Universidad Ricardo Palma

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

PROGRAMA DE ESPECIALIZACIÓN DATA SCIENCE NIVEL I

R + Python

MÓDULO I

Modelos de Segmentación: Algoritmo K - Means, K – Means ++

A nuestro recordado Maestro

Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

« Para poder **seguir** a veces hay que **empezar de nuevo**»

Agenda

- Segmentación de Clientes.
- Modelos No Supervisados.
- Análisis de Conglomerados :Objetivos
- Análisis de Conglomerados : Criterio de Inercia
- Ejemplo de Ilustración: Estudiantes
- Algortimo de K-Means: Objetivo
- Algortimo de K-Means: Método.
- Algortimo de K-Means: Elección de k.

SEGMENTACIÓN DE CLIENTES

SEGMENTACIÓN DE CLIENTES

- Es el proceso de dividir **clientes** en grupos basados en características comunes para que las compañías puedan mercadear cada grupo efectiva y apropiadamente.

- Los grupos o segmentos deben ser homogéneos intragrupos y heterogéneos intergrupos.

Modelos No Supervisados

- No hay una variable objetivo (Variable de Salida).
- No hay variables que ayudan a predecir a la variable de salida.

- Todas las variables tienen la misma importancia.
- Se busca la interdependencia de las variables.

Modelos no Supervisados

Métodos de agrupamiento o clustering

- **"Clustering":**(Clasificación no supervisada, aprendizaje no supervizado): Es similar a la clasificación (discriminación), excepto que los grupos no son predefinidos. El objetivo es **particionar** o **segmentar** un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos se forman basados en la similaridad de los datos o individuos en ciertas variables. Como los grupos no son dados a priori el experto debe dar una interpretación de los grupos que se forman.

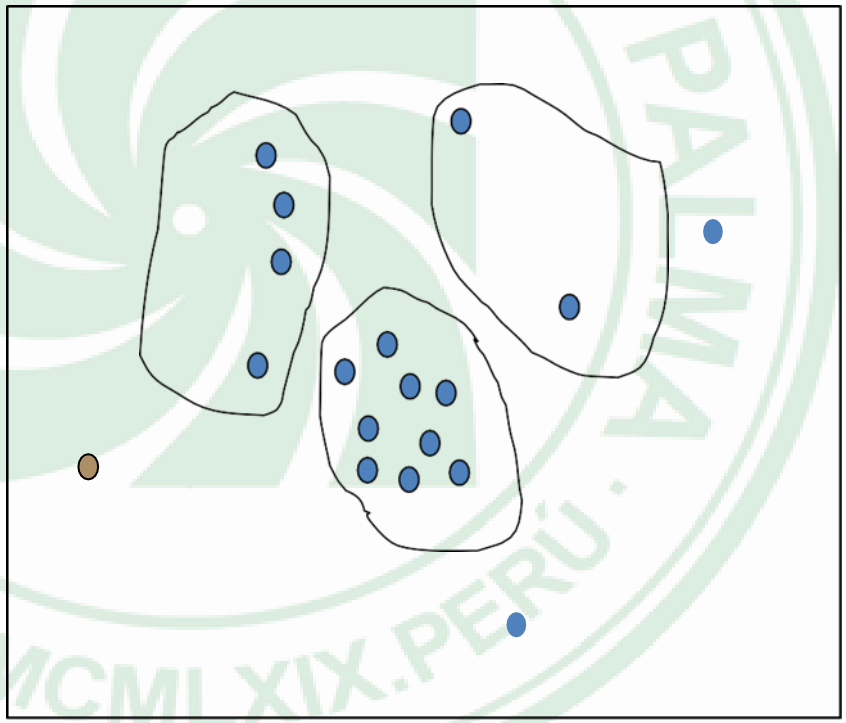
- <u>Métodos</u>:

  - Clasificación Jerárquica (grupos disjuntos).
  - Nubes Dinámicas – k-means (grupos disjuntos).
  - Clasificación Piramidal (grupos NO disjuntos).

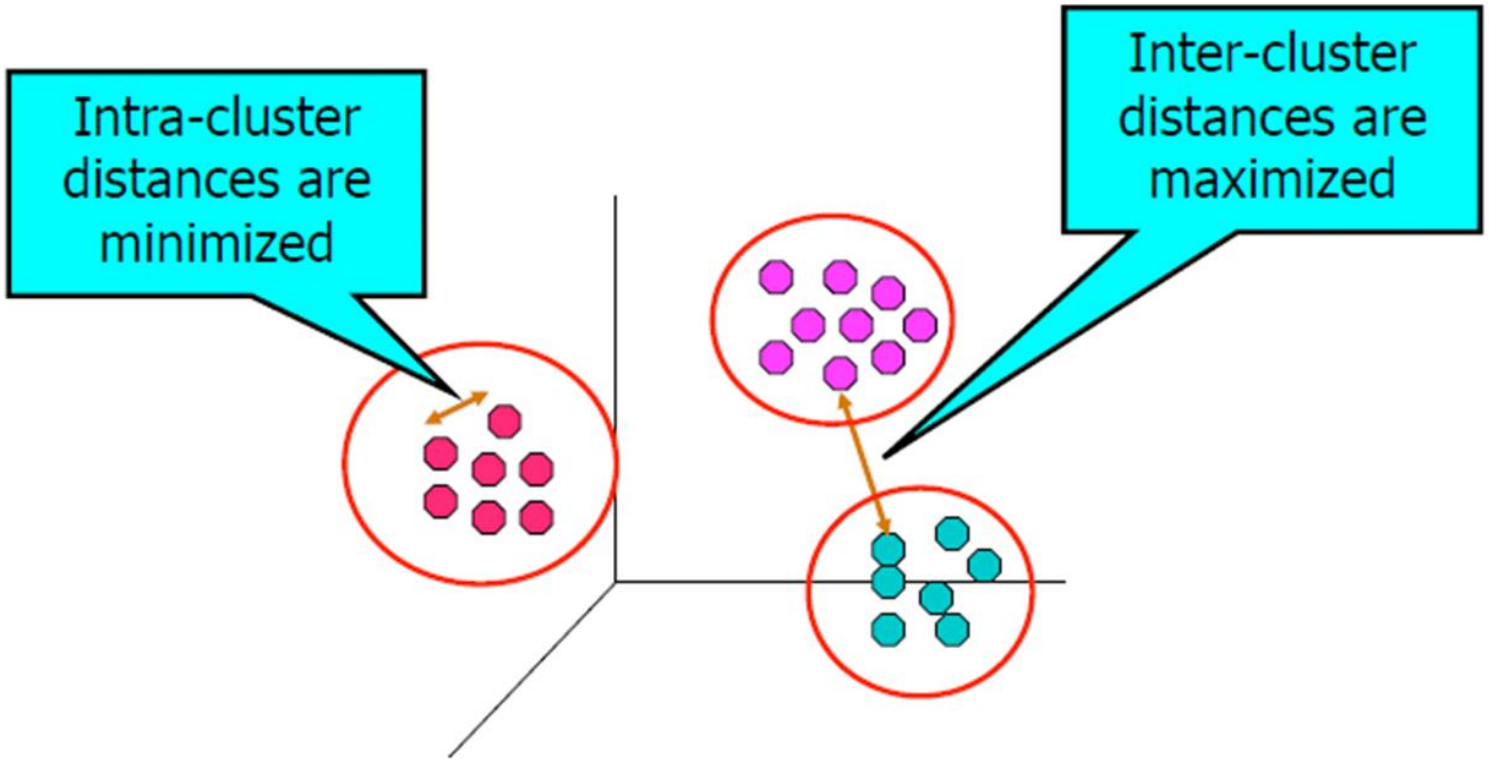
Análisis de Conglomerados : Objetivo

Objetivo:

Obtener clases lo más homogéneas posibles y tal que estén suficientemente separadas.



Análisis de conglomerados: Criterio de la inercia



Análisis de conglomerados: Criterio de la inercia

- Como se ha mencionado , se quiere obtener clases lo más homogéneas posibles y que estén suficientemente separadas. Este objetivo se puede concretar numéricamente a partir de la siguiente propiedad :
- Supóngase que se está en presencia de una partición P=(C_1, C_2, C_3,……C_k) de Ω , donde g_1, g_2, g_3, g_4,……., g_k son los centros de gravedad de las clases:

$$g_k = \frac{1}{|C_k|} \sum_{i \, \varepsilon \, C_k} x_i$$

- g es el centro de gravedad total:

$$g = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Análisis de conglomerados: Criterio de la inercia

- ***Inercia total*** de la nube de puntos:

$$I = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{g}||^2$$

- ***Inercia inter-clases***, es decir la inercia de los centros de gravedad respecto al centro de gravedad total:

$$B(P) = \sum_{k=1}^{K} \frac{|C_k|}{n} ||\mathbf{g}_k - \mathbf{g}||^2$$

Análisis de conglomerados: Criterio de la inercia

- *Inercia intra-clases, es decir la inercia al interior de cada clase :*

$$W(P) = \sum_{k=1}^{K} I(C_k) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$$

Teorema: Igualdad de Fisher

Inercia total = Inercia inter - clases

+

Inercia intra-clases

$$I = B(P) + W(P)$$

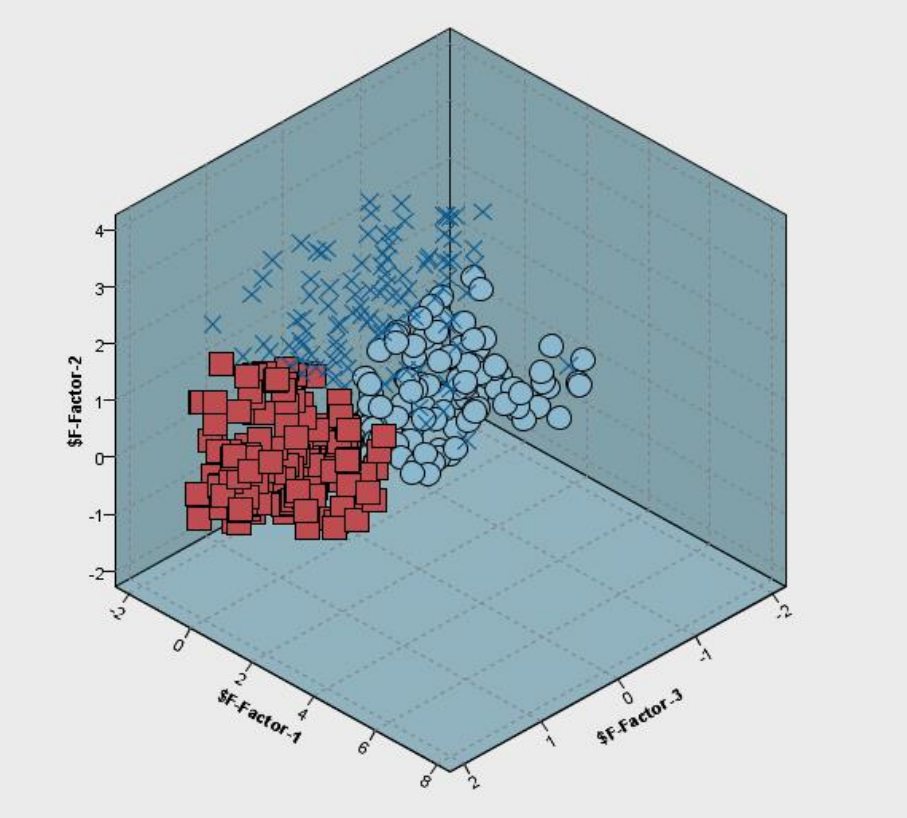
Objetivos del clustering

- ***Objetivo:*** Se quiere que ***B(P)* sea máxima** y ***W(P)* sea mínima.**

- Como la inercia *I(P)* es fija, dada la nube de puntos, entonces al maximizar *B(P)* se minimiza automáticamente *W(P)*.

- Por lo tanto, los dos objetivos (homogeneidad al interior de las clases y separación entre las clases) se alcanzan al mismo tiempo al querer minimizar *W(P)*.

ALGORITMO DE K - MEANS



Objetivo del Método K-means

- Así, el objetivo en el método de K-means es encontrar una partición *P* de Ω y representantes de las clases, tales que *W*(*P*) sea mínima.

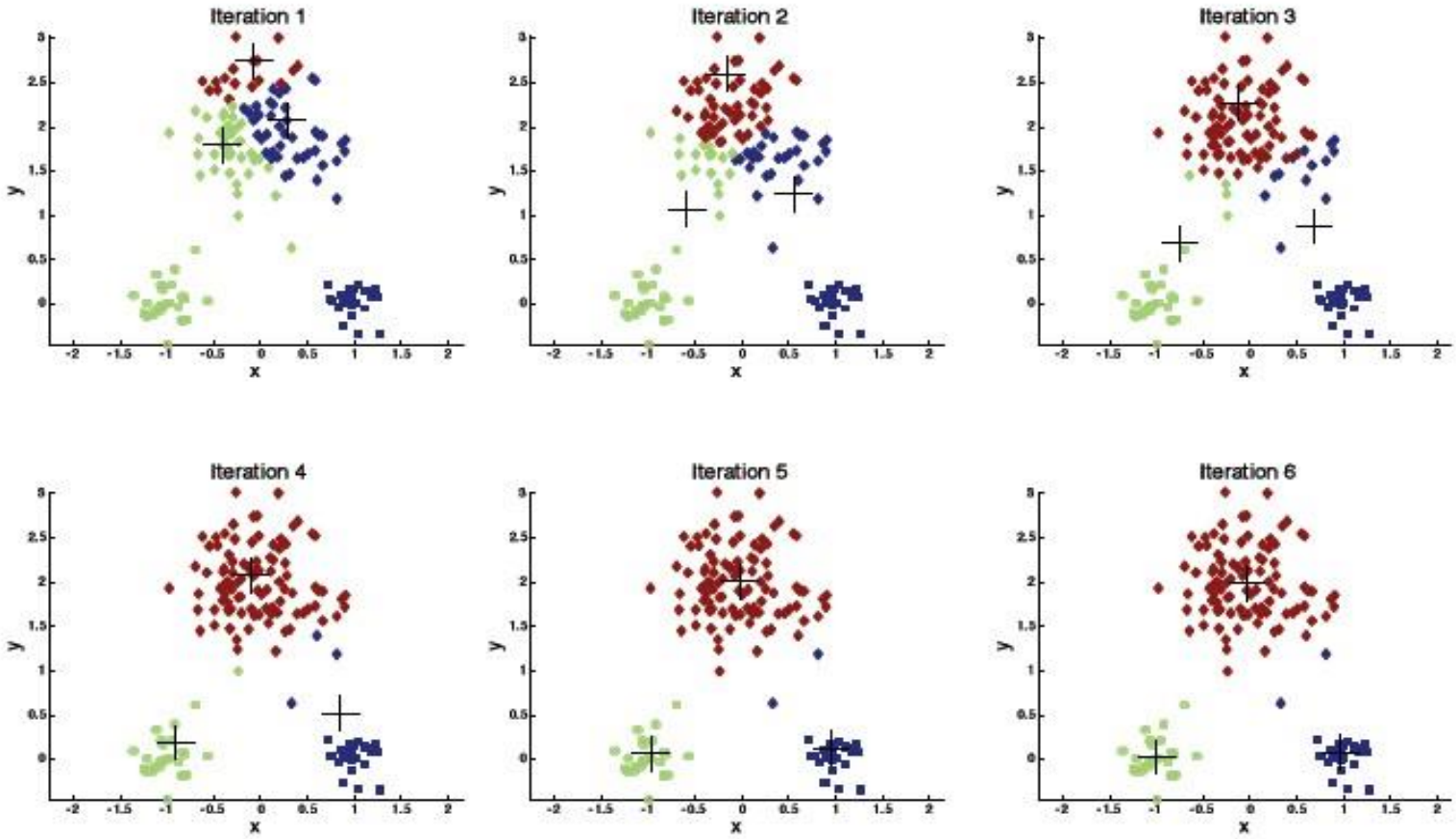
Método de K - Means

- ✓ Existe un poco de confusión en la literatura acerca del método de las k-medias, ya que hay dos métodos distintos que son llamados con el mismo nombre.
- ✓ Originalmente, Forgy propuso en 1965 un primer método de reasignación-recentraje que consiste básicamente en la iteración sucesiva, hasta obtener convergencia, de las dos operaciones siguientes:

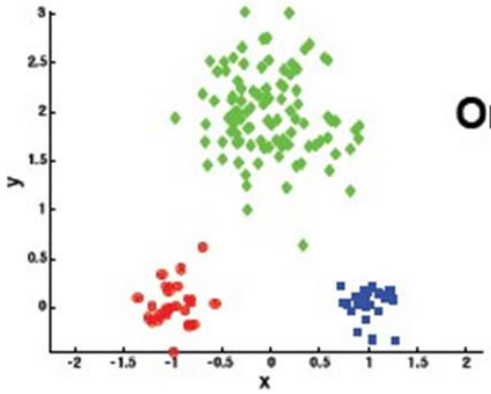
Método de K – Means : Proceso

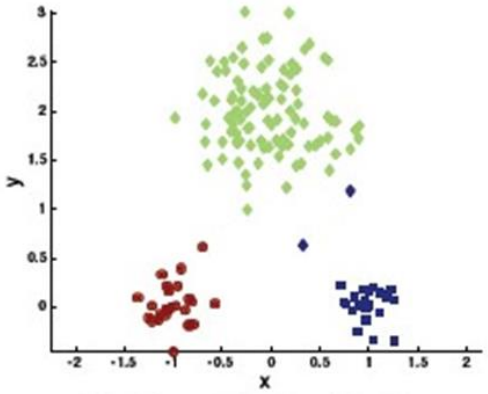
1. Representar una clase por su centro de gravedad, esto es, por su vector de promedios.
2. Asignar los objetos a la clase del centro de gravedad más cercano.

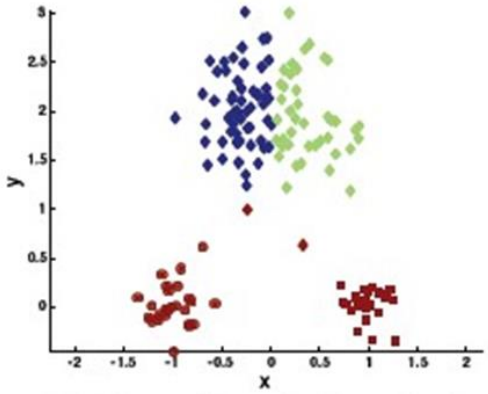
Método de K - Means



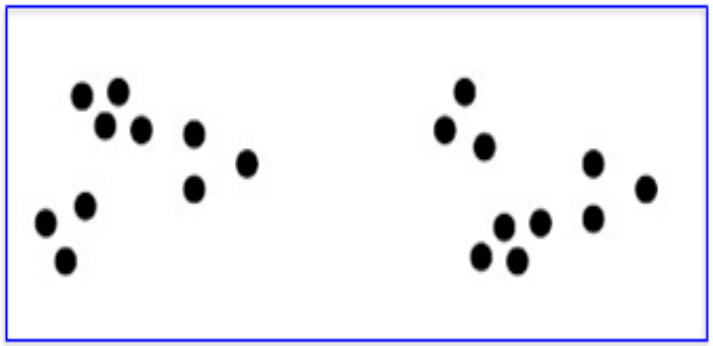
Método de K - Means



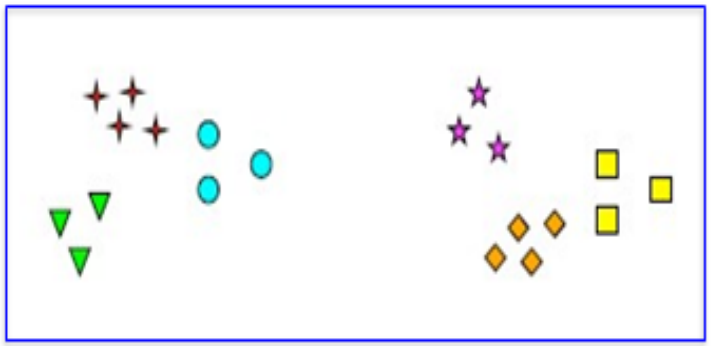




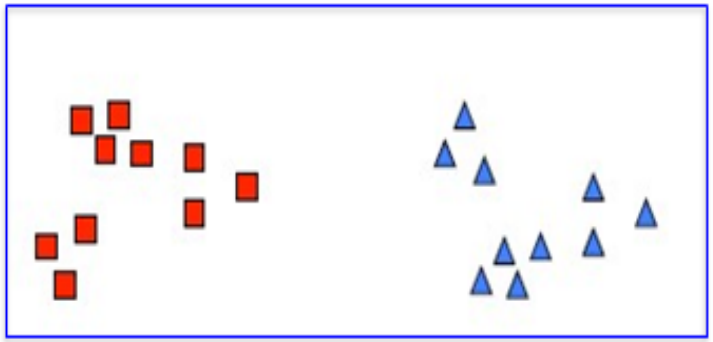
¿ Cuántos clústeres?



Datos originales



6 clústeres



2 clústeres

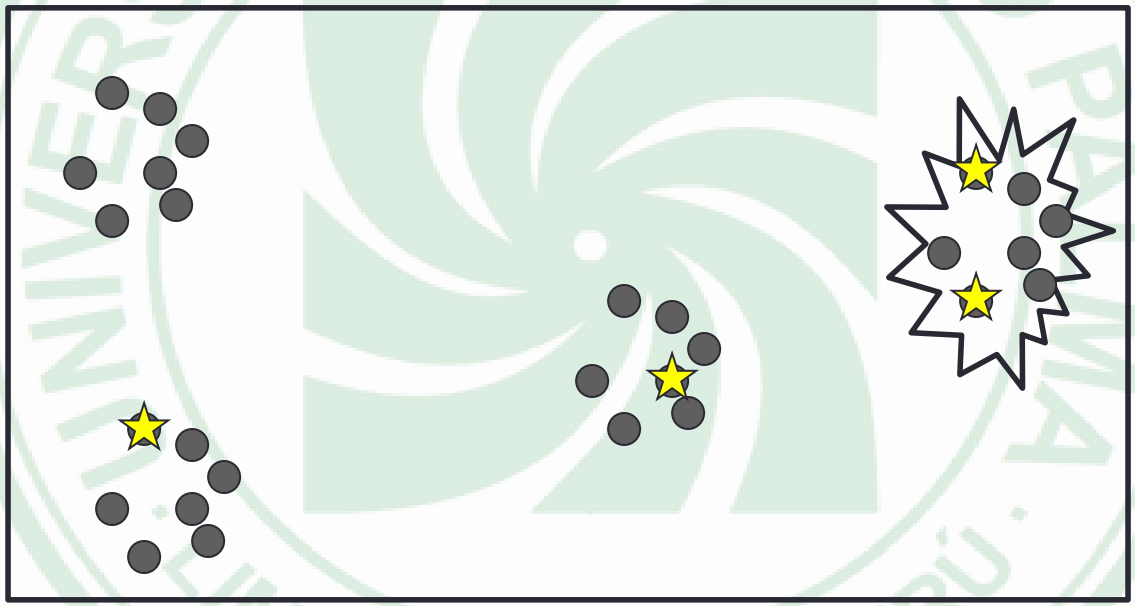
4 clústeres

ELECCIÓN DE K : PROBLEMA COMBINATORIO

- Es necesario hacer notar que, cuando se quiere obtener una partición en *K* clases de un conjunto con *n* individuos, no tiene sentido examinar *todas* las posibles particiones del conjunto de individuos en *K* clases.
- En efecto, se está en presencia de un problema combinatorio muy complejo; sólo para efectos de ilustración, mencionemos que el número de particiones en 2 clases de un conjunto de 60 elementos es aproximadamente 10^{18} y para 100 elementos en 5 clases anda por 10^{68} .

K-MEANS ++ CLUSTERING

ALGORITMO: INICIALIZACIÓN DE K - MEANS



K-MEANS ++ CLUSTERING

- En el análisis del conglomerados o aprendizaje no supervisado, *k*-means++ es un algoritmo que se utiliza para la selección de los valores iniciales (o "semillas") en el algoritmo *k*-means.

- Fue propuesto en 2007 por David Arturo y Sergei Vassilvitskii.

- Se utiliza para evitar los agrupamientos pobres a veces encontrados por el algoritmo k-means estándar.

K-MEANS ++ CLUSTERING

- El algoritmo exacto es como sigue:

  1.- Escoger un centro de entre los puntos de datos utilizando una variable aleatoria uniforme.

  2.- Para cada punto x, calcular D(x), que es la distancia entre x y el centro más cercano que ya ha sido seleccionado.

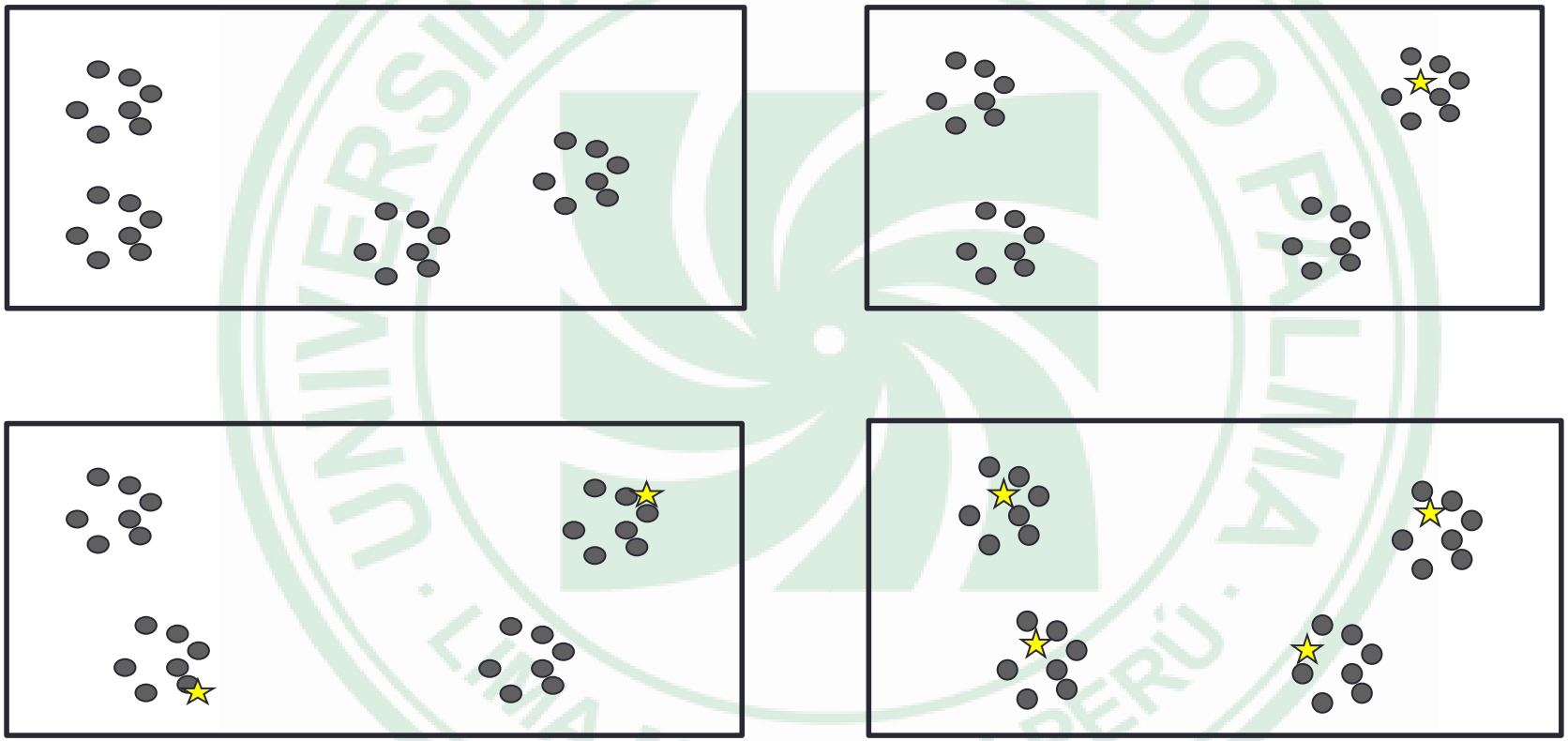
  3.- Escoger un nuevo punto al azar (con variable aleatoria uniforme) como nuevo centro, utilizando una distribución de probabilidad ponderada donde un punto x es escogido con la probabilidad proporcional a $D(x)^2$.

  4.- Repetir paso 2 y 3 hasta que se hayan seleccionado k centros.

- Ahora que los centros iniciales han sido elegidos, continuar utilizando ***k*-means clustering estándar.**

K-MEANS ++ CLUSTERING

K-MEANS++ INICIALIZACIÓN DE K – MEANS ++



Conclusiones

- Aunque la selección inicial en el algoritmo toma tiempo extra, k-means converge muy rápidamente después de la selección de puntos iniciales.

- Los autores probaron su método con conjuntos de datos reales y sintéticos y obtuvieron mejoras de 2-veces en la velocidad, y para ciertos conjuntos de datos, cerca de 1000 veces mejoras en error.

A seguir estudiando…….

- **Fuzzy C-Means Clustering** es una versión difusa del K-means, donde cada punto tiene un grado difuso de pertenecía a cada grupo.
- **Modelos de Mezclas Gausianas** entrenadas con el algoritmo de esperanza-maximización presentan una asignación probabilística a cada grupo, en vez de asignaciones deterministas.
- Algoritmos de filtrado utilizan **K Means-Trees** para mejorar la eficiencia en cada paso del algoritmo.
- El algoritmo **Spherical K-means** es bastante usado para datos direccionales.
- El algoritmo **Minkowski Metric Weighted K-means** trata el problema del ruido asignando pesos a las componentes de los vectores por grupos.

¡Gracias!

Comunidad Data Science Perú

Comunidad Data Science Perú