Lecture 5:
Aprendizaje No Supervizado
Aprendizaje y Minería de Datos para los Negocios

Ignacio Sarmiento-Barbieri

Universidad de los Andes

October 29, 2021

Agenda

1. Tipos de Aprendizaje
2. Reducción de Dimensión: PCA
   - Aspectos Operativos
3. Clusters
   - Medidas de distancia o disimilaridad
   - K-Medias
4. Break
5. `R para ML`

Tipos de Aprendizaje

- ML se divide en dos (¿?) ramas principales:
  1. Aprendizaje supervisado: Tenemos datos tanto sobre un resultado y como sobre las variables explicativas X.
     - Esto es lo más cercano al análisis de regresión que conocemos.
     - Si y es discreto, también podemos ver esto como un problema de clasificación.

Tipos de Aprendizaje

- ML se divide en dos (¿?) ramas principales:
  1. Aprendizaje supervisado: Tenemos datos tanto sobre un resultado y como sobre las variables explicativas X.
     - Esto es lo más cercano al análisis de regresión que conocemos.
     - Si y es discreto, también podemos ver esto como un problema de clasificación.
  2. Aprendizaje no supervisado: No tenemos datos sobre y, solo sobre X.
  3. Lo utilizamos generalmente para 2 tareas:
     - Permite reducir la dimensionalidad y explorar datos
     - Agrupar datos

Exploración de Caracteristicas

Ejemplo: Casas

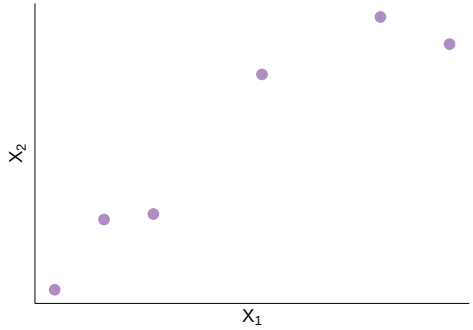
- Comencemos con un ejemplo simple
- Tenemos un grupo de 100 casas y características de estas casas
- Queremos hacer un análisis exploratorio y no sabemos donde comenzar
- Tenemos las siguientes variables

Size
Number of rooms
Number of bathrooms

Schools around
Crime rate

Exploración de Caracteristicas

- Supongamos que queremos visualizar n observaciones de las cuales tenemos k variables o atributos, representadas por $x_1, x_2, \ldots, x_k$ como parte de un análisis descriptivo.



Exploración de Caracteristicas

- Siguiendo el ejemplo anterior, n serian las 100 casas y k las 5 variables que miden caracteristicas de las casas.
- Una forma de hacer el análisis descriptivo es haciendo diagramas de dispersión para las n observaciones examinando dos variables por vez.
- El problema surge que vamos a tener

$$\left(\binom{5}{2} = \frac{5 \times 4}{2} = 10 \right) \tag{1}$$

- Si fuesen 10 tendriamos 45, y 20, 190!

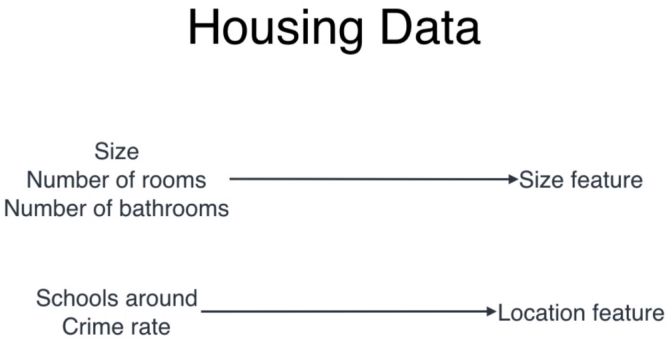
Reduccion de Dimension: PCA

- PCA: análisis de componentes principales
- Es una técnica de aprendizaje no supervisado que permite reducir la dimensionalidad de tales conjuntos de datos, aumentando la interpretabilidad, pero al mismo tiempo minimizando la pérdida de información.
- Intuitivamente, PCA plantea que cada observación se encuentra en un espacio $k - dimensional$, pero en el que no todas estas dimensiones son igualmente informativas.

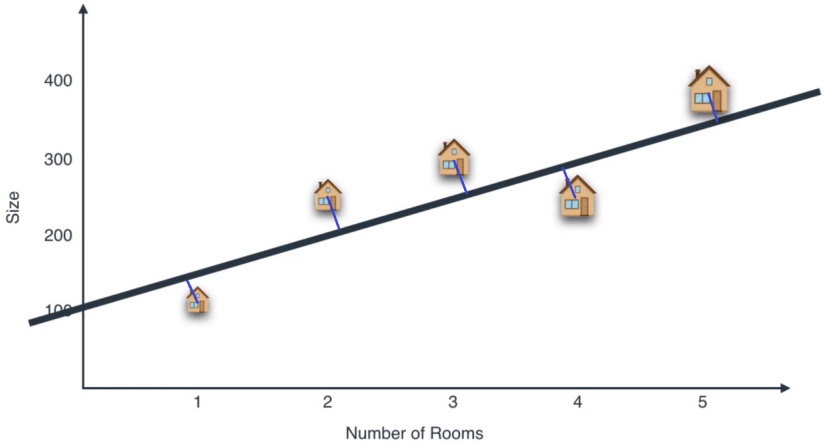
Reduccion de Dimension: PCA

- Por lo tanto, PCA busca representar los datos en un espacio de menor dimensión, reteniendo la mayor cantidad de información posible.
- Entonces estas nuevas dimensiones encontradas por PCA, llamadas componentes, es una combinación lineal de las variables originales.
- Se han desarrollado muchas técnicas para este propósito, pero el análisis de componentes principales es uno de los más antiguos y más utilizados.

Reduccion de Dimension: PCA



Reduccion de Dimension: PCA



Reduccion de Dimension: PCA

PCA

- El primer componente de un conjunto de variables $X_1, X_2, \ldots, X_k$ es

$$f_1 = \delta_{11}x_1 + \delta_{12}x_2 + \cdots + \delta_{1k}x_k \tag{2}$$

- f_1 denota al primer componente principal
- δ_{ij} se conocen como pesos o "loadings" del primer componente principal.
- Esta ecuación ilustra también el hecho de que el primer componente principal es una combinación lineal de las variables originales.

PCA

- El primer componente de un conjunto de variables $X_1, X_2, \ldots, X_k$ es

$$f_1 = \delta_{11}x_1 + \delta_{12}x_2 + \cdots + \delta_{1k}x_k \tag{2}$$

- f_1 denota al primer componente principal
- δ_{ij} se conocen como pesos o "loadings" del primer componente principal.
- Esta ecuación ilustra también el hecho de que el primer componente principal es una combinación lineal de las variables originales.
- La pregunta que naturalmente surge es: ¿Cómo se calculan estos componentes de forma tal que preserven la mayor información posible?

PCA

- La pregunta que naturalmente surge es: ¿Cómo se calculan estos componentes de forma tal que preserven la mayor información posible?
- Formalmente, supongamos que X es una matriz $n \times k$ que contiene los datos, es decir, las n observaciones de las k variables.
- Asumimos que cada una de las variables en X están centradas para tener media cero.
- La matriz X, a su vez, tiene una matriz de covarianza asociada denotada con $S = Var(X)$, que por definición es una matriz cuadrada de orden k.

PCA

- La tarea del primer componente principal es encontrar la combinación lineal de las variables originales que maximiza la varianza, es decir, preservan la mayor información posible.
- El objetivo es crear un índice f_1 que tiene la siguiente forma:

$$f_1 = X\delta_1' \tag{3}$$
$$= \delta_{11}x_1 + \delta_{12}x_2 + \cdots + \delta_{1k}x_k \tag{4}$$

- El problema consiste en elegir δ_1 óptimamente, ya que este indice va a ser la "mejor" combinación lineal de $x_1, x_2, \ldots, x_k$.
- Definimos como "mejor" a aquella combinación lineal que maximiza la varianza.

PCA

- Notando que

$$
\begin{aligned}
Var(f_1) &= Var(X\delta_1') && (5)\\
&= \delta_1 Var(X)\delta_1' && (6)\\
&= \delta_1 S\delta_1' && (7)
\end{aligned}
$$

- el problema se reduce a elegir δ_1 de forma que maximice $Var(X\delta_1)$.

PCA

- Notando que

$$
\begin{aligned}
Var(f_1) &= Var(X\delta_1') && (5)\\
&= \delta_1 Var(X)\delta_1' && (6)\\
&= \delta_1 S\delta_1' && (7)
\end{aligned}
$$

- el problema se reduce a elegir δ_1 de forma que maximice $Var(X\delta_1)$.
- Maximizar $\delta_1 S\delta_1'$ tiene como solución trivial llevar δ_1 a infinito $(\delta_1 \to \infty)$.

PCA

- Para que esta maximización tenga solución en la practica se impone una restricción adicional que normaliza δ_1:

$$\delta_1 \delta_1' = 1 \tag{8}$$

- Esto restringe a que la suma del cuadrado de los pesos o “loadings” sean igual a uno,

PCA

- El problema queda definido de la siguiente manera:

$$\max_{\delta_1} \ \delta_1 S \delta_1' \tag{9}$$

$$\text{sujeto a} \tag{10}$$

$$\delta_1 \delta_1' = 1 \tag{11}$$

- maximiza $\delta_1 S \delta_1'$ restringiendo que $\delta_1 \delta_1' = 1$. Escribiendo el Lagrangiano,

$$\mathcal{L} = \delta_1 S \delta_1' + \lambda_1 (1 - \delta_1 \delta_1') \tag{12}$$

PCA

- maximizamos esta expresión de la forma habitual derivando respecto a δ_1 e igualando a cero:

$$\frac{\partial \mathcal{L}}{\partial \delta_1} = S\delta_1' - \lambda_1 \delta_1' = 0 \tag{13}$$

- Reordenando:

$$S\delta_1' = \lambda_1 \delta_1' \tag{14}$$

- En el óptimo, δ_1 es el eigenvector correspondiente al eigenvalor λ. Premultiplicando la ecuación anterior por δ_1 y usando la restricción $\delta_1 \delta_1' = 1$:

$$\delta_1 S \delta_1' = \lambda_1 \tag{15}$$

PCA

$$\delta_1 S \delta_1' = \lambda_1 \tag{16}$$

- Para maximizar $\delta_1 S \delta_1'$ debemos elegir λ_1 igual al máximo eigenvalor de S y δ_1 igual al eigenvalor correspondiente.
- Notando además que $\delta_1 S \delta_1' = Var(X\delta_1')$, el problema de encontrar la mejor combinación lineal que reproduce la variabilidad en X se reduce a encontrar el mayor eigenvalor de S y su correspondiente eigenvector.

PCA

- Luego de calcular el primer componente principal f_1,
- podemos encontrar también el segundo componente principal, f_2:

$$f_2 = X\delta_2' \tag{17}$$
$$= \delta_{21}x_1 + \delta_{22}x_2 + \cdots + \delta_{2k}x_k \tag{18}$$

- El segundo componente principal, será la combinación lineal que tiene la máxima varianza de todas las combinaciones lineales ortogonales a f_1.
- En otras palabras, responde a la pregunta: ¿Cuál es la mejor combinación lineal de las variables $x_1, x_2, \ldots, x_k$ no correlacionada al primer componente principal?
- Intuitivamente, esta es la "segunda mejor" combinación lineal de $x_1, x_2, \ldots, x_k$, que no esta contenida en el primer componente.

PCA

El cálculo del segundo componente entonces responde al siguiente problema:

$$\max_{\delta_2} \ \delta_2 S \delta_2' \tag{19}$$

$$\text{sujeto a} \tag{20}$$

$$\delta_2 \delta_2' = 1 \tag{21}$$

$$y \tag{22}$$

$$\delta_2 \delta_1' = 0 \tag{23}$$

PCA

- La solución es el eigenvector asociado al segundo eigenvalor mas grande.
- Siguiendo esta lógica, por inducción es posible seguir calculando componentes cada uno ortogonal entre si y decrecientes en importancia.
- En general, para una matriz X con n observaciones y k variables tiene al menos el mínimo entre el número de observaciones menos 1 $(n-1)$ y el número de variables (k), componentes principales distintos.

$$\#PC = (min(n-1, k) \tag{24}$$

Escalar las variables

- Desde un punto estrictamente matemático no hay nada intrínsecamente incorrecto en hacer combinaciones lineales de variables con diferentes unidades de medida.
- Sin embargo, cuando usamos PCA buscamos maximizar varianza y la varianza se ve afectada por las unidades de medida.
- Esto implica que los componentes principales basados en la matriz de covarianza S van a cambiar si las unidades de medida de una o más variables cambian.
- Para que esto no suceda, es práctica habitual estandarizar las variables. Es decir, cada valor de X es recentrado y dividido por la desviación estándar:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{25}$$

Descomposición espectral

- La descomposición espectral o eigendecomposición es una forma de descomponer matrices.
- Descomponer una matriz significa que queremos encontrar un producto de matrices que sea igual a la matriz inicial.
- En el caso de la eigendecomposición, descomponemos la matriz inicial en el producto de sus eigenvectores y eigenvalores.

Descomposición espectral

- Veamos cómo se usan los eigenvectores y eigenvalores para descomponer una matriz.
- Tomando los eigenvectores de una matriz $A_{m\times m}$ podemos concatenarlos y colocarlos en una matriz P.
- Entonces P será una matriz cuyas columnas son los eigenvectores de A:

$$A = P\Lambda P' \tag{26}$$

- donde $\Lambda = diag(\lambda)$ es una matriz diagonal de los eigenvalores.
- Es importante notar que esta descomposición sólo es válida para matrices cuadradas, como lo es la matriz de covarianza. Por lo tanto, no puede ser utilizada para matrices no cuadradas.

Proporción de Varianza Explicada

- Una propiedad muy útil del procedimiento del cálculo de componentes principales es que la variabilidad total de X es la suma de los k eigenvalores de $V(X) = S$.
- Para demostrarlo supongamos que $\lambda_1, \ldots, \lambda_k$ son los eigenvalores de $V(X) = S$, ordenados de mayor a menor,
- $p_1, \ldots, p_K$ son los eigenvectores correspondientes.
- Adicionalmente llamemos P a la matriz cuyas columnas son estos eigenvectores.

Proporción de Varianza Explicada

- Supongamos también que $f_j = X\delta_j'$ es el $j - esimo$ componente principal, entonces

$$V(f_j) = \delta_j S \delta_j' \tag{27}$$
$$= p_j P \Lambda P' p_j' \tag{28}$$
$$= \lambda_j \tag{29}$$

- Este resultado nos dices que la varianza del $j - esimo$ componente principal es el $j - esimo$ eigenvalor ordenado de S.
- Usando este resultado podemos ver que la varianza total de X va a

Proporción de Varianza Explicada

- Entonces

$$tr(\Sigma) = tr(P\Lambda P') = tr(PP'\Lambda) = \sum_{j=1}^{k} \lambda_j = \sum_{j=1}^{k} V(f_j) \tag{30}$$

- Este resultado nos permite preguntarnos: ¿Cuánta información perdimos por proyectar los datos en unos cuantos componentes principales?
- ¿Cuánto de la varianza esta contenida en los primeros componentes principales?

$$\frac{\lambda_k}{\sum_{j=1}^{k} \lambda_j} \tag{31}$$

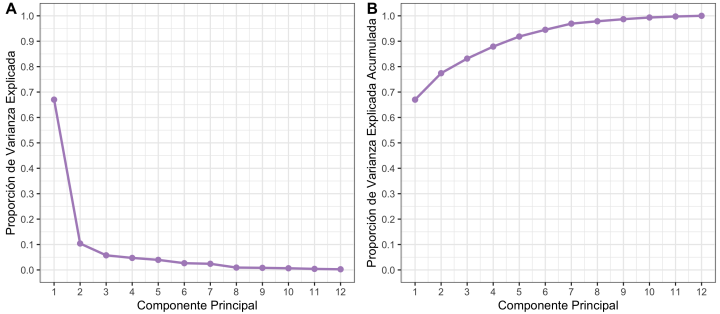
¿Cuántos componentes principales debemos seleccionar?

- Mencionamos que una matriz X de dimensión $(n \times k)$ tiene en general $min(n-1, k)$ componentes principales distintos.
- En la práctica generalmente no estamos interesados en todos los componentes, sino más bien quedarnos con los primeros que nos permitan visualizar o interpretar datos.
- En efecto, nos gustaría quedarnos con el mínimo numero que nos permita una buena comprensión de los datos.

¿Cuántos componentes principales debemos seleccionar?

- La pregunta natural que surge aquí es si hay una forma establecida para determinar el número de componentes principales a utilizar.
- Desafortunadamente, no existe una forma objetiva aceptada en la literatura de responderla. Sin embargo, hay tres enfoques simples que pueden servir de guía para decidir el número de componentes principales relevantes.
  Estos son:
  - Examen visual de un gráfico de sedimentación (scree plot).
  - Proporción de varianza explicada.
  - Criterio de Kaiser.

Examen visual de un gráfico de sedimentación (screeplot)



Proporción de varianza explicada

- Otro enfoque a menudo utilizado en la práctica, es imponer un umbral a priori y elegir los componentes principales en base a esta.
- Por ejemplo podríamos definir un umbral del 90%, o 70%
- El umbral a definir dependerá de la aplicación, el contexto, y el conjunto de los datos.

- Típicamente se utilizan umbrales entre el 70% y el 90%.

Criterio de Kaiser

- El criterio de Kaiser es otro enfoque ampliamente utilizado para evaluar el número máximo de componentes principales.
- Este sugiere que solo se retengan los componentes principales cuyos eigenvalores sean mayores a 1.
- La idea es que se retengan aquellos componentes cuyos eigenvalues sean superiores a la media de los eigenvalues:

$$\lambda_h > \frac{\sum_j^k \lambda_j}{k} \tag{32}$$

- Dado que los datos están estandarizados tenemos que $\sum_j^k \lambda_j = k$, por lo que es equivalente a buscar los eigenvalues mayores a uno.

Clusters

Análisis de Clusters

- El análisis de clústeres (también llamado segmentación de datos) es una de las principales aplicaciones de los algoritmos de aprendizaje no supervisado
- consiste dividir las observaciones de un conjunto en un número m de grupos de tal manera que todos los puntos dentro de un mismo grupo estén más estrechamente relacionados entre sí que los puntos asignados a diferentes grupos.
- un concepto central en todo análisis de clústeres es la noción de similaridad o disimilaridad entre observaciones la cual podrá variar según las decisiones del investigador.
- Existen algunas guías para escoger una adecuada medida de distancia, sin embargo, existe un buen grado de subjetividad

Matrices de distancia o disimilaridad

- Muchas veces los datos son presentados directamente en términos de proximidad entre pares de objetos.
- Este tipo de datos generalmente se representa en una matriz D de tamaño $N \times N$ en donde N es el número de objetos y cada elemento d_{ij} corresponde a la proximidad entre el objeto i y el objeto j.
- En muchos casos esta matriz se usa como insumo en los algoritmos de clustering.
- La mayoría de algoritmos suponen que la matriz de disimilaridad debe ser no negativa y con ceros en la diagonal (es decir que la distancia de un objeto a si mismo es 0).

Distancia

- Entre las medidas de distancia más comunes se encuentra la distancia euclideana:

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2 \tag{33}$$

- Muy útil para variables continuas
- Para atributos no cuantitativos (como las variables categóricas) la distancia al cuadrado no es apropiada.

Distancia

- Para variables cuantitativas es usual definir el error entre ellas como una función monótona creciente del valor absoluto de su diferencia:

$$d(x_i, x_{i'}) = l(|x_i, x_{i'}|) \tag{34}$$

- Para variables ordinales: Este tipo de variables usualmente se representa como un continuo de enteros y el dominio de estos es considerado un conjunto ordenado.

  $\frac{i-1/2}{M}$, $i = 1, \cdots, M$

- En el orden prescrito de sus valores originales. Luego se tratan como variables cuantitativas en esta escala.

K-Medias

K-Medias

- K-Medias es quizas el algoritmo mas conocido para hacer agrupamiento
- Es un algoritmo de descenso iterativo.
- Este se usa cuando todas las variables en el conjunto de datos son numéricas y se utiliza la distancia euclidiana cuadrática para definir la disimilaridad entre observaciones:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2 \tag{35}$$

K-Medias

- En este caso particular la función de pérdida a minimizar para garantizar que los puntos más cercanos entre sí estén dentro de un mismo segmento es:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}||x_i - x_{i'}||^2 \tag{36}$$

$$= \sum_{k=1}^{K} N_k \sum_{C(i)=k}||x_i - \bar{x_k}||^2 \tag{37}$$

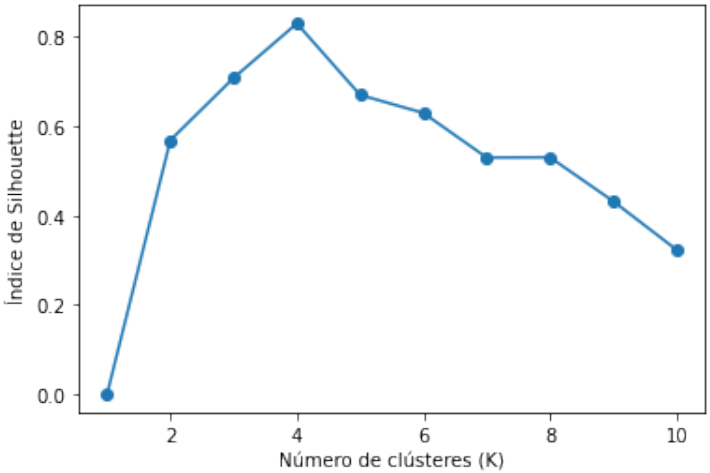
- En donde $\bar{x_k} = (\bar{x_{1k}}, \cdots, \bar{x_{pk}})$ es el vector de medias asociado al clúster k-ésimo y $N_k = \sum_{i=1}^{N}(C(i) = k)$.

K-Medias

- Para minimizar la expresión pasada se usa el siguiente algoritmo:
  1. Para una determinada asignación de clusters K, la varianza total de los conglomerados se minimiza con respecto a $\{m_1, \cdots, m_K\}$ obteniendo las medias de los clusters asignados actualmente $\bar{x_s} = argmin_m \sum_{i \in S} ||x_i - m||^2$.
  2. Dado un conjunto actual de medias $\{m_1, \cdots, m_K\}$, la varianza total de los conglomerados se minimiza asignando cada observación a la media del conglomerado más cercana. Es decir,
  $$C(i) = argmin_{1 \leq k \leq K} ||x_i - m_k||^2$$
  3. Se repiten los pasos 1 y 2 hasta que las asignaciones no cambien.

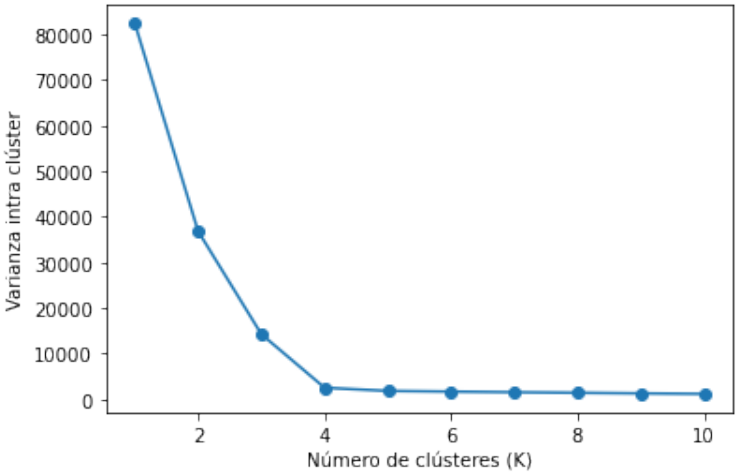
Aspectos operativos

Gráfica de Codo



Aspectos operativos

Coeficiente de Silhouette

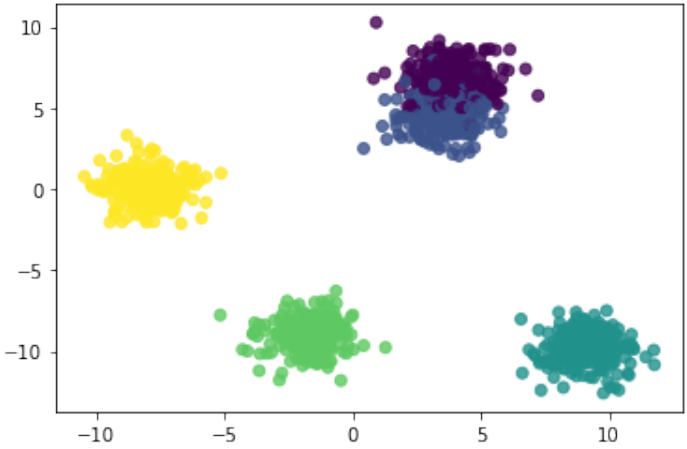


Aspectos operativos

- Cada uno de los pasos 1 y 2 reduce el valor de la varianza total de los conglomerados, por lo que la convergencia está asegurada.
- Sin embargo, el resultado puede representar un mínimo local subóptimo.
- Por ende se recomienda iniciar el algoritmo con varias opciones aleatorias diferentes para las medias iniciales y elegir la solución que tenga el valor más pequeño de la función objetivo.

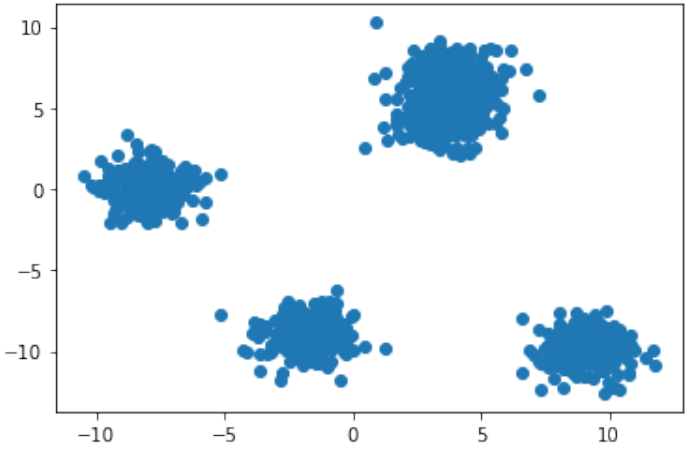
Ejemplo

Simulación



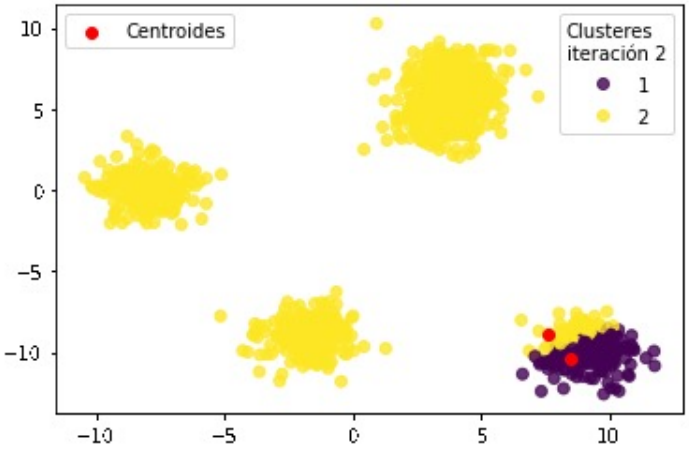
Ejemplo

Observamos



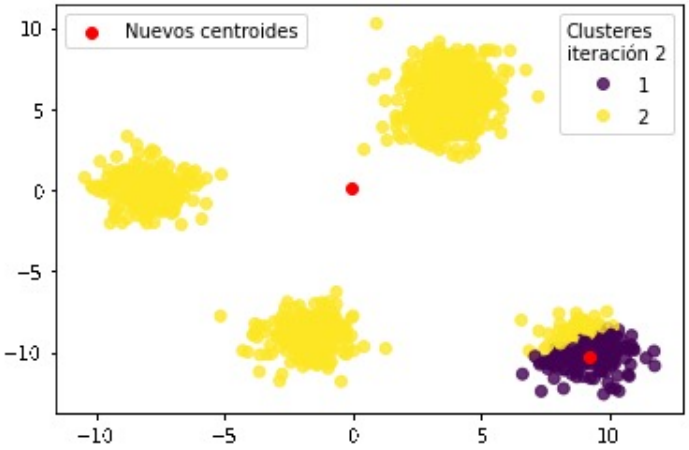
Ejemplo

Proceso K=2



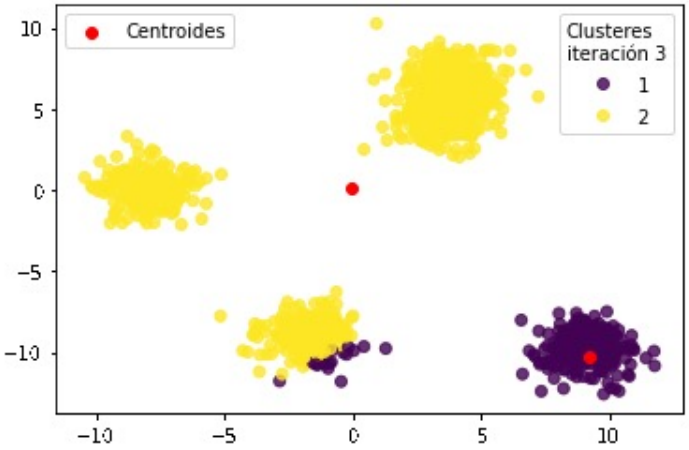
Ejemplo

Proceso K=2



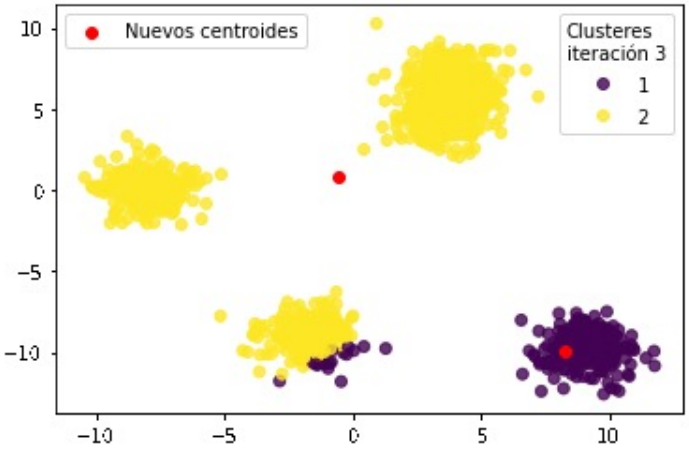
Ejemplo

Proceso K=2



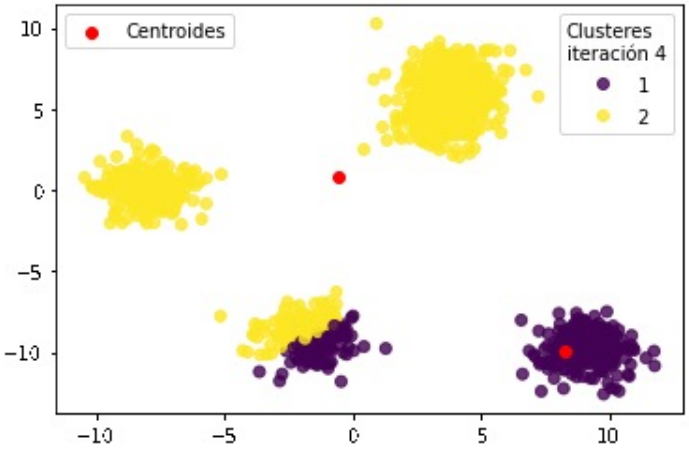
Ejemplo

Proceso K=2



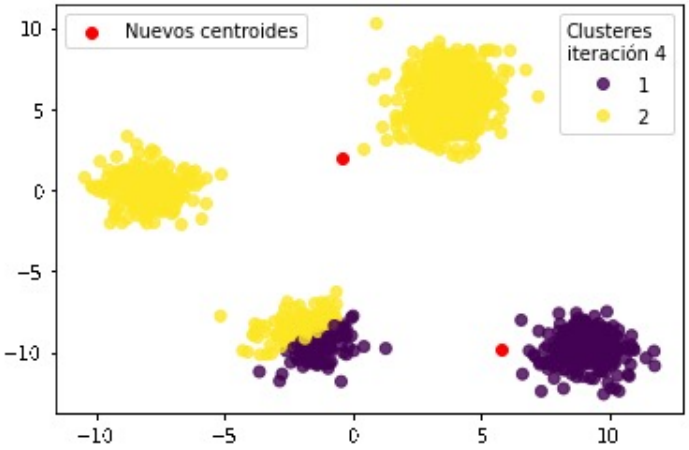
Ejemplo

Proceso K=2



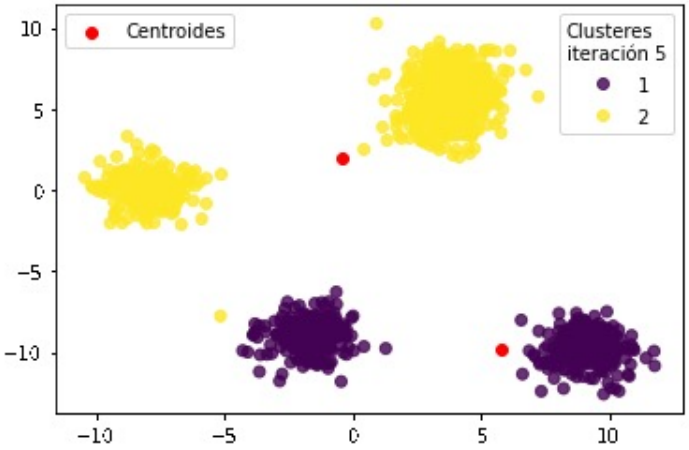
Ejemplo

Proceso K=2



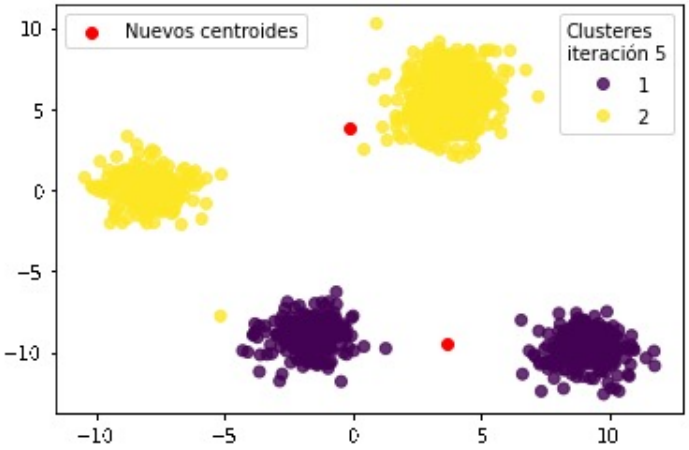
Ejemplo

Proceso K=2



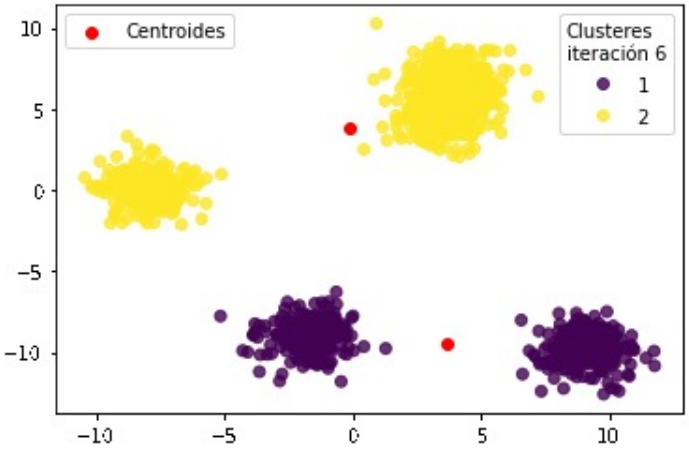
Ejemplo

Proceso K=2



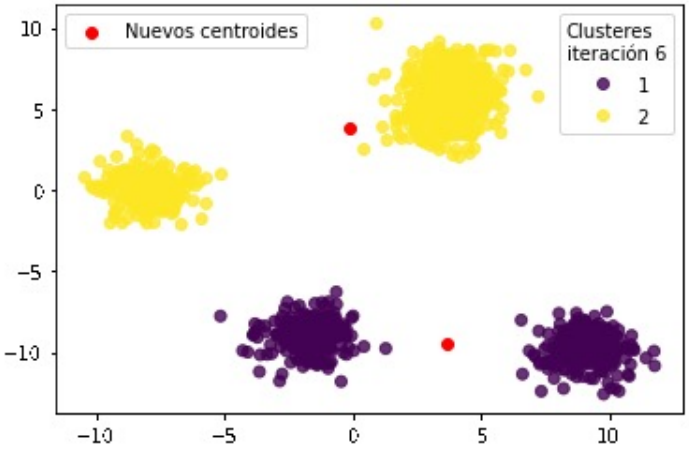
Ejemplo

Proceso K=2



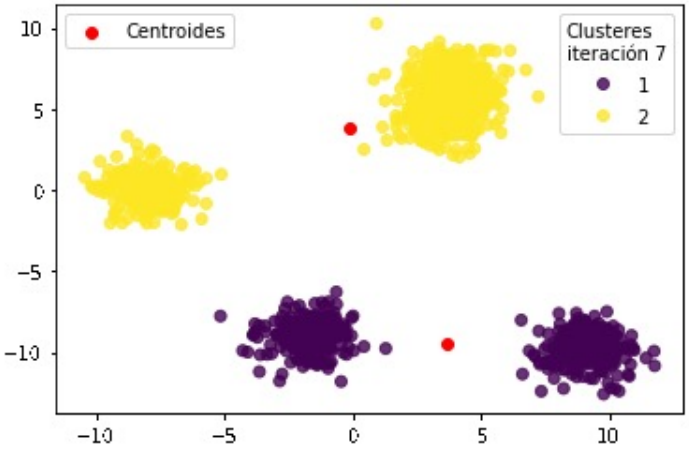
Ejemplo

Proceso K=2



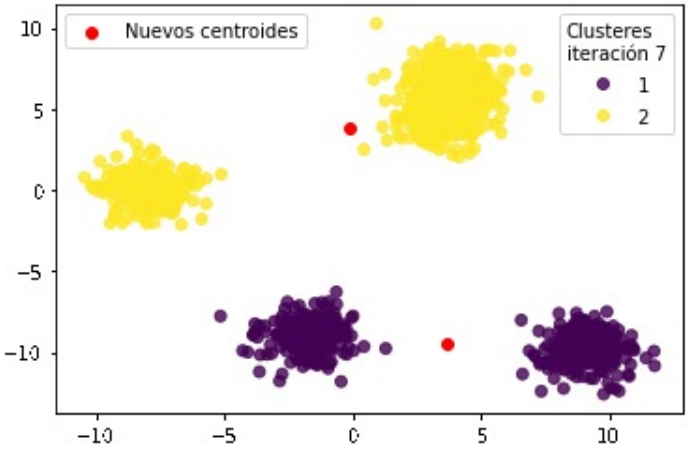
Ejemplo

Proceso K=2



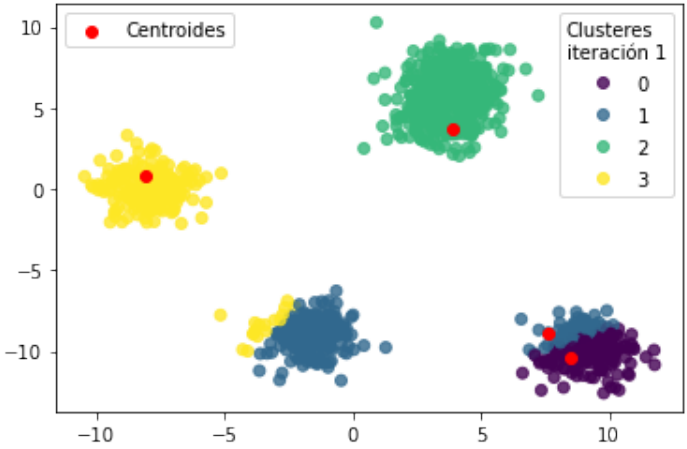
Ejemplo

Proceso K=2

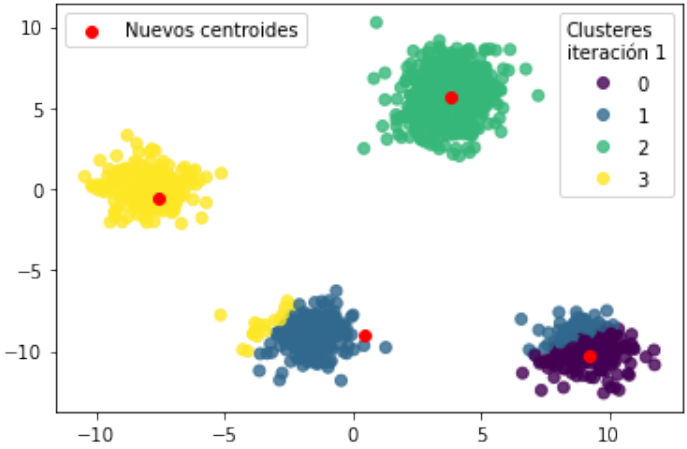


Ejemplo

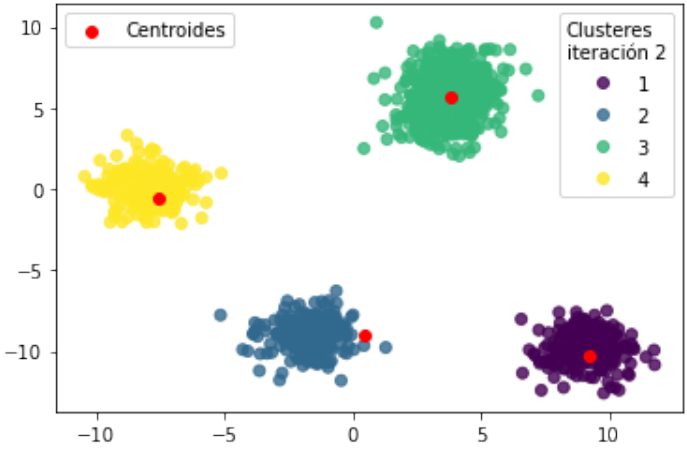
Proceso K=4



Ejemplo
Motivación

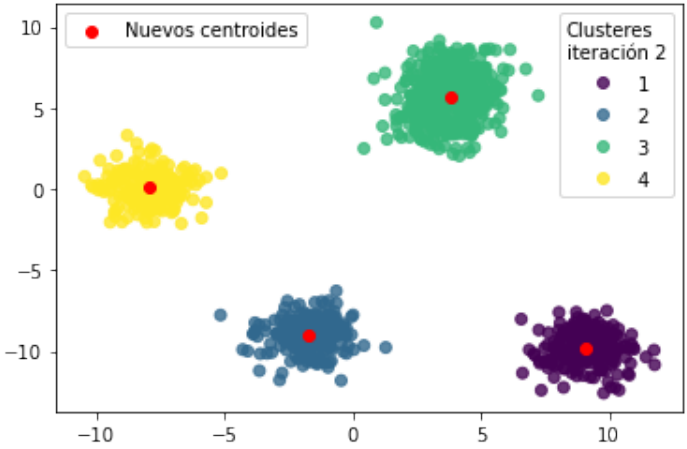


Ejemplo
Motivación

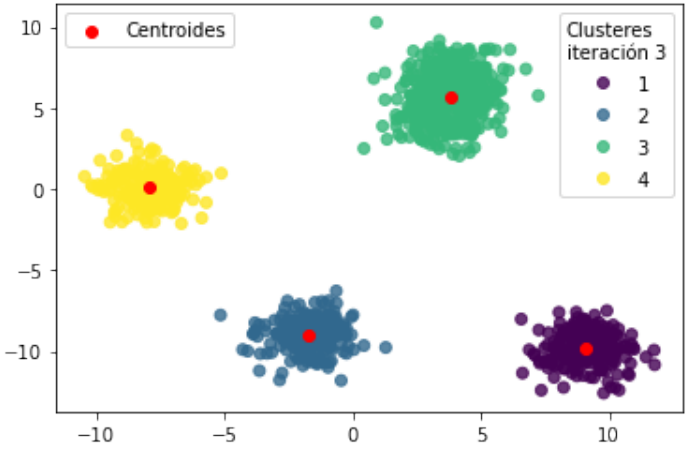


Ejemplo

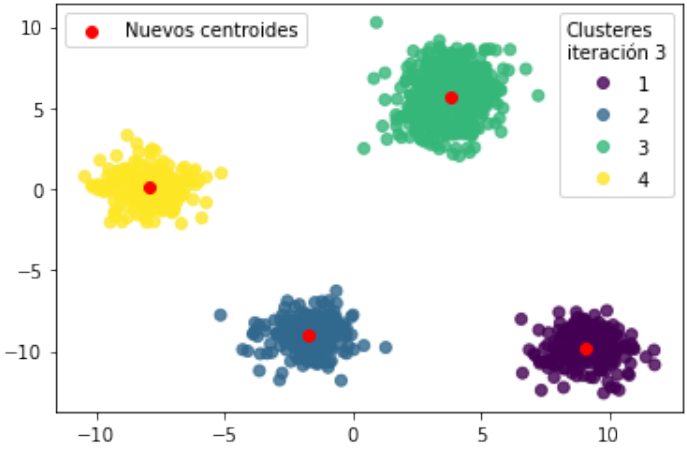
Motivación



Ejemplo
Motivación

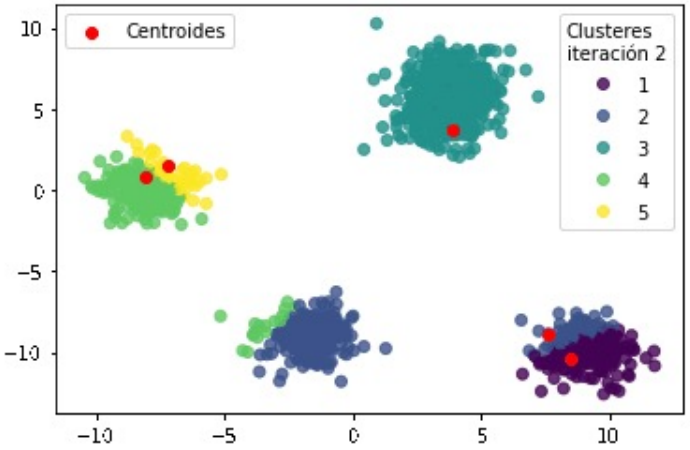


Ejemplo
Motivación



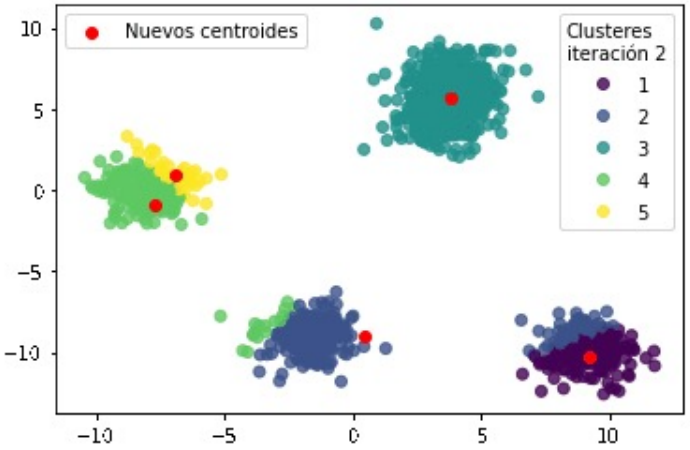
Ejemplo

Proceso K=5



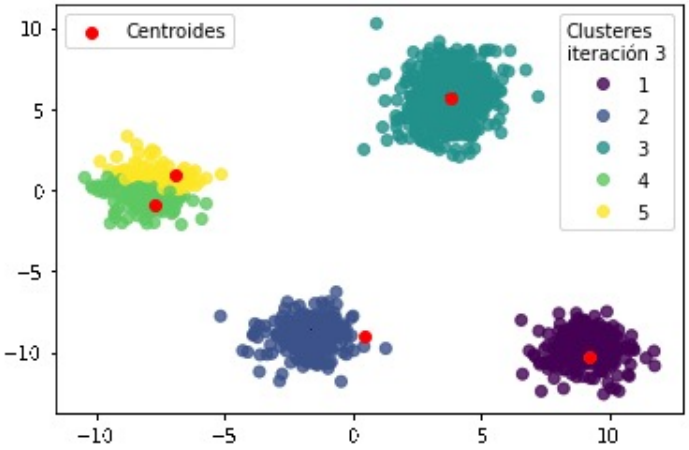
Ejemplo

Proceso K=5



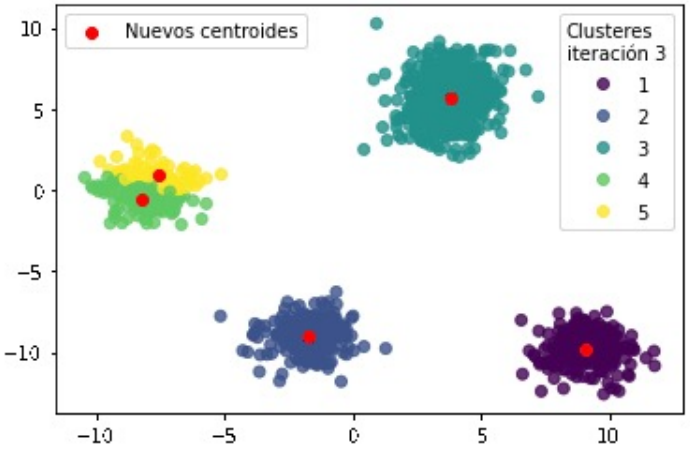
Ejemplo

Proceso K=5



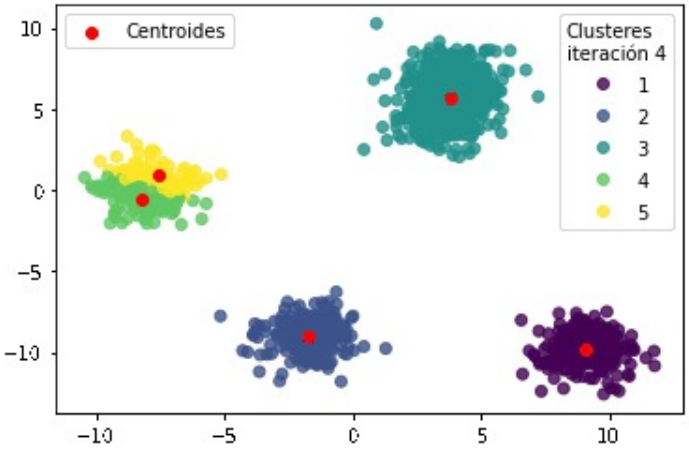
Ejemplo

Proceso K=5



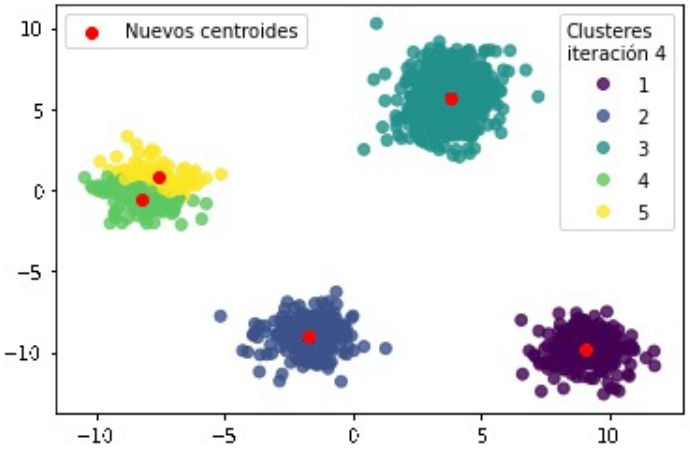
Ejemplo

Proceso K=5



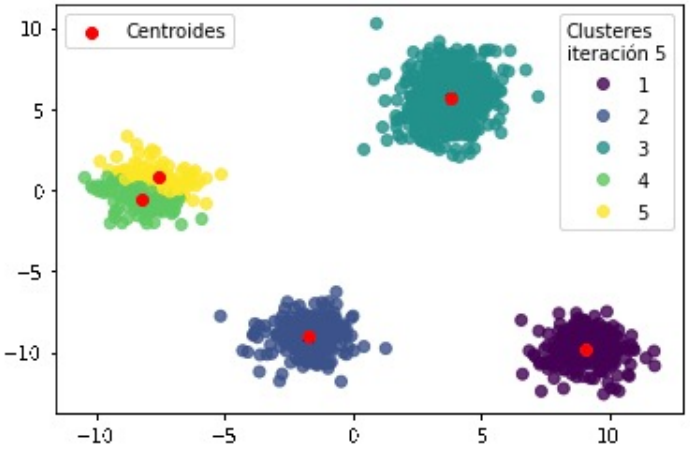
Ejemplo

Proceso K=5



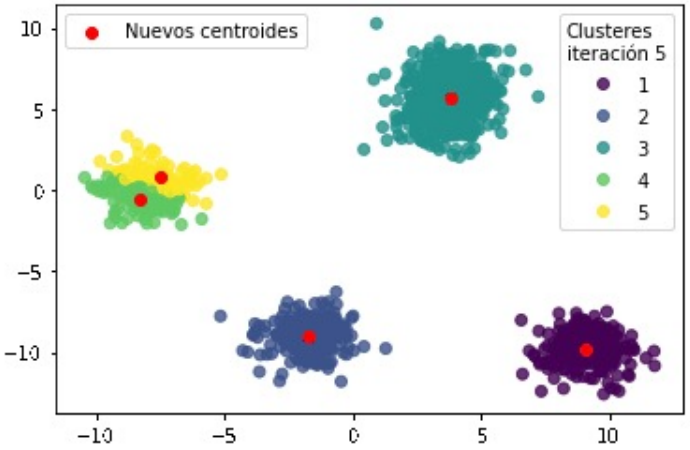
Ejemplo

Proceso K=5



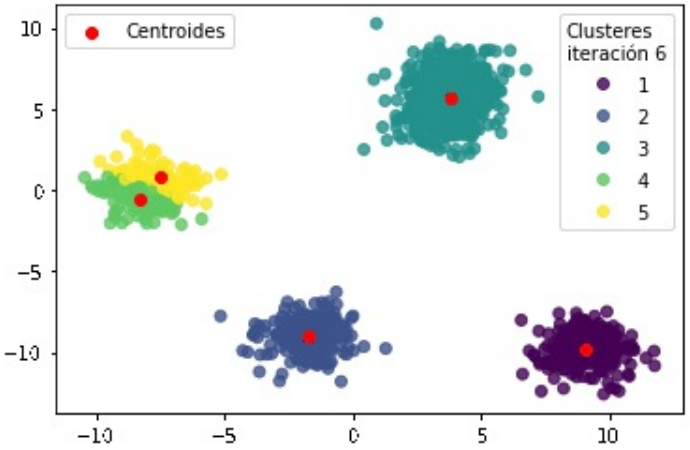
Ejemplo

Proceso K=5



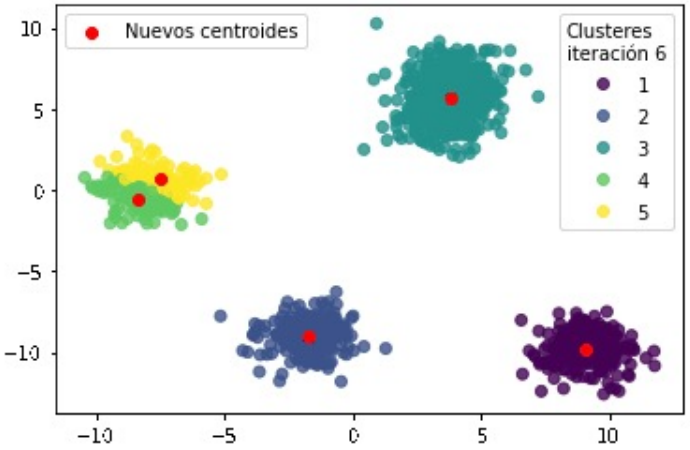
Ejemplo

Proceso K=5



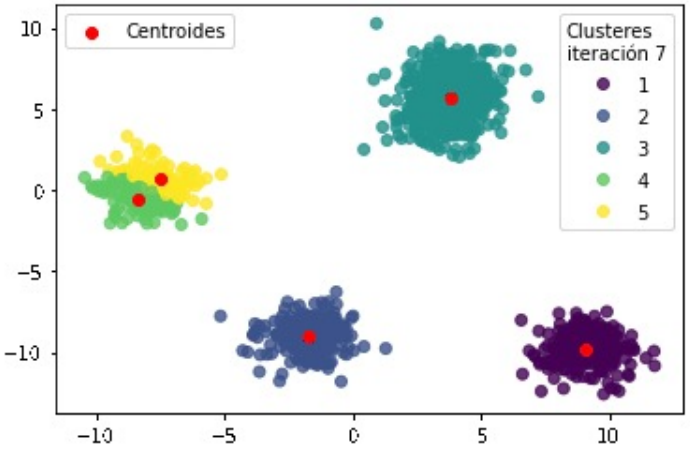
Ejemplo

Proceso K=5



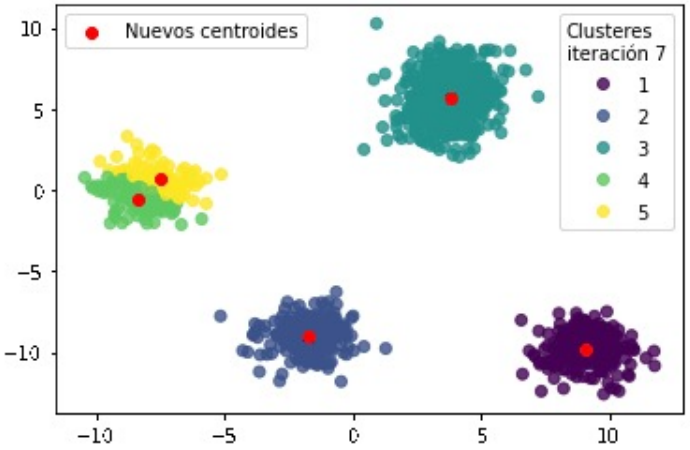
Ejemplo

Proceso K=5



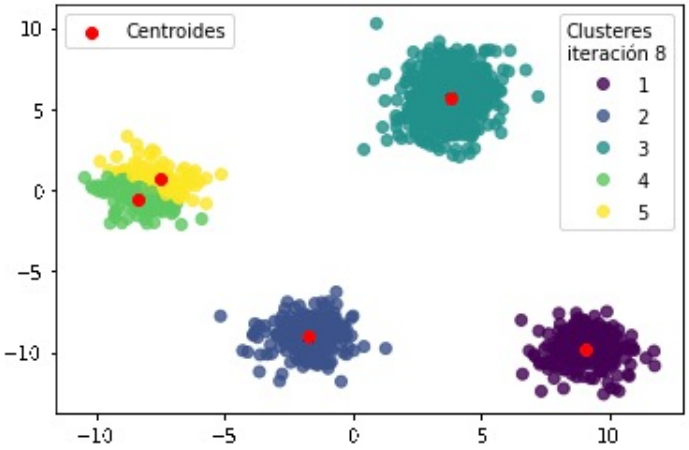
Ejemplo

Proceso K=5



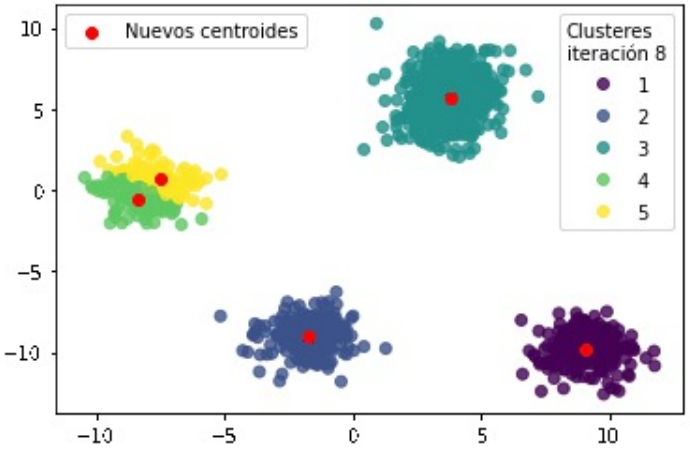
Ejemplo

Proceso K=5



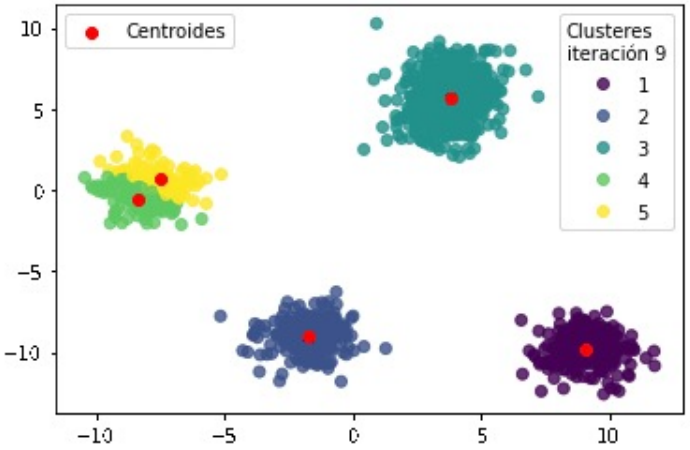
Ejemplo

Proceso K=5



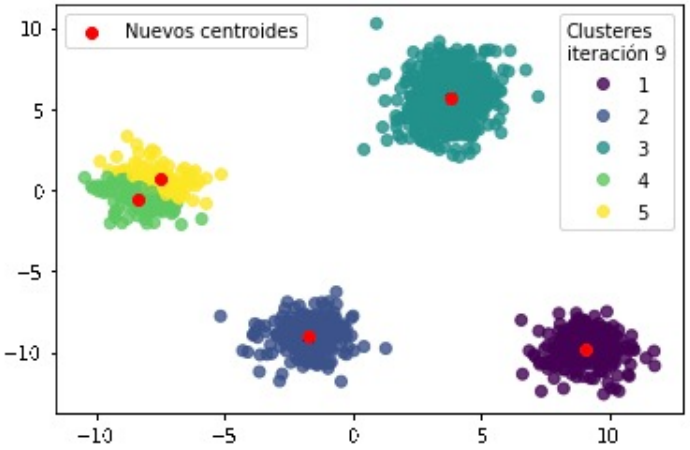
Ejemplo

Proceso K=5



Ejemplo

Proceso K=5



Volvemos en 15 mins con R

R para ML

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/