Introducción al Aprendizaje No Supervisado

1. El concepto de distancia y similaridad.
2. Introducción al aprendizaje no supervisado.
3. K-Means.

El concepto de distancia y similaridad.

- Una medida de similitud cuantifica qué tan próximos están dos objetos, entregando generalmente el valor 0 para aquellos que no tienen relación alguna. A mayor valor de la medida de similitud, mayor es la proximidad o parecido entre dos objetos.

- Las medidas de distancias están íntimamente relacionadas a las de similitud, pero de manera inversa. Esto es, a mayor valor de la medida, más lejanos son los puntos considerados. Cada objeto tendrá distancia igual a 0 al ser comparado consigo mismo.

Existen medidas para cuantificar la proximidad de objetos representados en espacios con dimensiones numéricas, binarias, categóricas, ordinales y mezclas de estos. Por ejemplo:

- Mediciones del largo y ancho de petalos y sepalos de distintas flores una misma especie correspondería a objetos numéricos en 4 dimensiones.

- Se tienen registros de color de ojos, de pelo y nivel de escolaridad correspondería a objetos representados con atributos categóricos en un espacio con 3 dimensiones.

Para un conjunto de datos con n objetos, se utiliza la matriz de distancia (o similitud en caso contrario) que contiene las distancias medidas entre todos los pares de objetos:

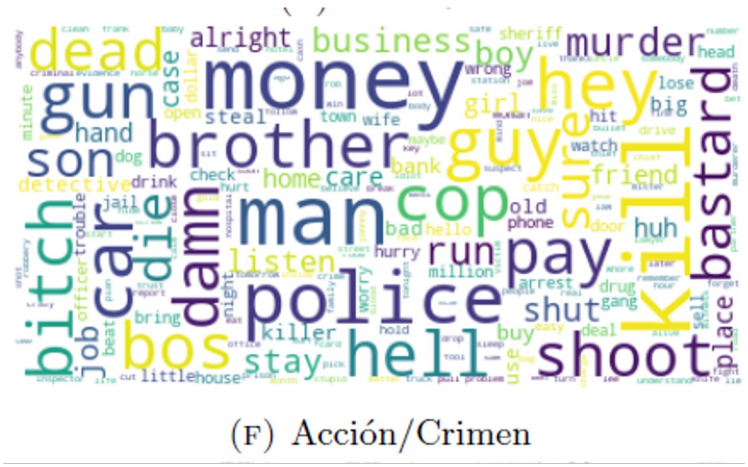
$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Introducción al aprendizaje no supervisado.

Aprendizaje no supervisado

¿Por qué estudiamos aprendizaje No-Supervisado?

Es más **fácil** conseguir datos y más **barato**, es más que nada data generada con una máquina (no hay que pagarle a alguien para identificar clases o chequear el output)



(F) Acción/Crimen

detección de tópicos

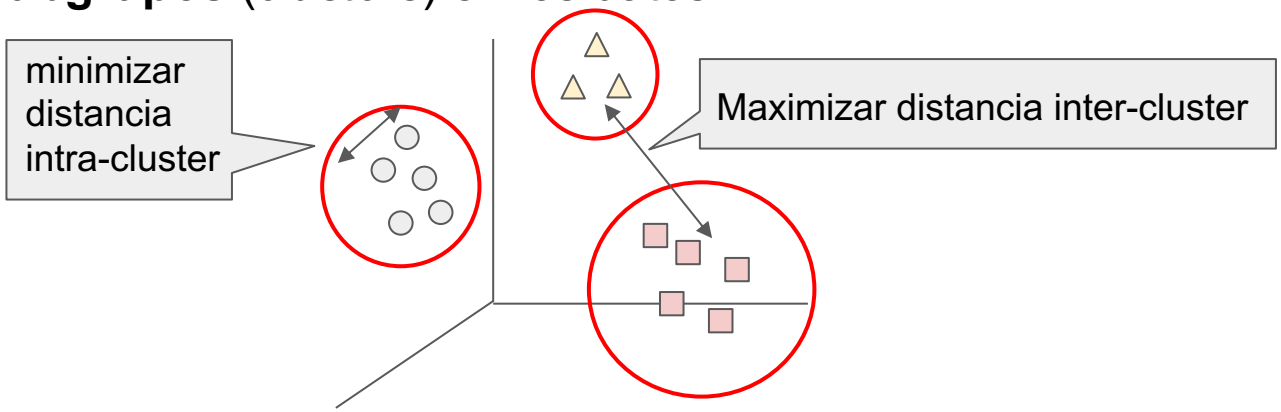
Recomendación/Publicidad

Aprendizaje no supervisado

Clustering

Proceso de agrupar un conjunto de objetos en múltiples grupos (*clusters*), de manera que los objetos ubicados dentro de un *cluster* tengan alta similitud entre ellos y que a su vez sean muy disímiles respecto de los objetos en los otros *clusters*.

Encontrar **subgrupos** (*clústers*) en los datos



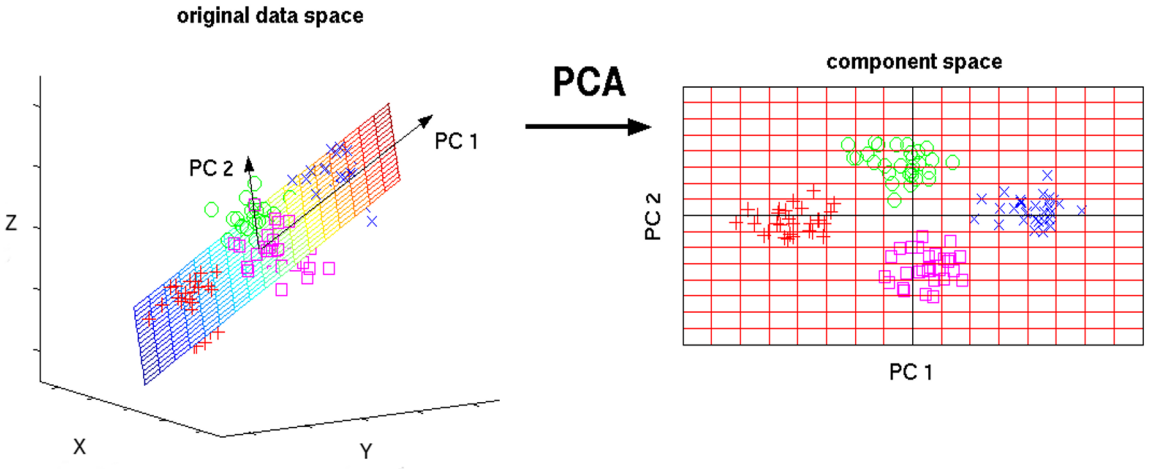
Observaciones dentro de un cluster **similares**
Observaciones entre clusters **no similares**

Aprendizaje no supervisado

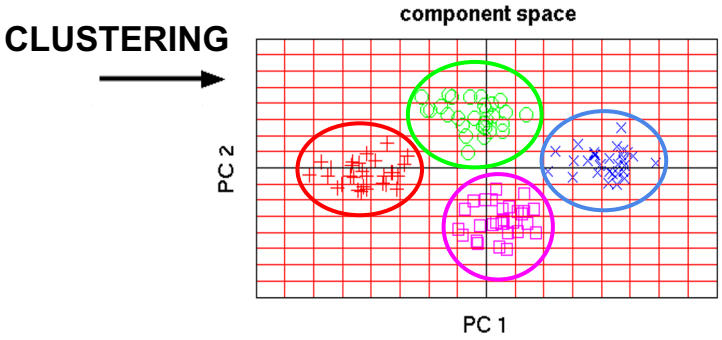
¿Recuerdan PCA? dijimos que era reducción de dimensionalidad.

Esto también es aprendizaje no supervisado, pero ahora nos centraremos en clustering.

Reducir dimensión maximizando la varianza

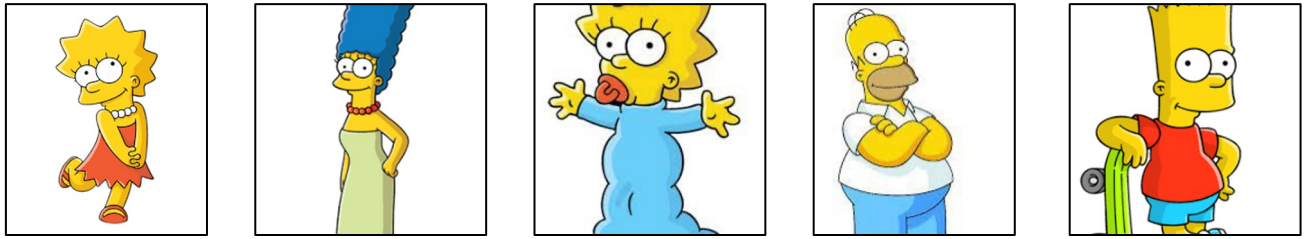


Encontrar grupos homogéneos

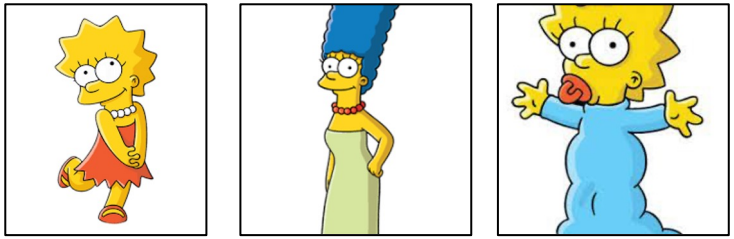


Aprendizaje no supervisado

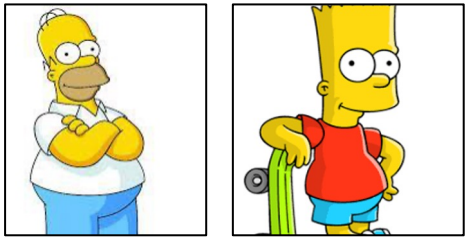
Clustering - forma natural de agrupar los datos



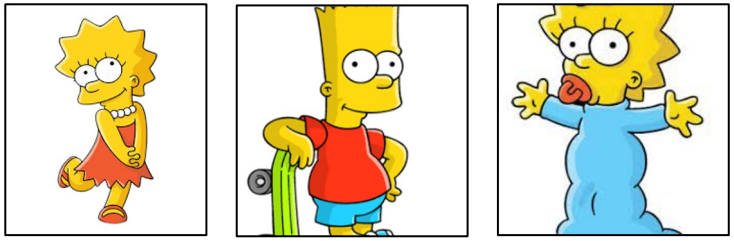
Mujeres



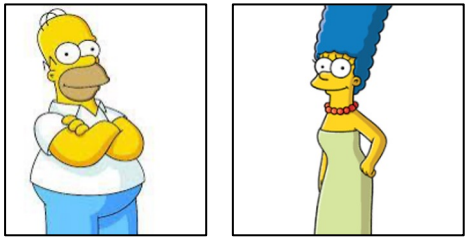
Hombres



Niñ@s

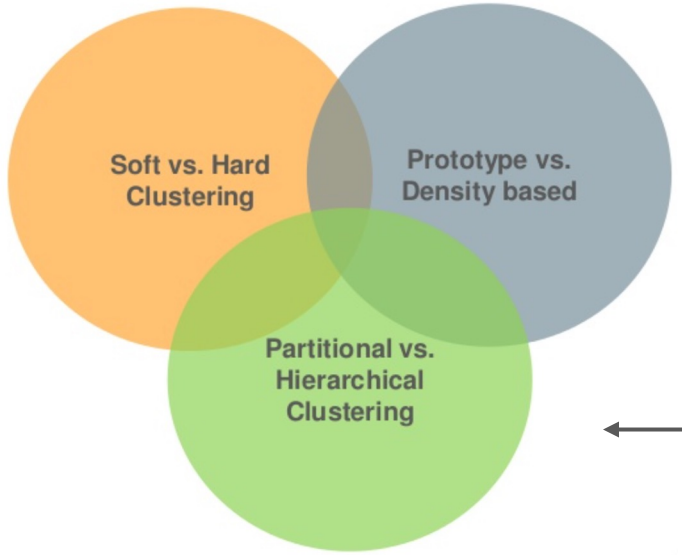


Adultos

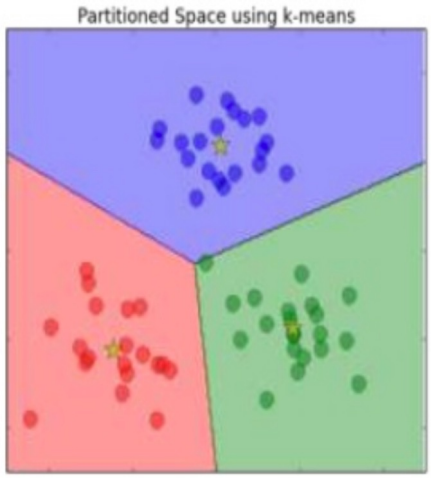


Aprendizaje no supervisado

Hay muchos métodos de clusterización y distintos criterios de división

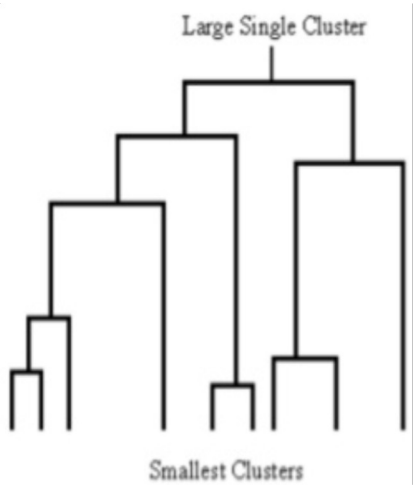


Partición



-particiona el espacio
-encuentra todos los clusters simultáneamente

Jerárquico



-genera una jerarquía de clusters anidados

K-Means

K-Means

Solamente puede ser aplicado cuando la media de un conjunto de datos está definida. Obtiene un conjunto de *k* grupos, todos ellos disjuntos.

- Para el caso de variables nominales, existe el método *K*-Modas, en el cual se reemplaza la media por la moda.

- Otra extensión es *K*-Medoide en que el representante es siempre un punto del conjunto de datos.

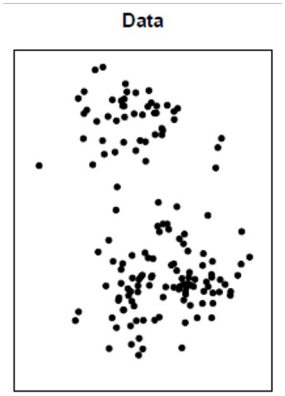
La gran desventaja de estos métodos radica en la necesidad de especificar la cantidad de clusters (*k*).

K-Means: Esquema

Inicialización random

sklearn.cluster.KMeans

Damos el número de clusters ***k*** que queremos obtener



init='random'

elige aleatoriamente ***k*** samples como centroids

init='k-means++'

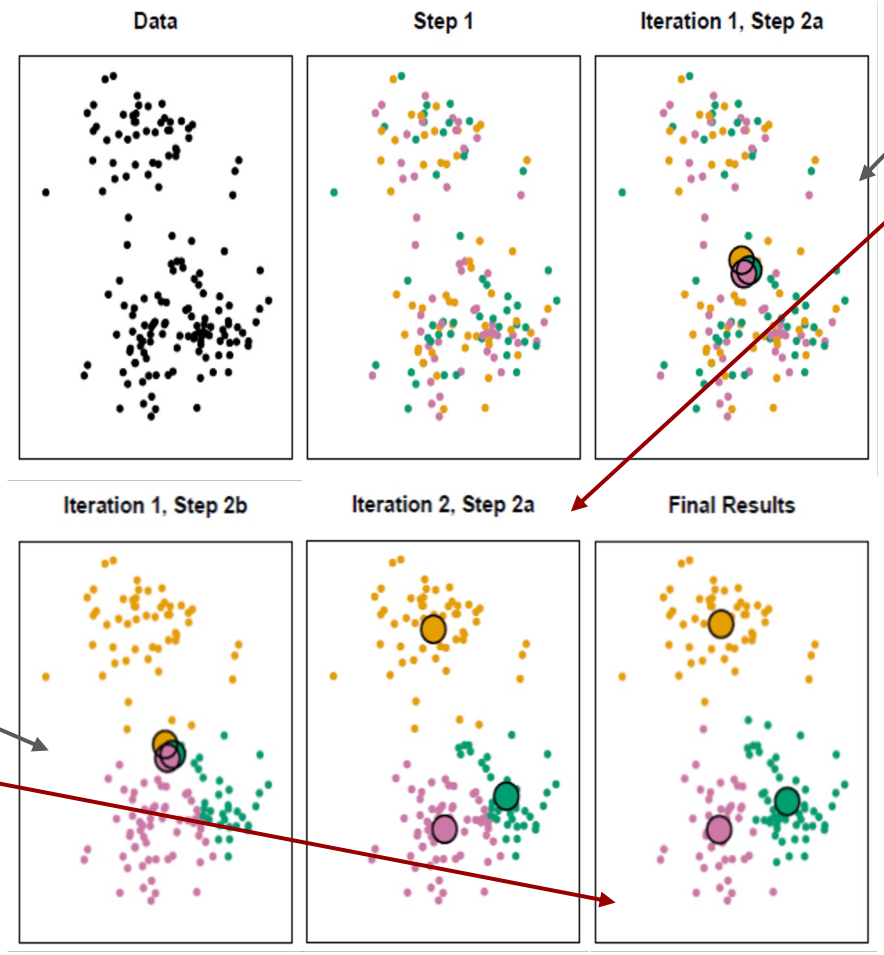
mayor velocidad de convergencia

1. Escoge aleatoriamente un dato y lo asigna como centroid
2. Para los otros datos x, calcula D(x), distancia entre x y el centro más cercano que ya ha sido seleccionado.
3. Escoge un nuevo punto al azar como nuevo centroid, utilizando una distribución de probabilidad ponderada donde un punto x es escogido con la probabilidad proporcional a $D(x)^2$.
4. Repite paso 2 y 3 hasta que se hayan seleccionado k centroids.

K-Means: Esquema

Inicialización random

Damos el número de clusters ***k*** que queremos obtener



Computa los ***centroids*** (centros) de cada cluster como el promedio de las features de sus samples

le asigna a cada sample la etiqueta del cluster cuyo centroid es más cercano (distancia euclídea al cuadrado)

Termina cuando en una iteración no hay cambio de etiqueta o se llega a un máximo de iteraciones 'max_iter'

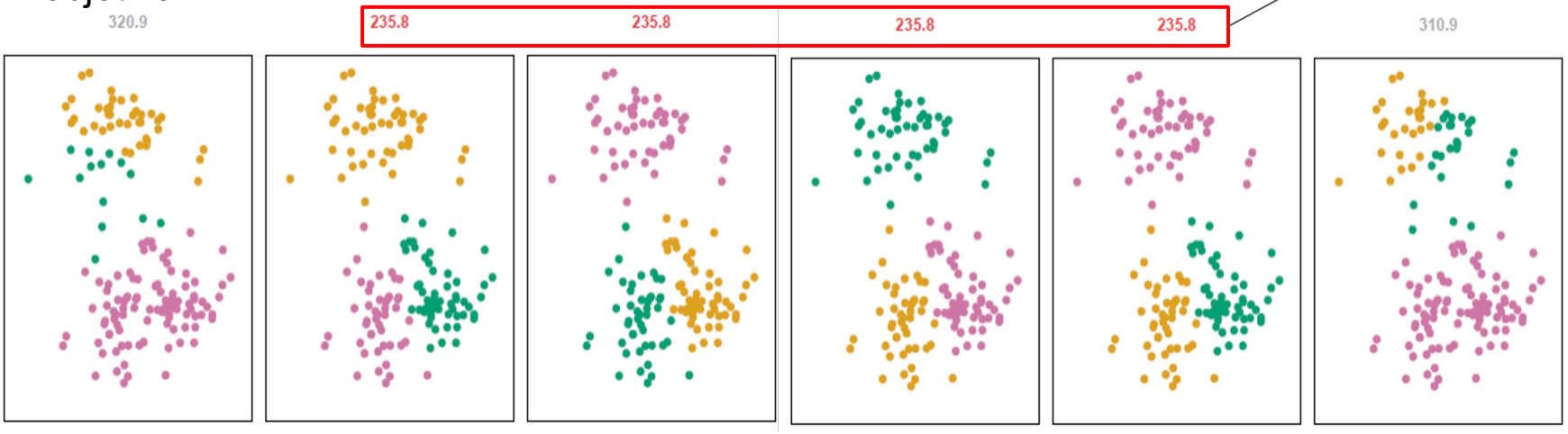
K-means: Función objetivo

Buena clusterización es la que minimiza la varianza entre datos de un mismo cluster

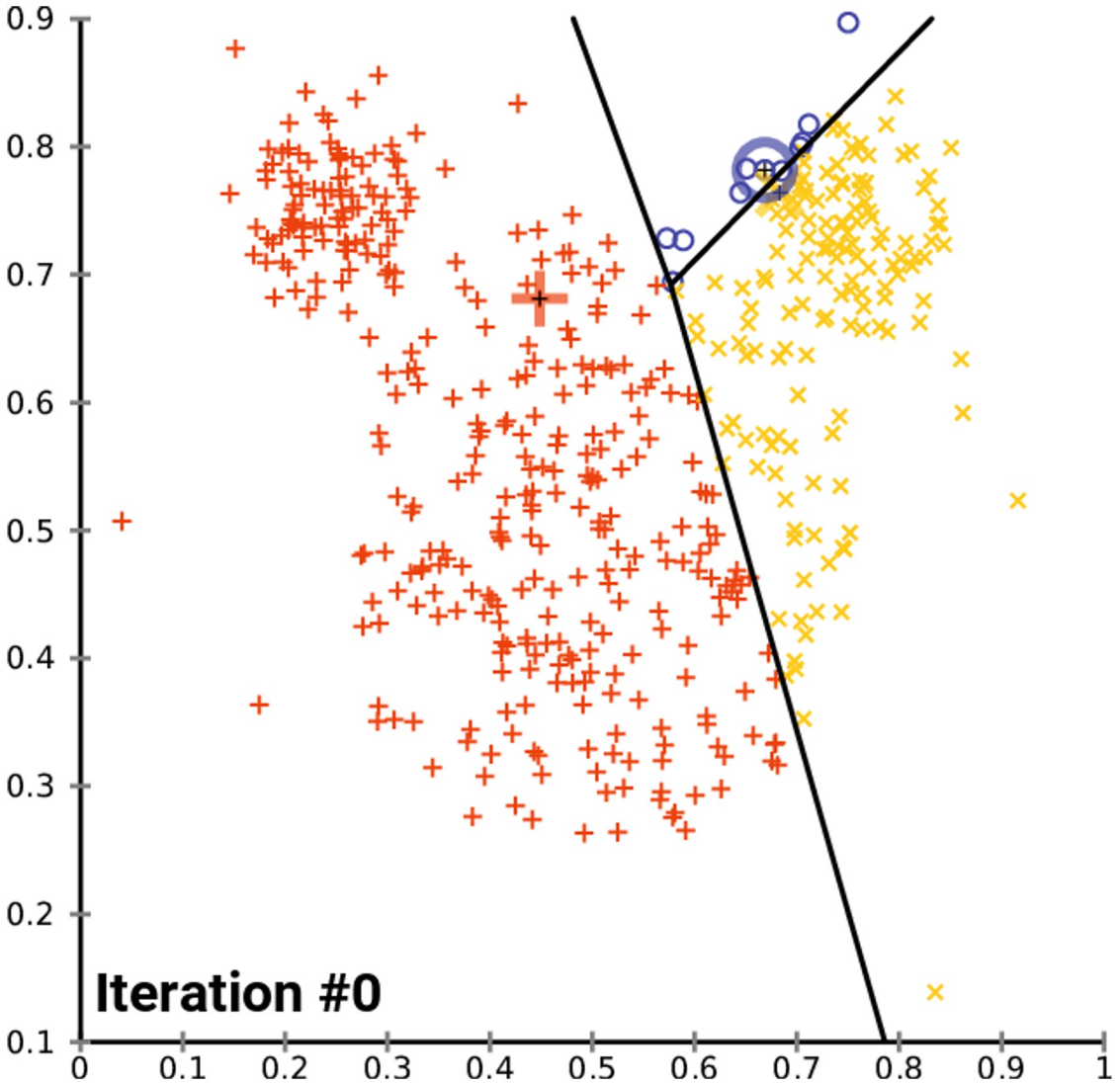
$$\text{SSE} = \underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Básicamente K-means es un algoritmo de optimización de esta función objetivo

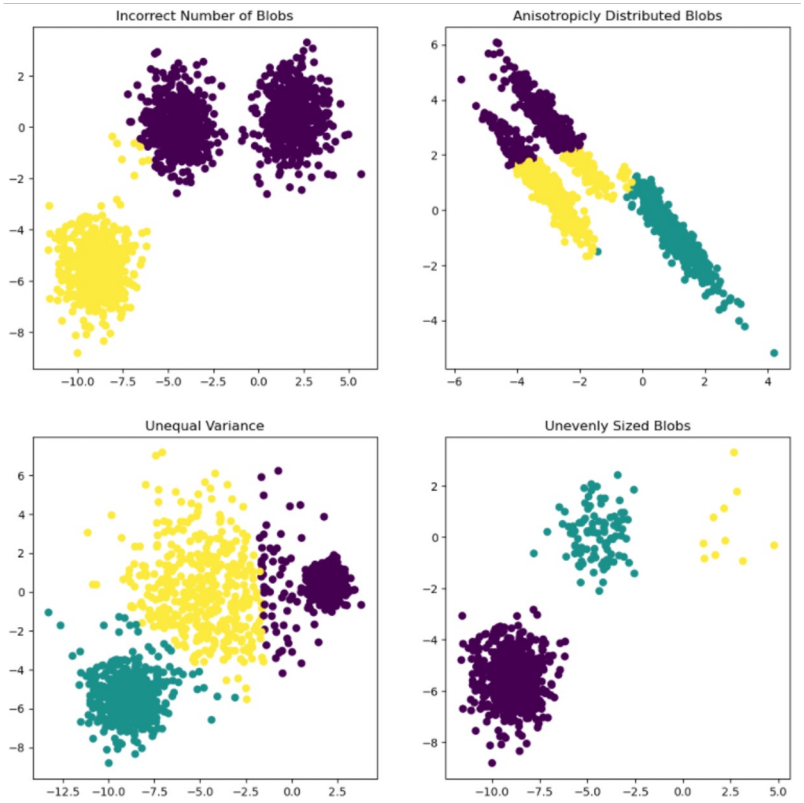
Elige alguna de estas 4 inicializaciones



distintas inicializaciones del mismo modelo con los mismos datos ‘n_init’


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #0

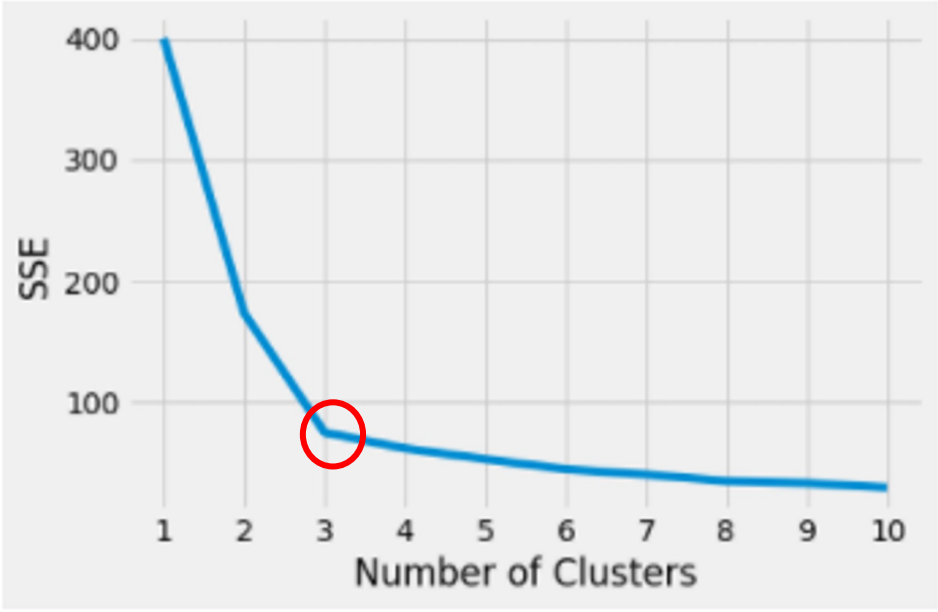
+ Simple y Fácil de implementar
+ Orden del algoritmo es lineal
- Depende de la inicialización
- Tiende a caer en un mínimo local
- Sensible a outliers
- Los clusters tienen que tener forma esférica
- No se puede aplicar a data categórica



No es trivial elegir ***k*** en la mayoría de los dataset reales. No hay algún método que funcione siempre.

Un método es el método Elbow (el método del codo)

Se acumulan las sumas de diferencias al cuadrados de todos los grupos y se grafican para distintos valores del parámetro **k**. Finalmente, se escoge visualmente aquel valor para el cual la caída en la suma total es marginal.



Otro método es el método de Silhouette (coeficiente de Silhouette)

- Medida de cuán similar es un dato, a los datos de su cluster, en comparación a los datos del cluster más cercano.
- Su valor va [-1,1]
- 1 indica que el dato está bien emparejado en su propio cluster y mal emparejado con los datos de otros clusters.

- **K-Means:**

  sklearn.cluster.KMeans

- **Método Elbow:**

  from yellowbrick.cluster import KElbowVisualizer

- **Coeficiente de Silhouette:**

  sklearn.metrics.silhouette_score