



Universidad **Ricardo Palma**

RECTORADO
PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

TALLER DE ESTADÍSTICA PARA LA CIENCIA DE DATOS



R + Python



MÓDULO 4 : ANALISIS MULTIVARIADO I

Cluster Jerárquicos y No Jerárquicos



A nuestro recordado Maestro

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



TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”

TALLER DE ESTADÍSTICA PARA CIENCIA DE DATOS

EXPOSITORES



José Antonio Cárdenas Garro
UNMSM

MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma



André Omar Chávez Panduro
UNMSM

MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

**Predictive Modelling
Specialist**



Data Scientist



**Portfolio and
Consumption Analyst**



**Customer Intelligence
Analyst**



Data Analyst



Data Analyst



Correo : josecardenasgarro@gmail.com
LinkedIn : www.linkedin.com/in/jos%C3%A1-antonio-c%C3%A1rdenas-garro-599266b0

Correo : andrecp38@gmail.com /
09140205@unmsm.edu.pe
LinkedIn : www.linkedin.com/in/andr%C3%A9-ch%C3%A1vez-a90078b9



TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

A nuestro recordado Maestro

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



TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”

« Si el **Plan** no funciona ,
cambia el **Plan** pero **No Cambies** los
OBJETIVOS NI METAS»



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



TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”

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

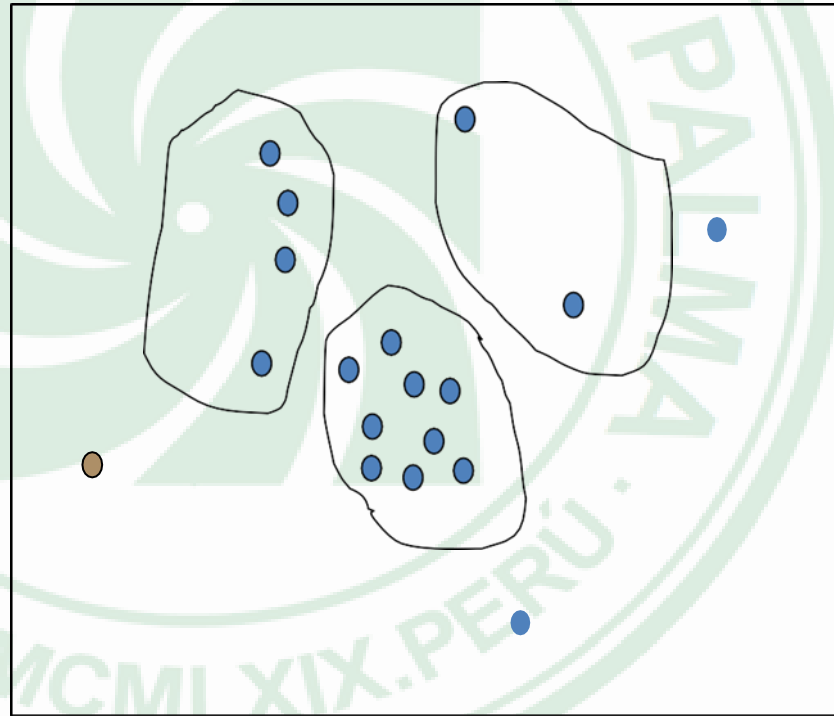
- **“Clustering”**: (Clasificación no supervisada, aprendizaje no supervisado): Es similar a la clasificación (discriminación), excepto que los grupos no son predefinidos. El objetivo es **particionar** o **segmentar** un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos 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.
- Métodos:
 - Clasificación Jerárquica (grupos disjuntos).
 - Nubes Dinámicas – k-means (grupos disjuntos).
 - Clasificación Piramidal (grupos NO disjuntos).



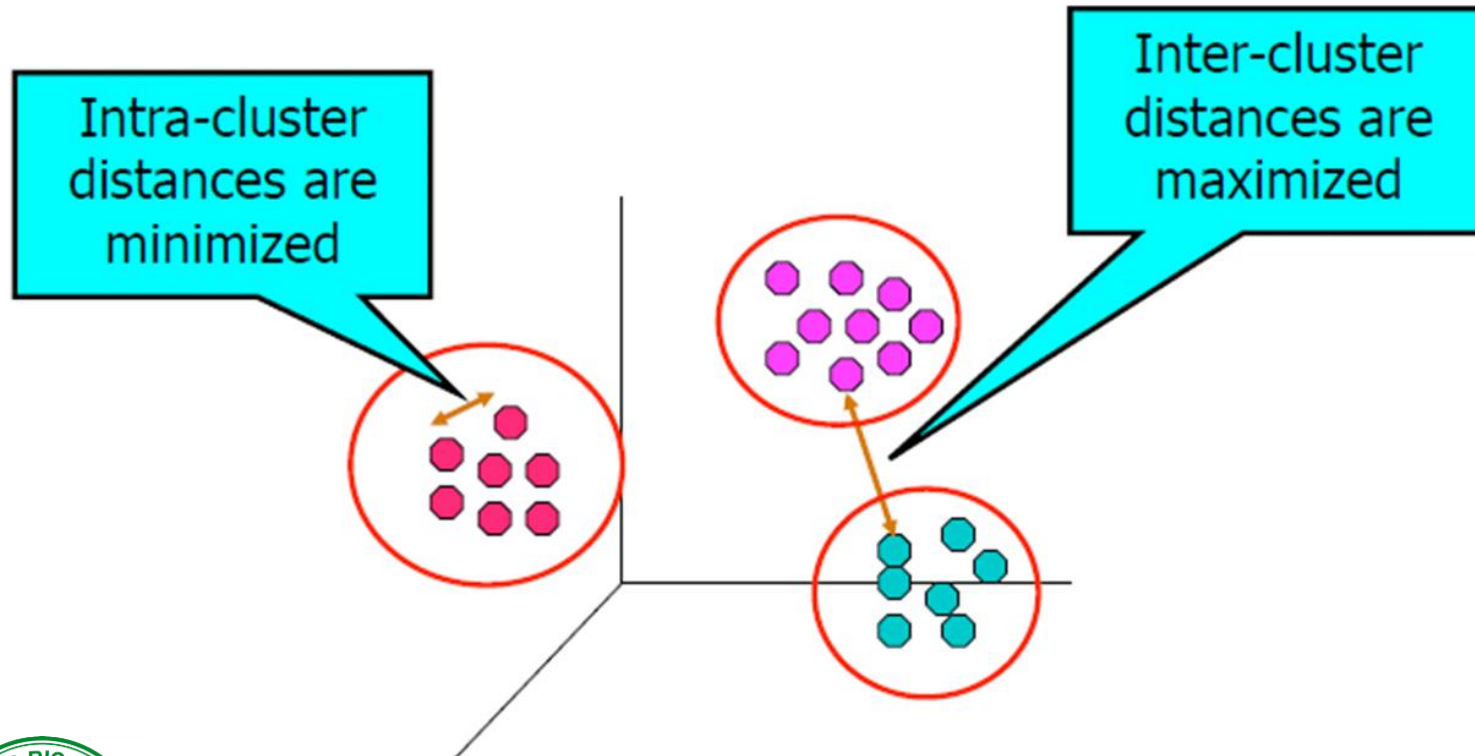
Análisis de Conglomerados : Objetivo

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 $P=(C_1, C_2, C_3, \dots, C_k)$ de Ω , donde $g_1, g_2, g_3, g_4, \dots, g_k$ son los centros de gravedad de las clases:

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

- 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 :*

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



ANÁLISIS DE CONGLOMERADOS: CRITERIO DE LA INERCIA

Teorema: Igualdad de Fisher

Inercia total = Inercia inter - clases
+
Inercia intra-clases

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

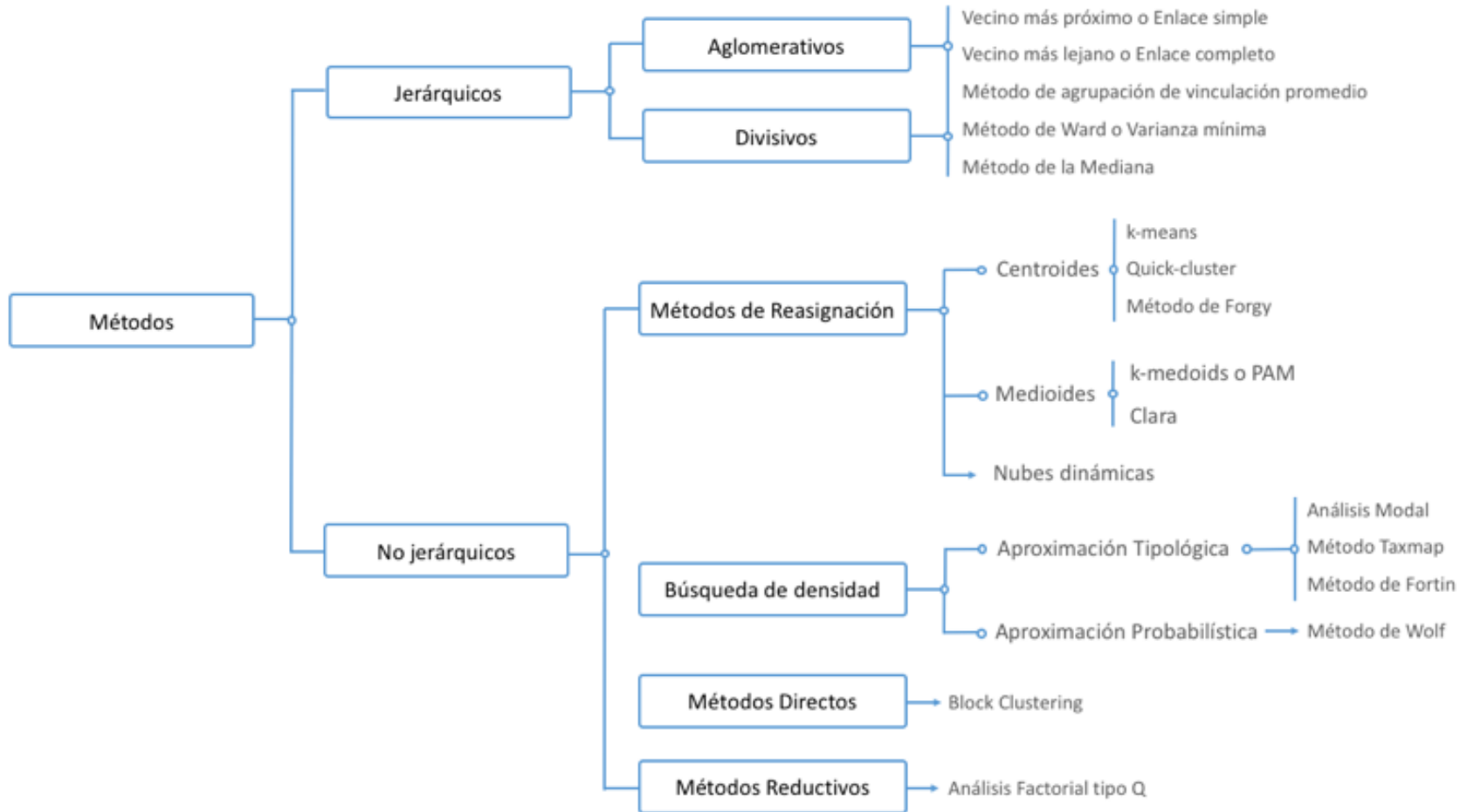


Objetivos del clustering

- ✓ **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 $W(P)$.



Tipos de Clustering



CLUSTER JERARQUICOS

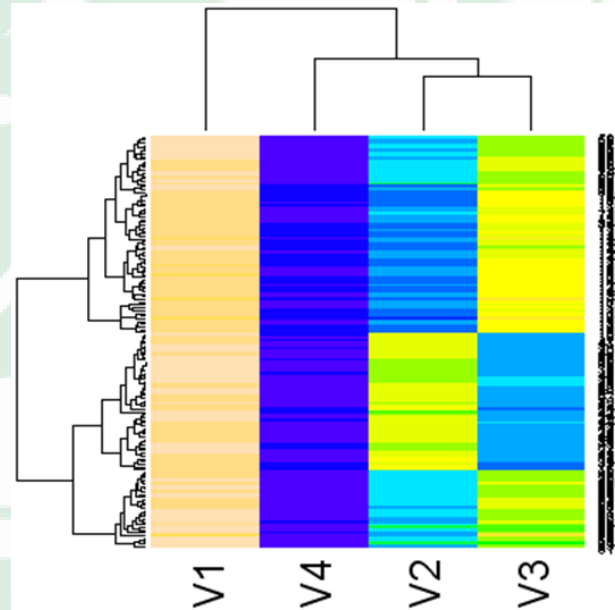
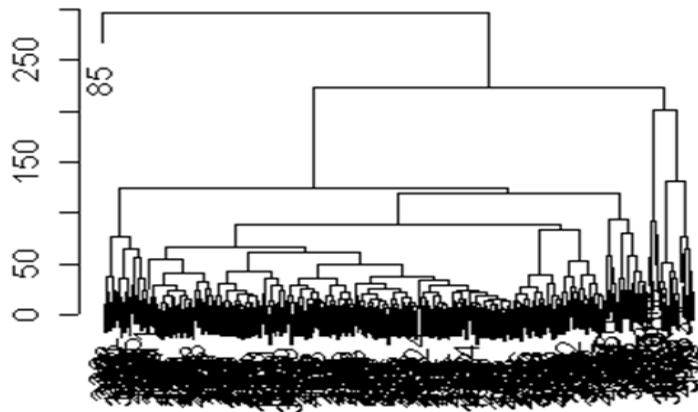
- En estos algoritmos se generan sucesiones ordenadas (jerarquias) de conglomerados. Puede ser juntando cluster pequeños en mas grande o dividiendo grandes clusters en otros mas pequeños. La estructura jerárquica es representada en forma de un árbol y es llamada **Dendograma**.

Se dividen en dos tipos:

- **Algoritmos jerárquicos aglomerativos** (bottom-up, inicialmente cada instancia es un cluster). AGNES
- **Algoritmos jerárquicos divisivos** (top-down, inicialmente todas las instancias estan en un solo cluster). DIANA.

Dendrograma

Cluster Dendrogram

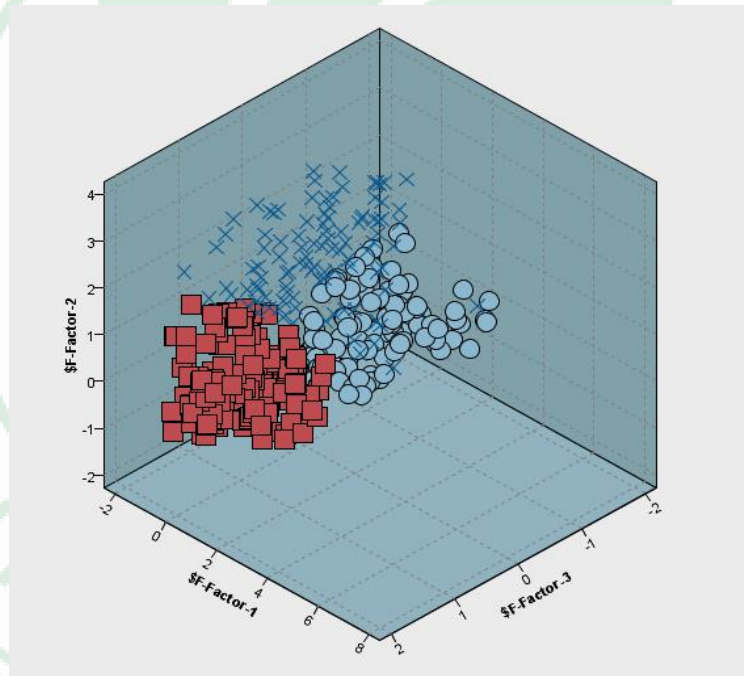


Dendograma

- Los dendrogramas son fáciles de interpretar pero pueden conducir a falsas conclusiones por las siguientes razones:
- El dendograma correspondiente a un conglomerado jerárquico no es único, puesto que por cada junte de clusters (merge) uno necesita especificar que sub-árbol va a la derecha y cuál a la izquierda.
- La estructura jerárquica del Dendograma no representa con certeza las verdaderas distancias entre los objetos distintos del conjunto de datos.

CLUSTER NO JERARQUICOS

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:

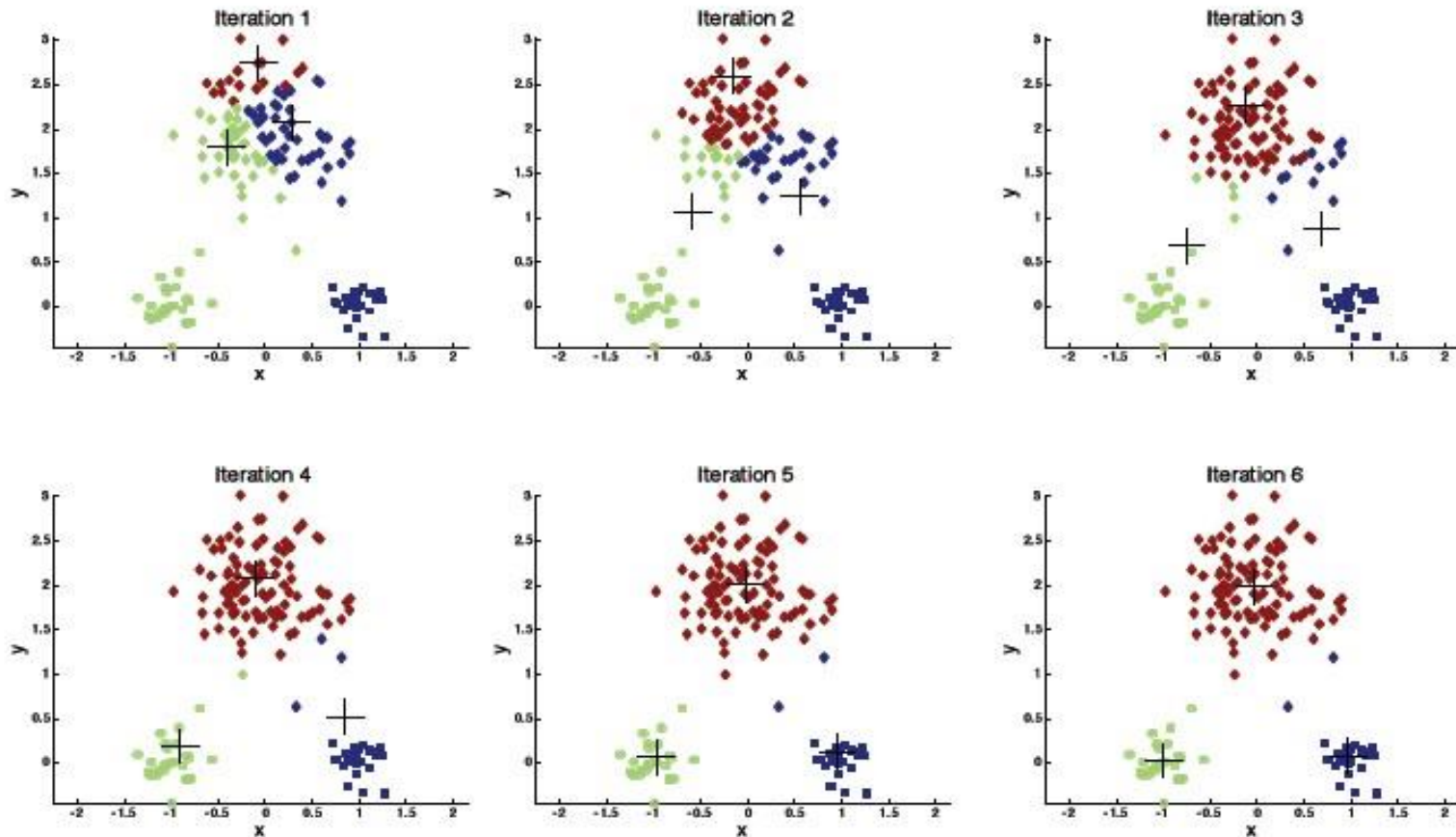


Método de K – Means : Proceso

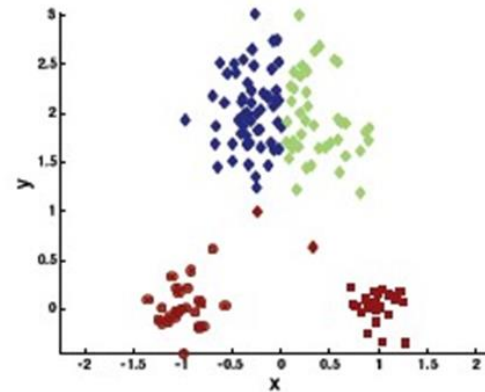
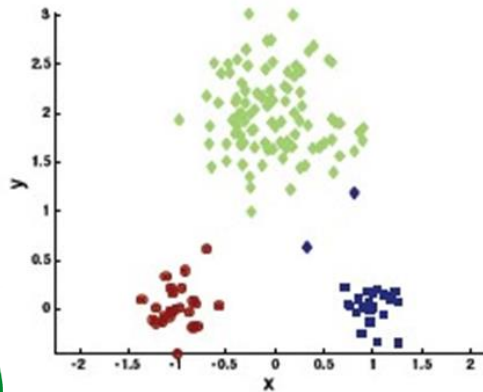
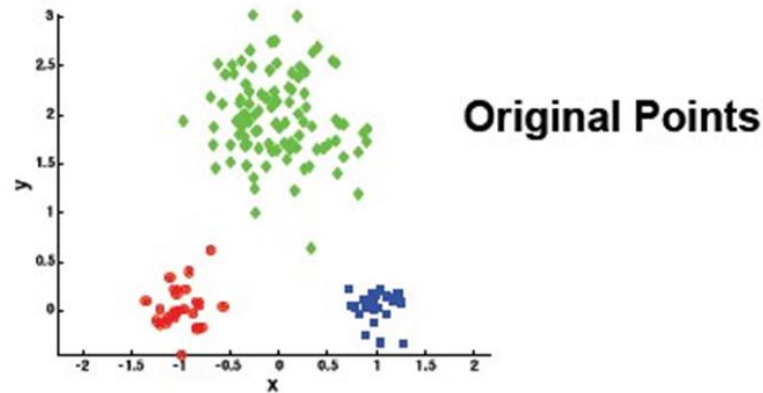
1. Representar una clase por su centro de gravedad, esto es, por su vector de promedios.
2. Asignar los objetos a la clase del centro de gravedad más cercano.



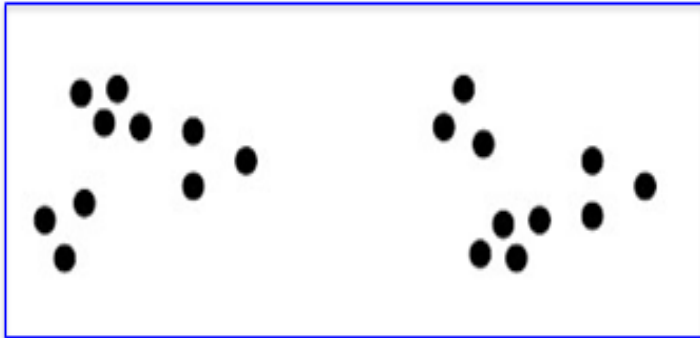
Método de K - Means



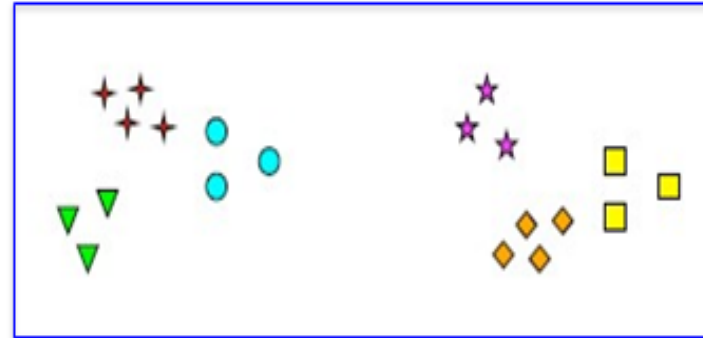
Método de K - Means



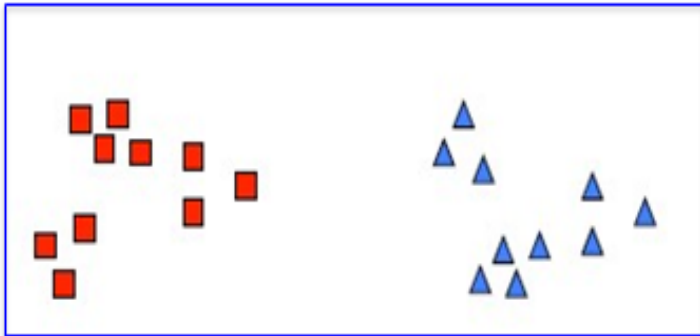
¿ Cuántos clústeres?



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} .





¡Gracias!

TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”