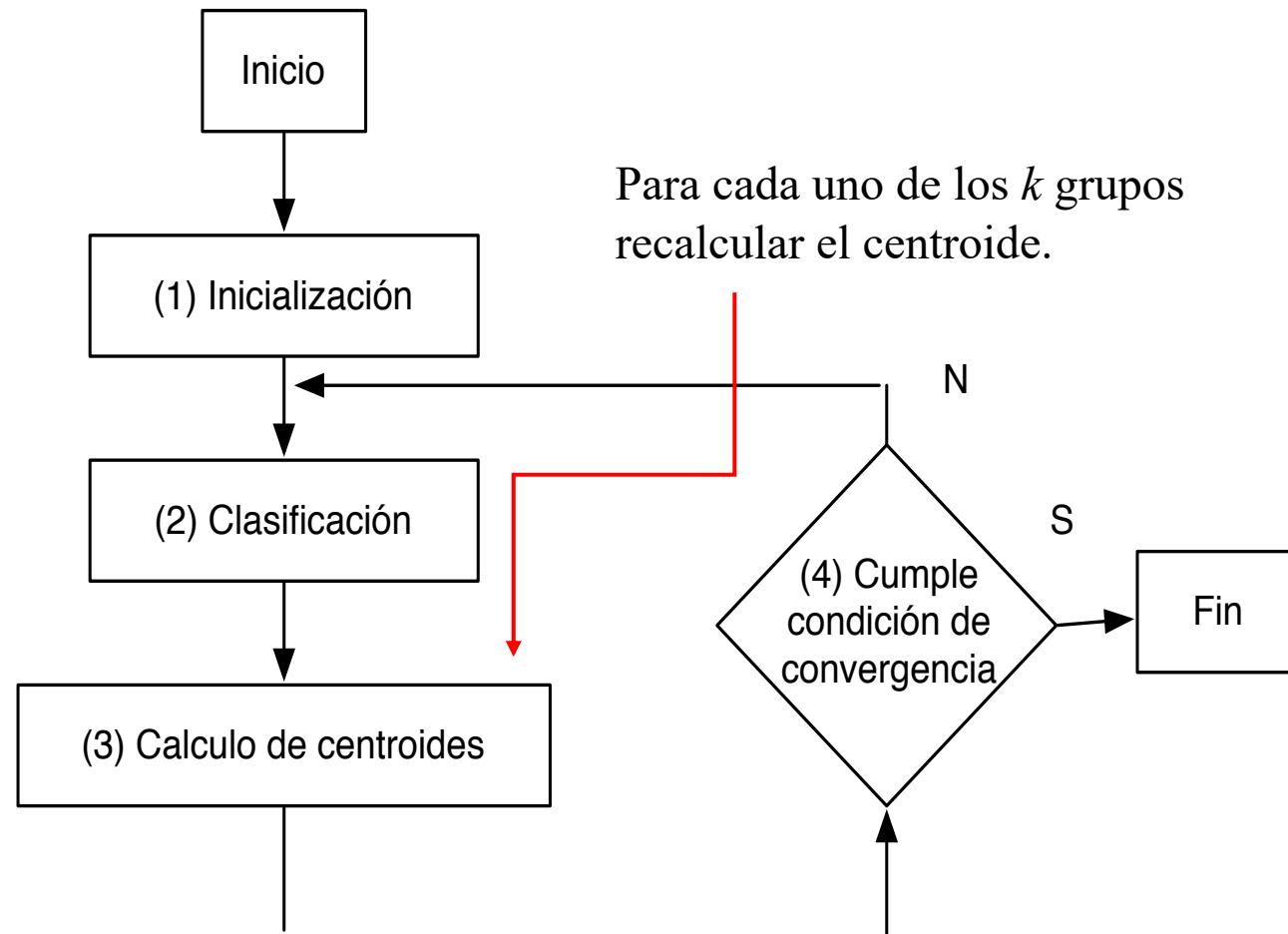
- 1. Introducción
(10 minutos)
- 2. Algoritmo K-Means.
(5 minutos)
- 3. K-Means en R.
(20 minutos)
- 4. Variantes de K-Means en R.
(10 minutos)
- 5. K-Means dinámico.
(5 minutos)
- 6. Receso.
(15 minutos)
- 7. Ejercicios con benckmark.
(45 minutos)
- 8. Comentarios finales.
(10 minutos)





Algoritmo K-Means

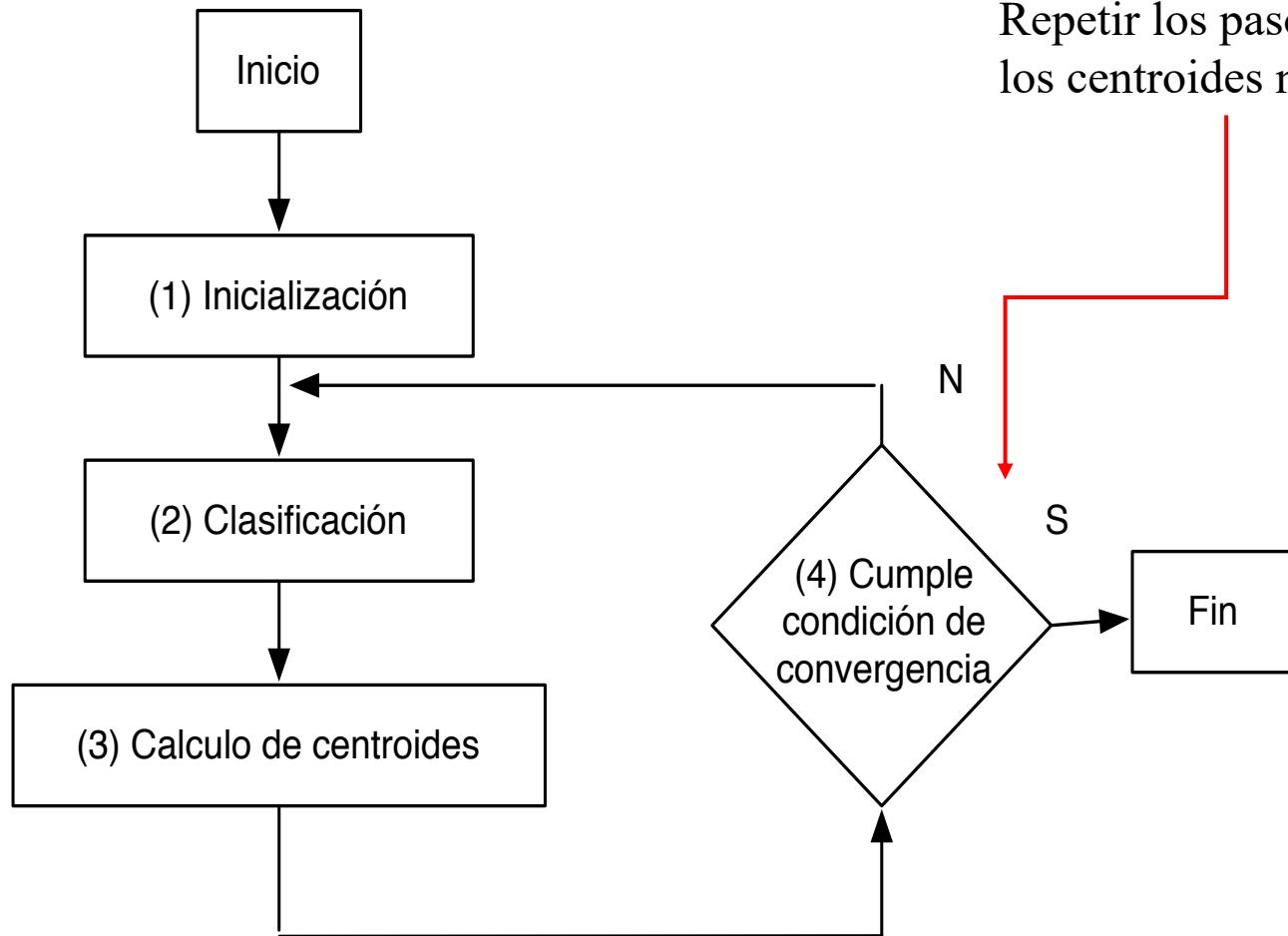
- 1. Introducción
(10 minutos)
- 2. Algoritmo K-Means.
(5 minutos)
- 3. K-Means en R.
(20 minutos)
- 4. Variantes de K-Means en R.
(10 minutos)
- 5. K-Means dinámico.
(5 minutos)
- 6. Receso.
(15 minutos)
- 7. Ejercicios con benckmark.
(45 minutos)
- 8. Comentarios finales.
(10 minutos)





1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

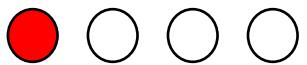
Algoritmo K-Means



Repetir los pasos 2 y 3 hasta que los centroides no cambien.

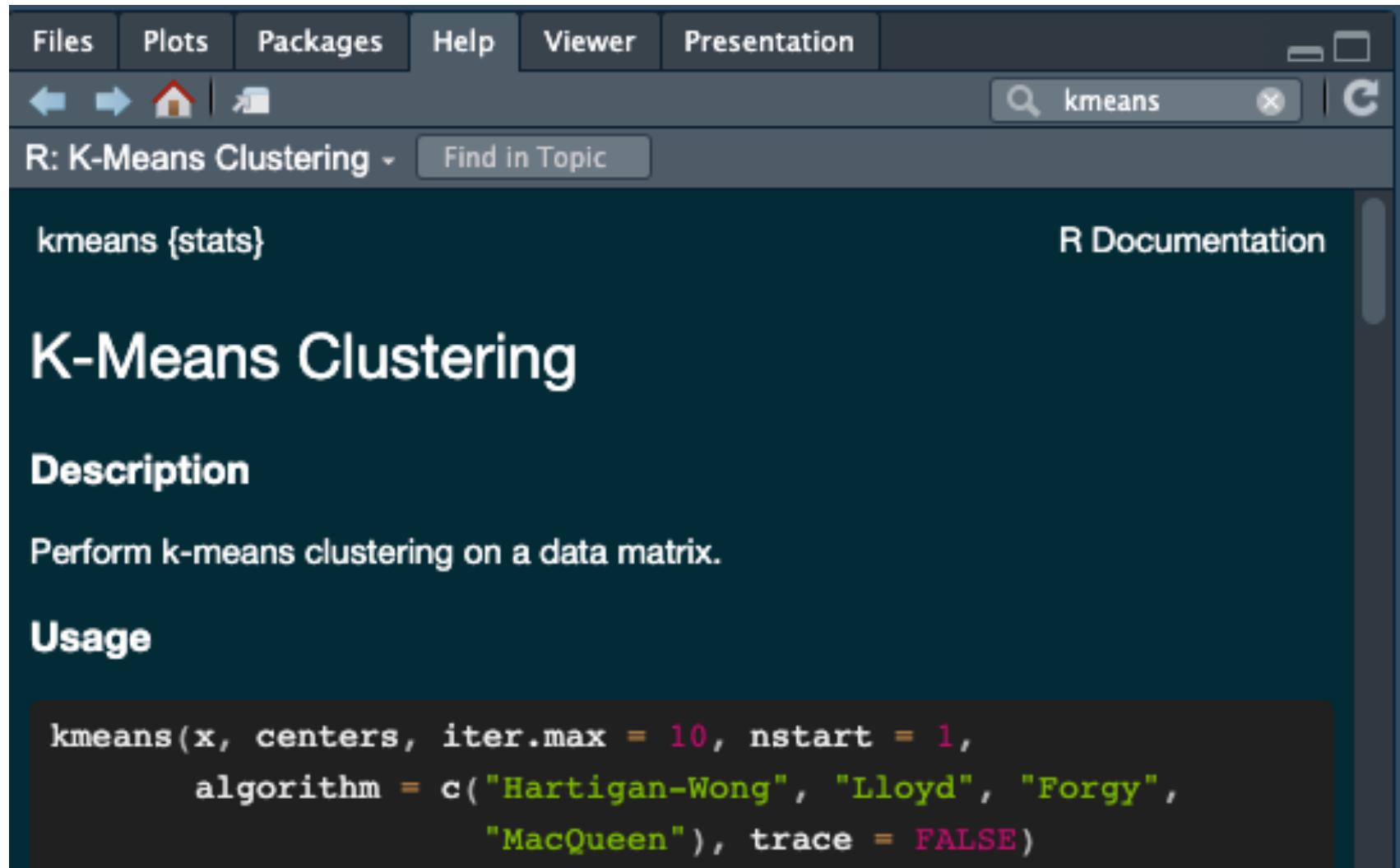
Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. **K-Means en R. 20 minutos**
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



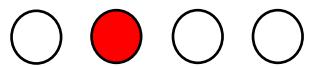
K-Means en R

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

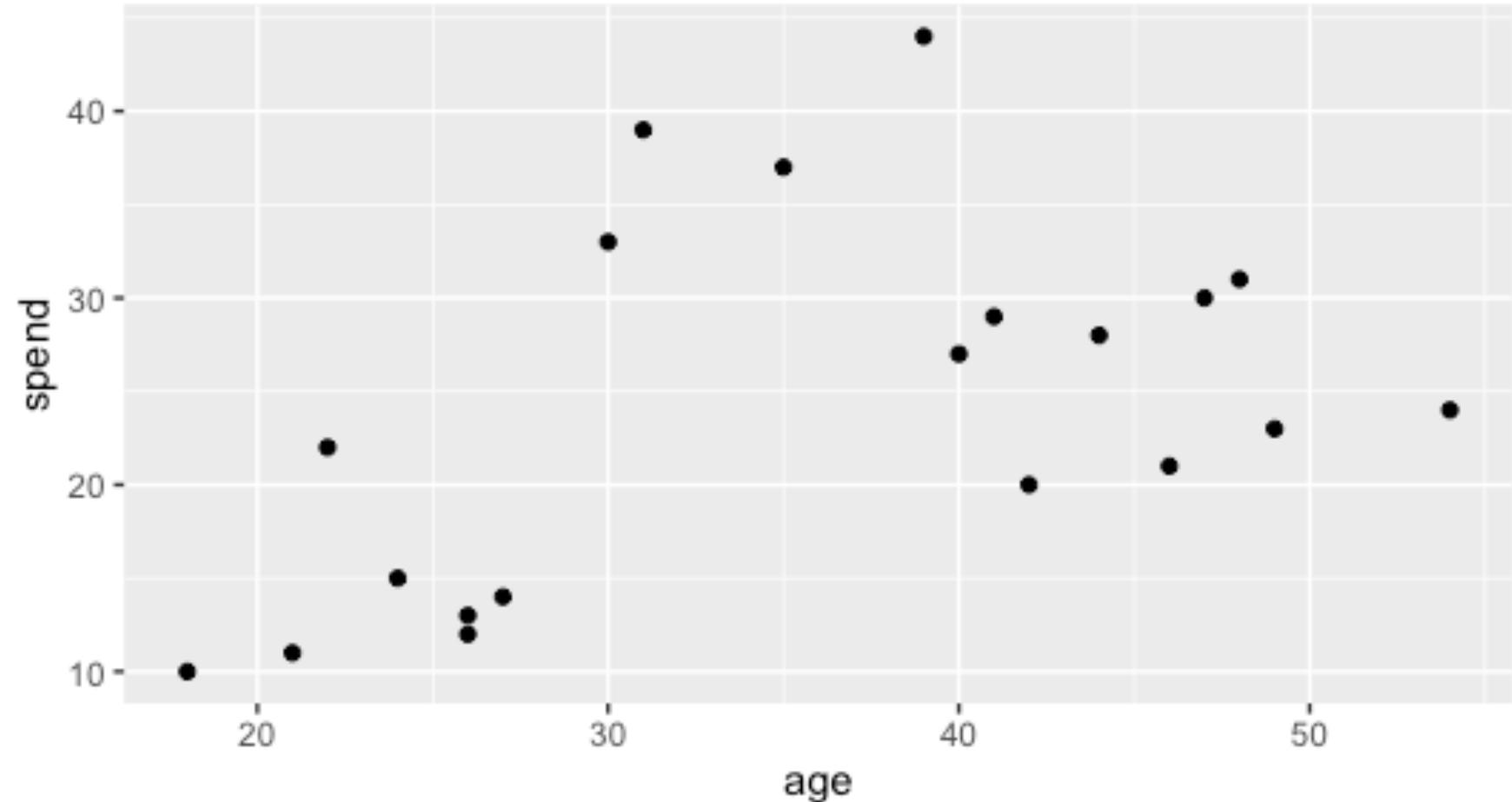


A screenshot of the R documentation for the `kmeans` function. The top navigation bar includes 'Files', 'Plots', 'Packages', 'Help', 'Viewer', and 'Presentation'. A search bar shows 'kmeans'. The main content area shows the title 'K-Means Clustering' and a 'Description' section stating 'Perform k-means clustering on a data matrix.' Below is a 'Usage' section with the R code:

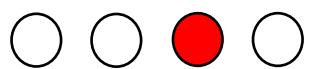
```
kmeans(x, centers, iter.max = 10, nstart = 1,  
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",  
                   "MacQueen"), trace = FALSE)
```



K-Means en R

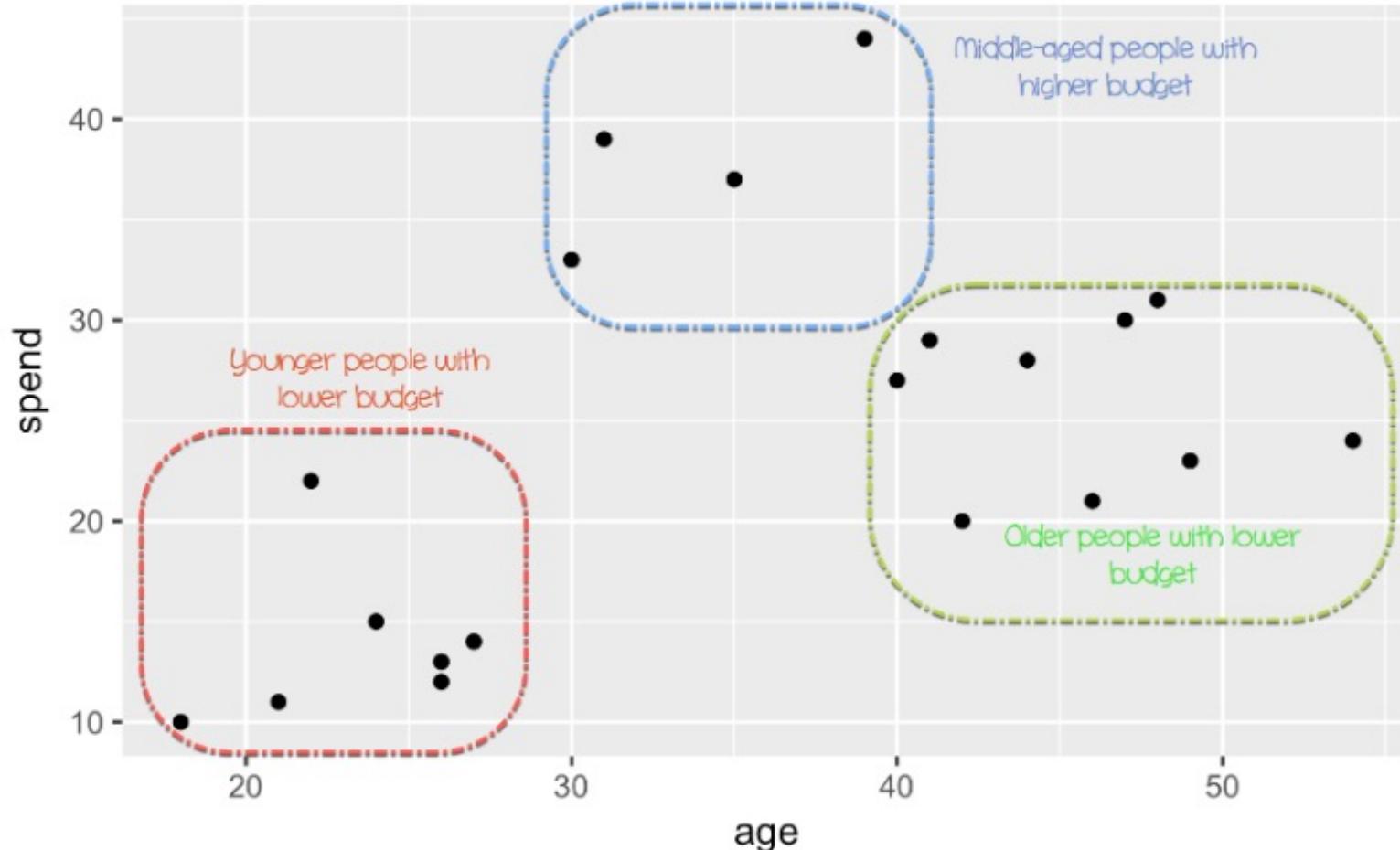


1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
- 3. K-Means en R.**
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



K-Means en R

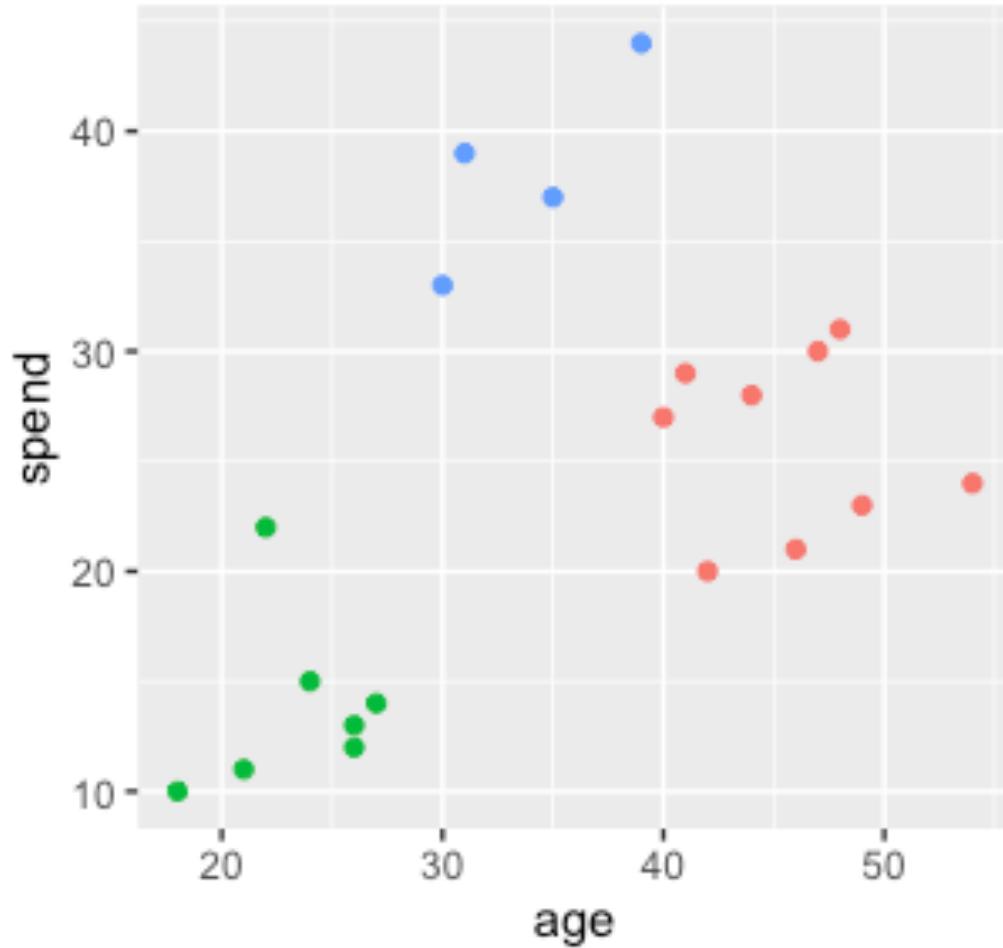
1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)





K-Means en R

- 1. Introducción
(10 minutos)
- 2. Algoritmo K-Means.
(5 minutos)
- 3. K-Means en R.**
(20 minutos)
- 4. Variantes de K-Means en R.
(10 minutos)
- 5. K-Means dinámico.
(5 minutos)
- 6. Receso.
- (15 minutos)
- 7. Ejercicios con benckmark.
(45 minutos)
- 8. Comentarios finales.
(10 minutos)

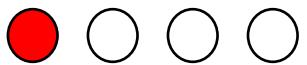


as.factor(out_kmeans\$cluster)



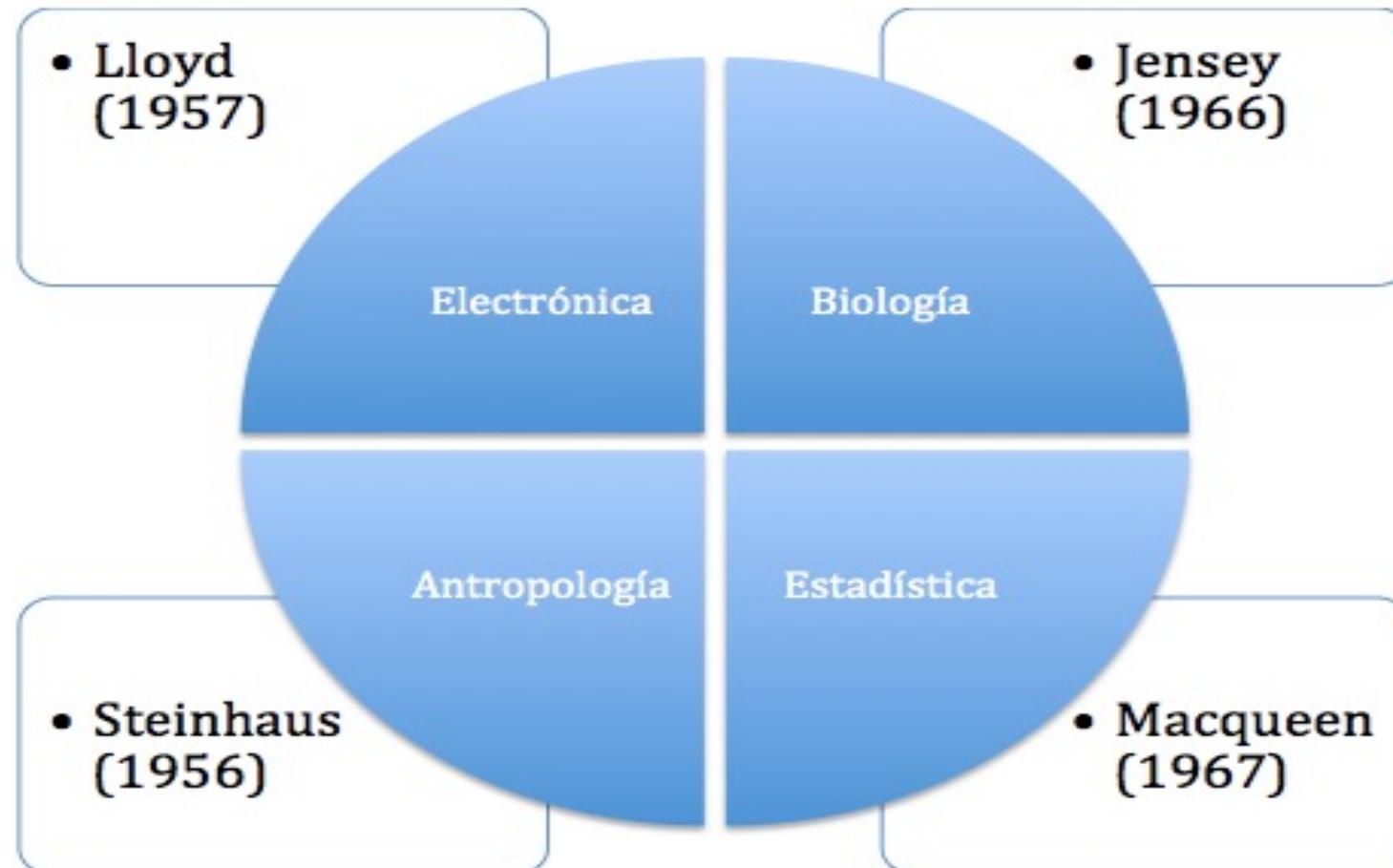
Agenda

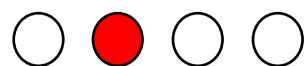
1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



Orígenes del algoritmo K-Means

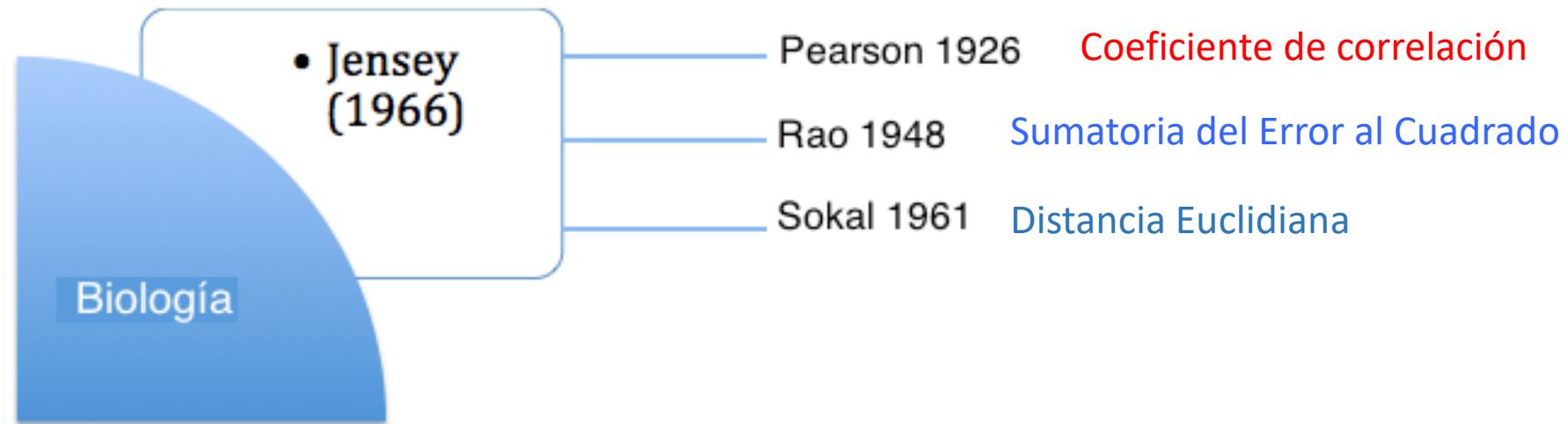
1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



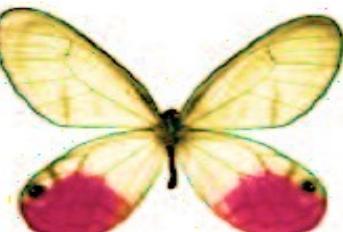


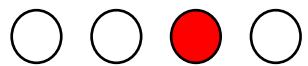
Orígenes del algoritmo K-Means

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



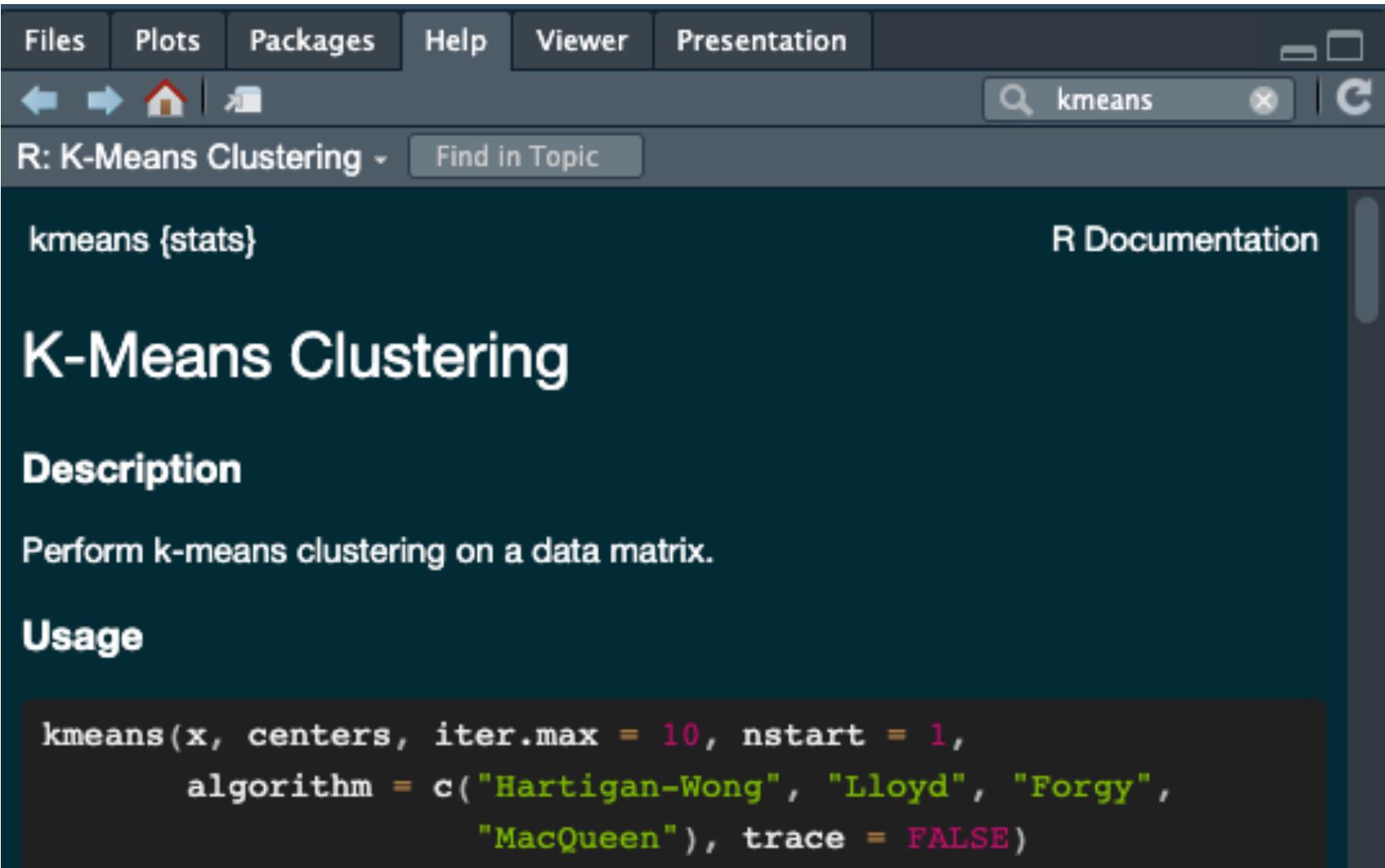
Taxonomy of genus *Phyllota* Benth. (Papilionaceae)



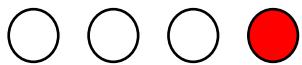


Variantes de K-Means en R

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



The screenshot shows the R Documentation interface. The top navigation bar includes 'Files', 'Plots', 'Packages', 'Help', 'Viewer', and 'Presentation'. A search bar contains the text 'kmeans'. The main content area displays the 'kmeans {stats}' topic under the heading 'R: K-Means Clustering'. It includes sections for 'Description' (Perform k-means clustering on a data matrix) and 'Usage' (A code block showing the function signature: `kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace = FALSE)`).



Variantes de K-Means en R

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
- 4. Variantes de K-Means en R.**
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

2.1.1 Lloyd

For Lloyd's algorithm (i.e., the basic K-means algorithm), let K be a set of k centroids, and for each centroid μ in K , let $G(\mu)$ denote its neighborhood (i.e., the set of data points for which μ is the nearest neighbor).

Each stage of Lloyd's algorithm moves each centroid μ to the centroid of $G(\mu)$ and then updates $G(\mu)$ by recalculating the distance from each object to its nearest centroid. These steps are repeated until convergence [18].

2.1.2 Forgy

This is essentially the basic K-means algorithm, except for the initialization of the centroids. This variant randomly selects k objects and uses these as the initial centroids [19].

2.1.3 MacQueen

This algorithm is fundamentally the same as the basic K-means. It adjusts all cluster centroids to the mean of their respective centroid μ each time an object x_i changes cluster membership $G(\mu)$ [20].

2.1.4 Hartigan-Wong

This algorithm assigns each object x_i to one of K groups or clusters to minimize the sum of squares within the cluster as shown by Eq. (1) [21]:

$$Sum(k) = \sum_{i=0}^n \sum_{j=0}^d (x(i,j) - x(k,j))^2, \quad (1)$$

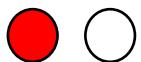
where $x(k, j)$ is the average of the objects belonging to the cluster. The main difference with respect to the basic K-means consists in the objective function: in this algorithm the objective function is Eq. (1), while in K-means it is the Euclidean distance [22].

Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. **K-Means dinámico. 5 minutos**
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



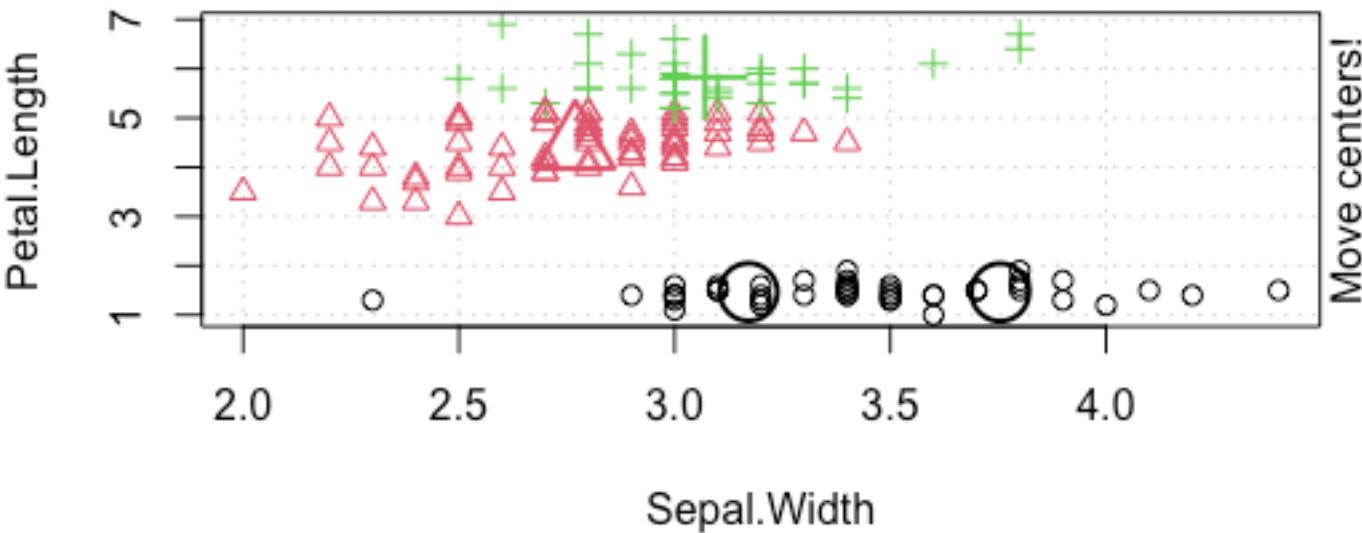
1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
- 5. K-Means dinámico.**
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

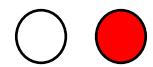


K-Means dinámico

```
## 6. K-Means dinámico

set.seed(2345)
library(animation)
iris
kmeans.ani(iris[2:3], 4)
```

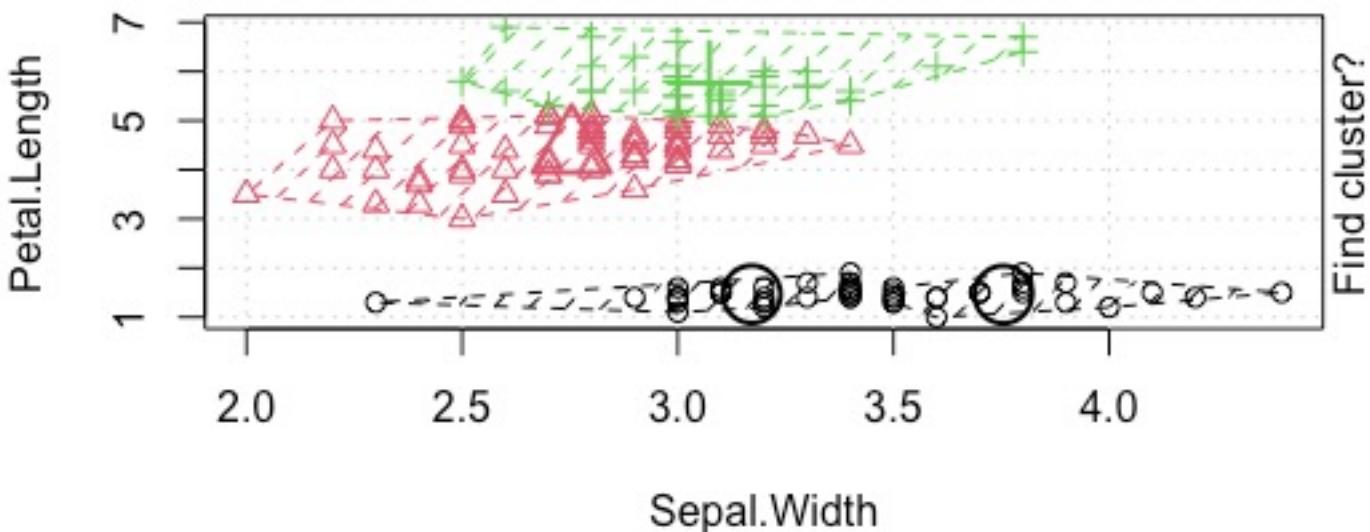




K-Means dinámico

```
## 6. K-Means dinámico  
  
set.seed(2345)  
library(animation)  
iris  
kmeans.ani(iris[2:3], 4)
```

1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
- 5. K-Means dinámico.**
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)



Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
7. Ejercicios con benckmark.
(45 minutos)
8. Comentarios finales.
(10 minutos)

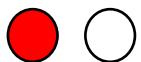


TAKE A LITTLE
COFFEE
BREAK



Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos



1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
- 7. Ejercicios con benckmark.**
(45 minutos)
8. Comentarios finales.
(10 minutos)

Clustering basic benchmark

Cite as:

P. Fänti and S. Sieranoja

K-means properties on six clustering benchmark datasets
Applied Intelligence, 48 (12), 4743-4759, December 2018

<https://cs.joensuu.fi/sipu/datasets/>

```
@misc{ClusteringDatasets,  
author = {Pasi Fr\"anti and Sami Sieranoja},  
title = {K-means properties on six clustering benchmark datasets},  
journal = {Applied Intelligence}  
year = {2018},  
volume = {48},  
number = {12},  
pages = {4743--4759},  
url = {http://cs.uef.fi/sipu/datasets/},  
}
```



1. Introducción
(10 minutos)
2. Algoritmo K-Means.
(5 minutos)
3. K-Means en R.
(20 minutos)
4. Variantes de K-Means en R.
(10 minutos)
5. K-Means dinámico.
(5 minutos)
6. Receso.
(15 minutos)
- 7. Ejercicios con benckmark.**
(45 minutos)
8. Comentarios finales.
(10 minutos)

Ejercicios con benckmark

1. Cargar un conjunto de datos

```
# variable <- read.csv("Nombre_del_archivo.csv")
```

Id	Read from a file
1	diamond9 <- read.csv("diamond9.csv")
2	longsquare <- read.csv("longsquare.csv")
3	spherical_4_3 <- read.csv("spherical_4_3.csv")
4	spherical_5_2 <- read.csv("spherical_5_2.csv")
5	triangle1 <- read.csv("triangle1.csv")

2. Particionar el conjunto de datos con el algoritmo K-Means

3. Visualizar el resultado de las particiones

Agenda

1. Introducción. 10 minutos
2. Algoritmo K-Means. 5 minutos
3. K-Means en R. 20 minutos
4. Variantes de K-Means en R. 10 minutos
5. K-Means dinámico. 5 minutos
6. Receso. 15 minutos
7. Ejercicios con benckmark. 45 minutos
8. Comentarios finales. 10 minutos

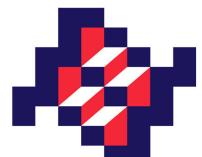


R-Ladies MX

Resultados con información descriptiva y predictiva de modelación epidemiológica, con patrones espacio temporales.



Universidade Federal
de Campina Grande



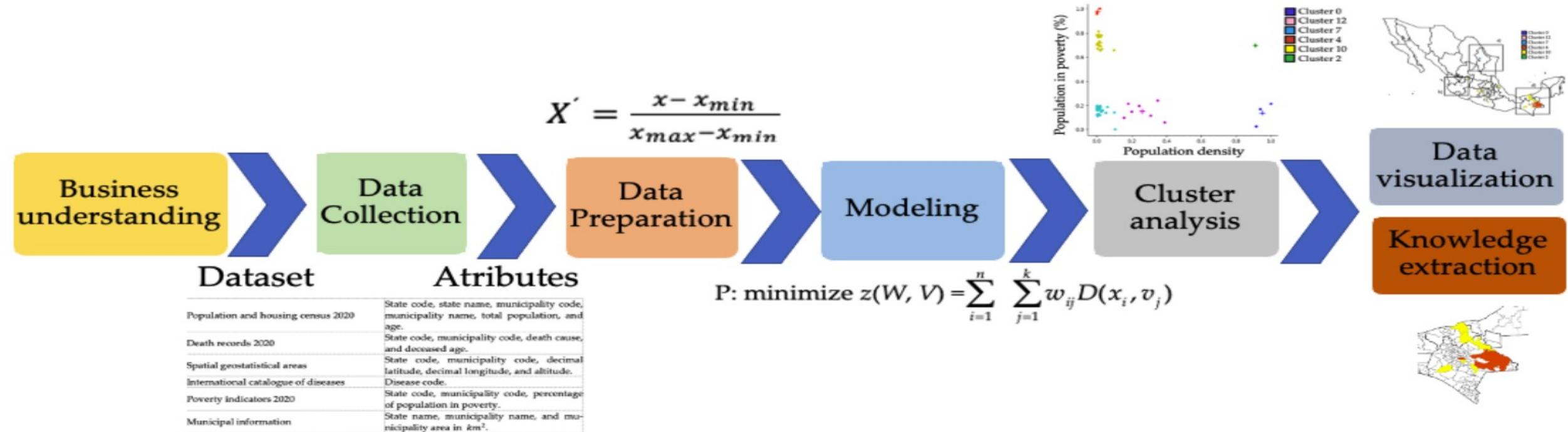
INSP



UNIVERSIDAD
COMPLUTENSE
MADRID

Resultados con información descriptiva y predictiva de modelación epidemiológica, con patrones espacio temporales.

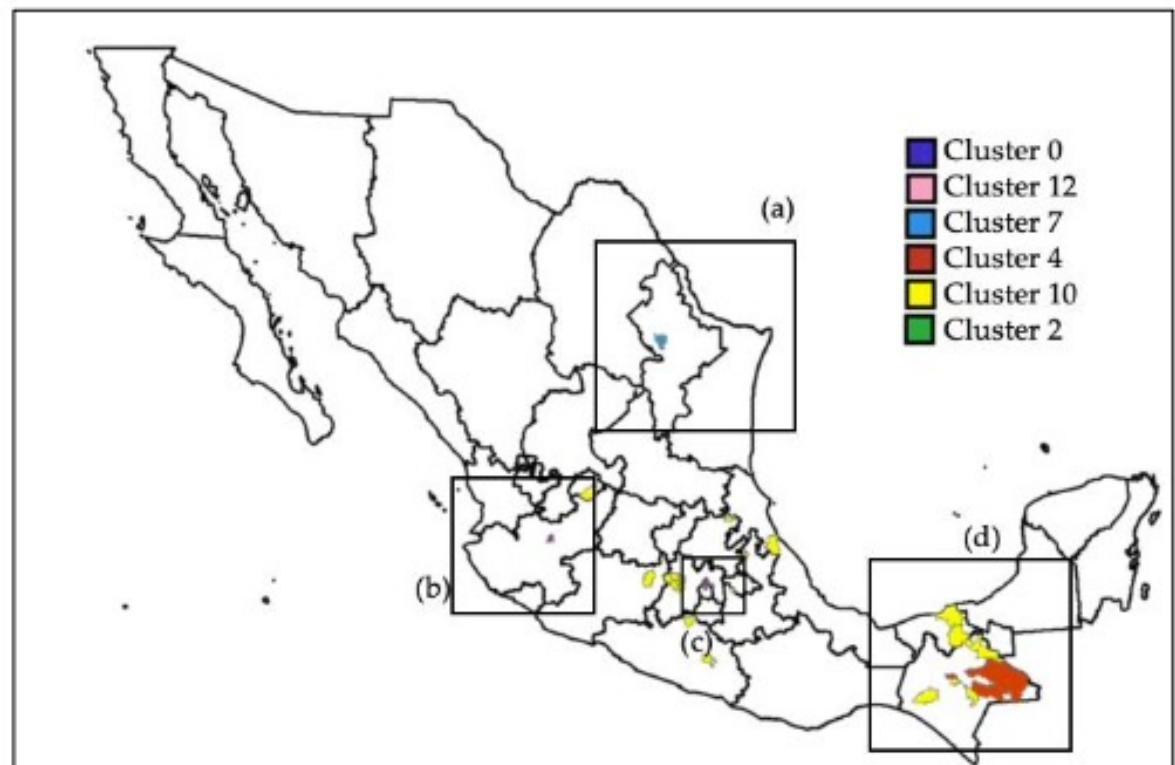
Mathematics, 2022. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico



Resultados con información descriptiva y predictiva de modelación epidemiológica, con patrones espacio temporales.

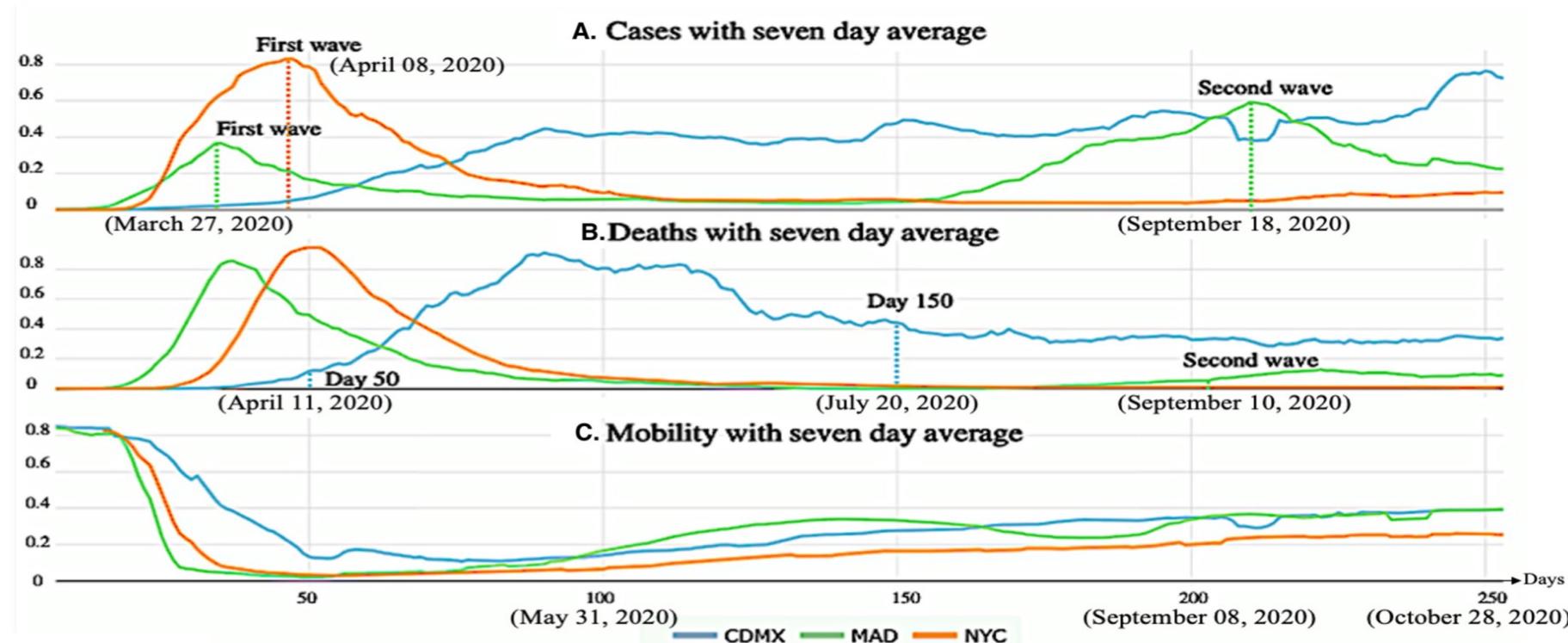
Mathematics, 2022. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico

Cluster	Population Density	% of Population in Poverty	Number of Municipalities
0	0.9138	0.0264	3
12	0.7223	0.2059	6
7	0.1995	0.1509	7
2	0.9108	0.6982	1
10	0.0089	0.7676	19
4	0.0009	0.9582	3



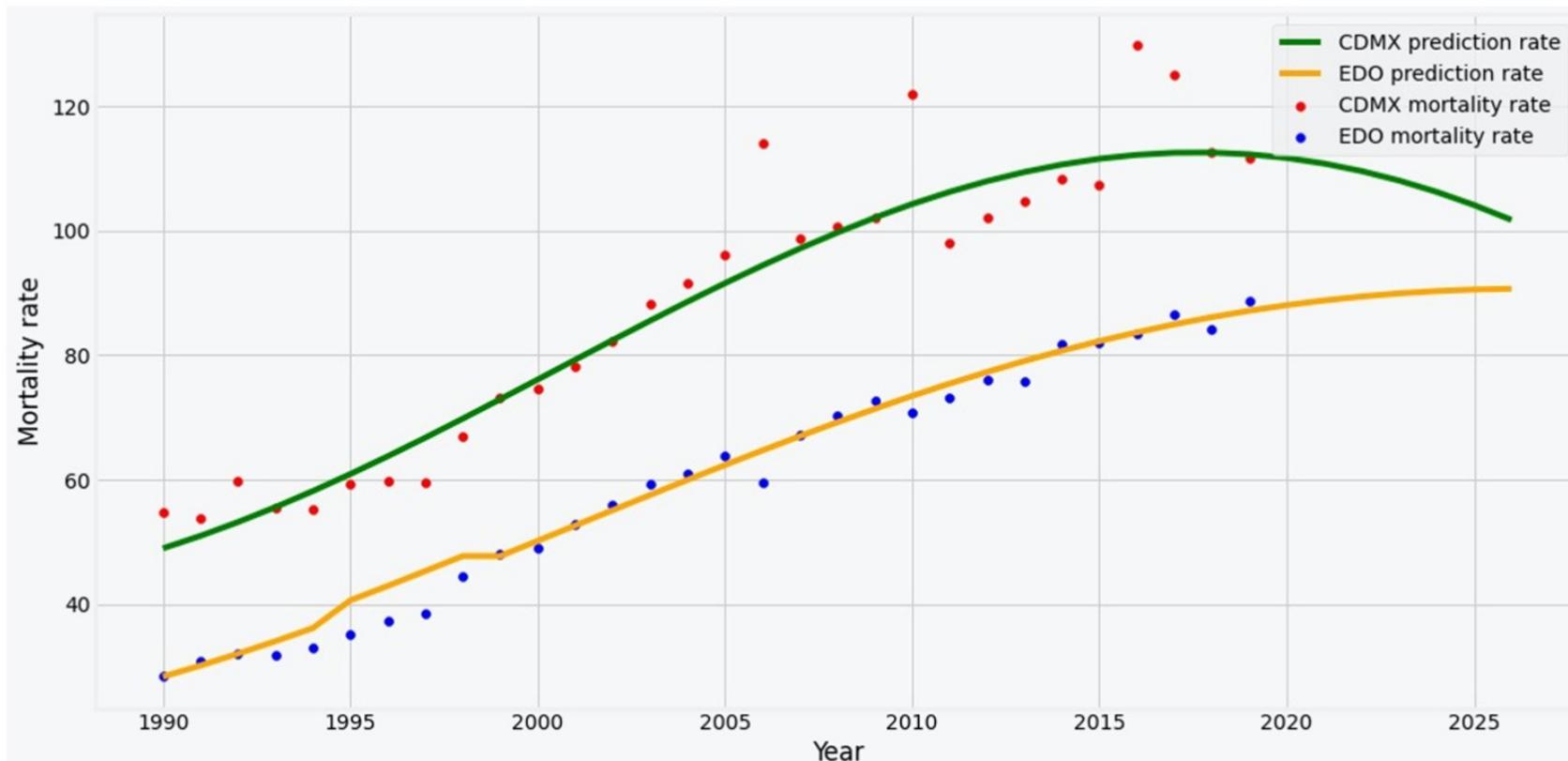
Resultados con información descriptiva y predictiva de modelación epidemiológica, con patrones espacio temporales.

Plos One, 2022. Correlation between mobility in mass transport and mortality due to COVID-19 A comparison of Mexico City, New York, and Madrid from a data science perspective.



Resultados con información descriptiva y predictiva de modelación epidemiológica, con patrones espacio temporales.

Springer LNCS, 2021 Prediction of Diabetes Mortality in Mexico City Applying Data Science





R-Ladies MX

muchas
gracias

A collage of hexagonal icons representing various R packages. The packages shown include: Shiny (blue), R Studio (light blue), gplot2 (grey), dplyr (orange), markdown (pink), knitr (red), lubridate (green), stringr (grey), devtools (light blue), tidyverse (orange), rmarkdown (yellow), and rlang (purple). The text 'muchas gracias' is overlaid on the bottom right of the collage.



**Dra. Nelva Nely
Almanza Ortega**

**Investigadora
Catedrática
Conahcyt**



R-Ladies MX



CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES
CIENCIAS Y TECNOLOGÍAS