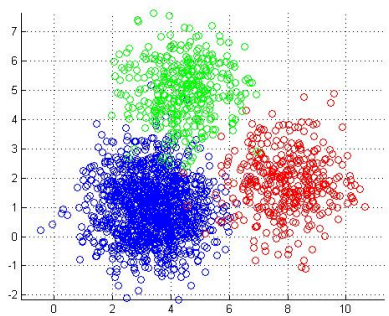
AGRUPACIÓN

Elizabeth León Guzmán

Research Group on Data Mining – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

2021



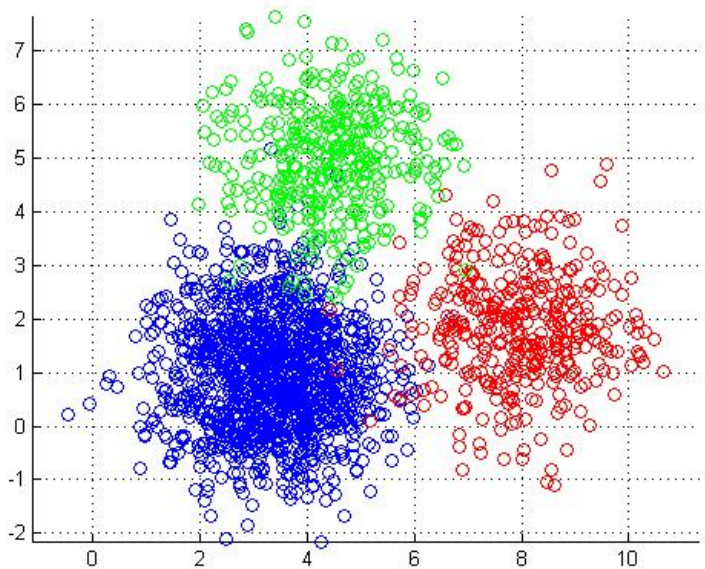
Agenda

Agenda

Definición

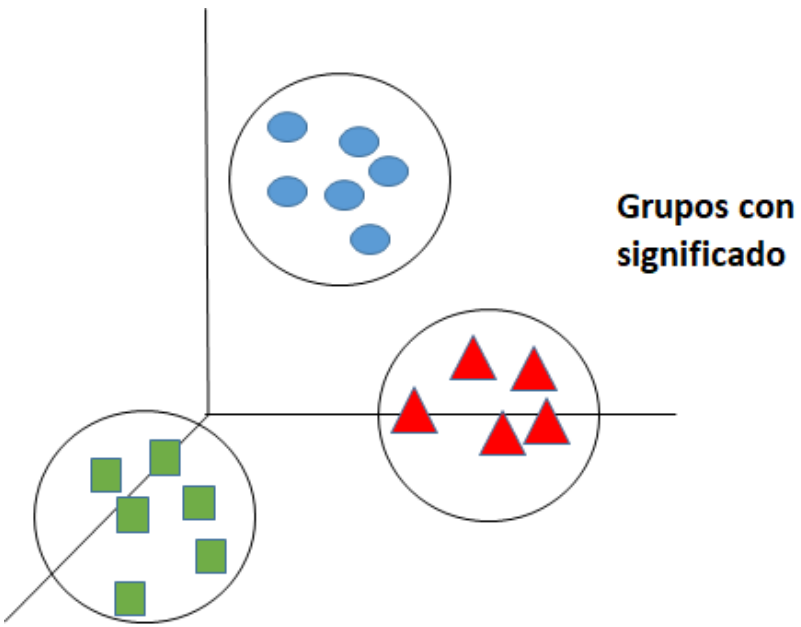
Proceso para

- Dividir los datos en grupos (clusters), de tal forma que los grupos capturan la **estructura natural** de los datos
- Dividir datos en grupos (clusters) de tal forma que datos que pertenecen al mismo grupo son **similares**, y datos que pertenecen a diferentes grupos son diferentes



Definición

- Los grupos con significado (clases) indican como las personas analizan y describen el mundo
- Los humanos tienen la habilidad de dividir los objetos en grupos (agrupación) y asignar objetos particulares a esos grupos (clasificación). Ej: los niños dividen objetos en fotografías: edificios, vehículos, gente, animales, plantas



Definición

El proceso de agrupar tiene en cuenta todas las variables implicadas.

Ejemplo

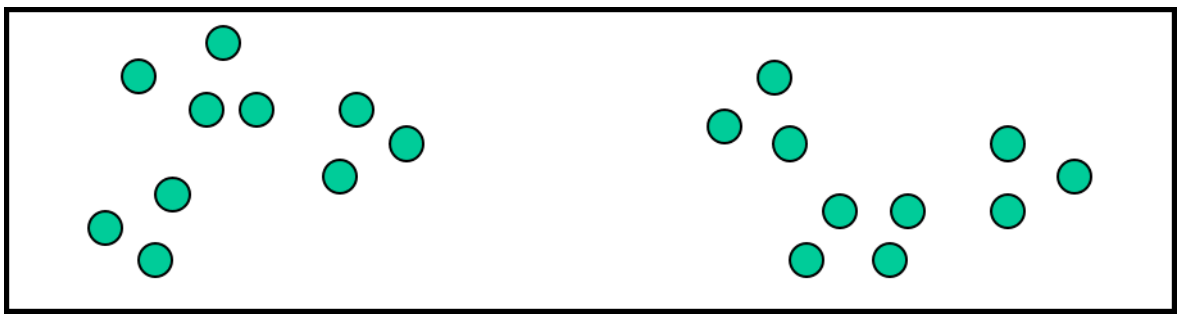
Agrupar los estudiantes de un curso de acuerdo a la similitud entre ellos.
Variables: edad, genéro, profesión, altura, peso, promedio académico, semestre, ciudad, colegio, asistencia, etc.
"Cada grupo corresponde a un perfil de estudiantes"

Características

- La arbitrariedad en el número de clusters es el mayor problema en clustering.
- Sistema visual del humano (espacio Euclideano).
- Grupos tienen diferentes formas, tamaños en un espacio n-dimensional
- Definición de cluster es impreciso y la mejor definición depende de la naturaleza de los datos y de los resultados deseados
- Aprendizaje NO supervisado

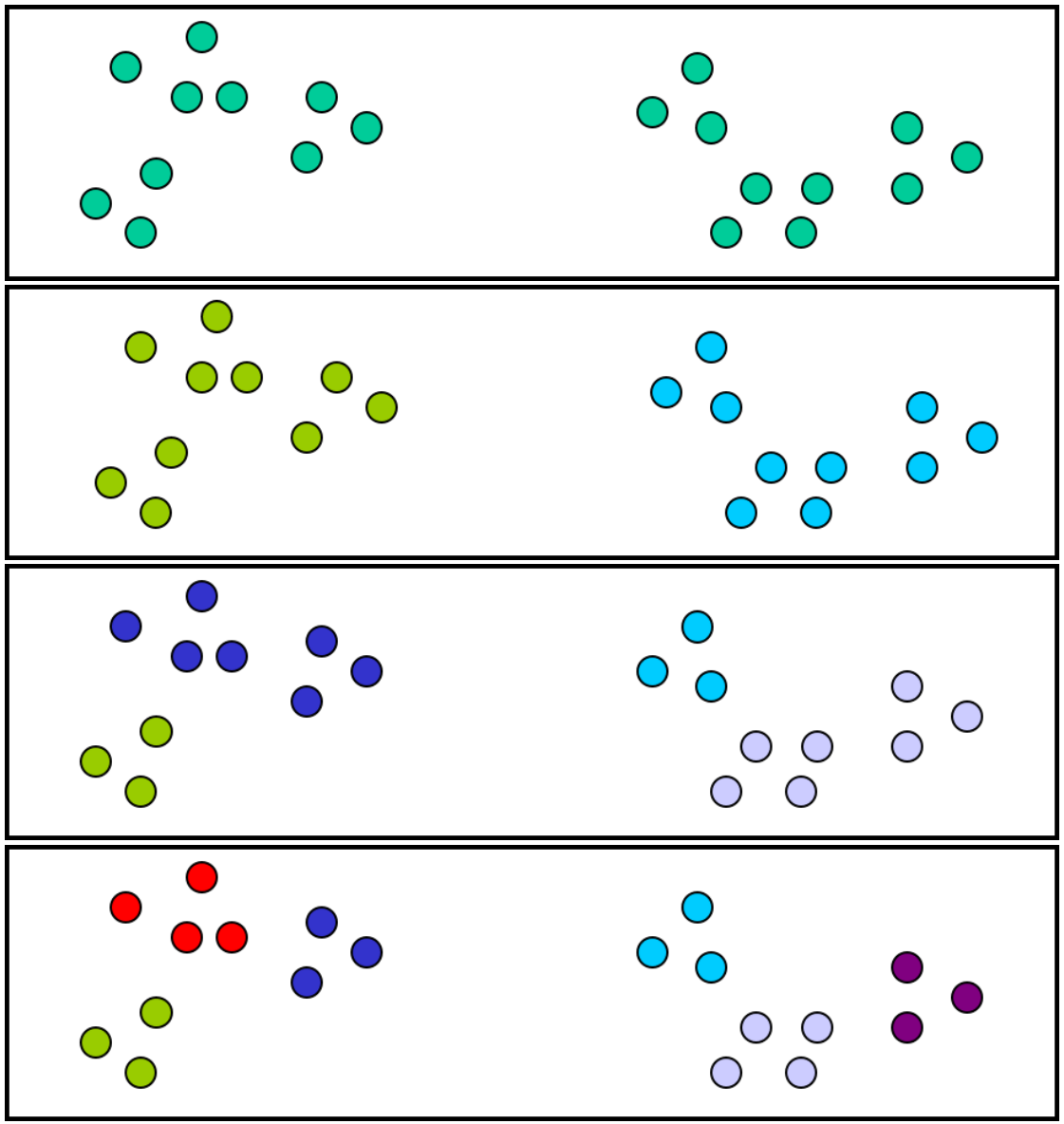
Arbitrariedad en número de "clusters"

¿Cuántos grupos (clusters) se deben encontrar?



Diferentes formas de agrupar el mismo conjunto de datos

Arbitrariedad en número de "clusters"



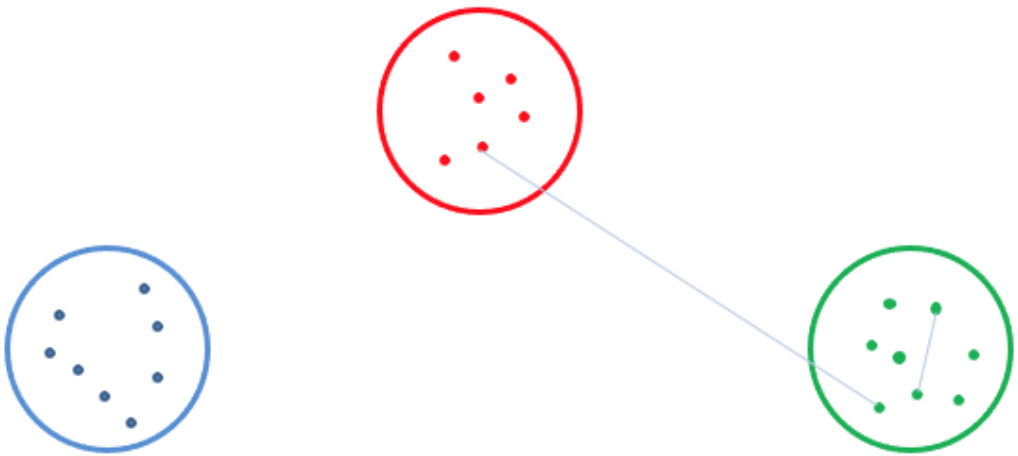
Tipos de Grupos/Clusters

- Clusters bien separados
- Clusters basados en el centro
- Clusters contiguos
- Clusters basados en densidad
- De propiedad o Conceptual
- Descrito por una Función Objetivo

Clusters bien separados

Un cluster es un conjunto de puntos en el que cualquier punto en el cluster es más cercano a cualquier otro punto en el cluster que cualquier otro punto que no esté en el cluster

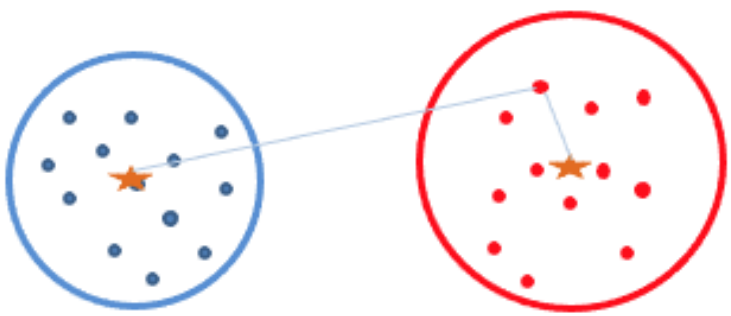
- La distancia entre dos puntos cualquierea de diferentes grupos es mas grande que la distancia entre dos puntos del mismo grupo.
- En ocasiones se usa un *umbral* para indicar que todos los objetos en un grupo deben ser cercanos



Clusters basados en un prototipo (centro)

Un cluster es un conjunto de objetos en el que un objeto está más cerca al prototipo (generalmente el centro del cluster), que al prototipo de otro cluster.

- Para datos continuos, el centro de un cluster frecuentemente es llamado **centroide** que corresponde con el promedio de todos los puntos del cluster.
- Para datos categoricos, el prototipo es el **medoid**, corresponde con el punto más representativo del cluster.
- Por lo general, el prototipo es el punto más central, por lo que los clusters corresponden a clusters *basados en el centro*

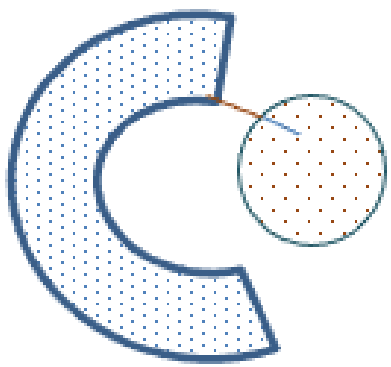


Clusters basados en grafos

Los datos son representados como un grafo, donde los nodos son los objetos y los enlaces representan las conecciones entre los objetos. Un grupo de objetos estan conectados entre sí, pero no estan conectados con objetos fuera del cluster.

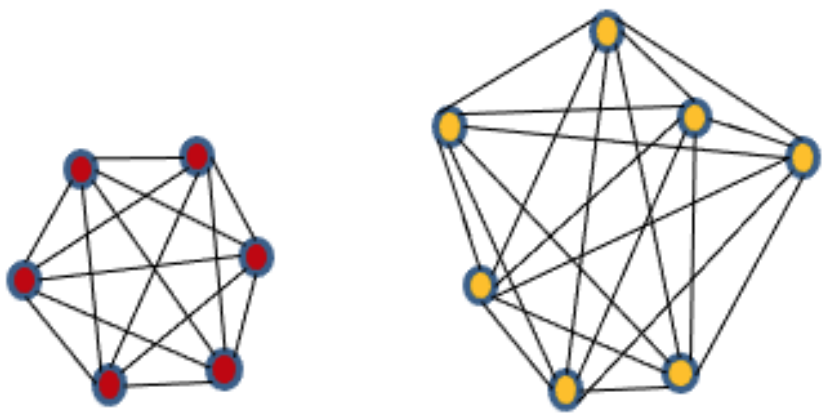
Clusters basados en contiguidad

Un cluster es un conjunto de puntos donde cada punto en el cluster está más próximo al menos a otro punto en el cluster que a cualquier otro punto que no pertenezca al cluster.



Clusters basados en grafos

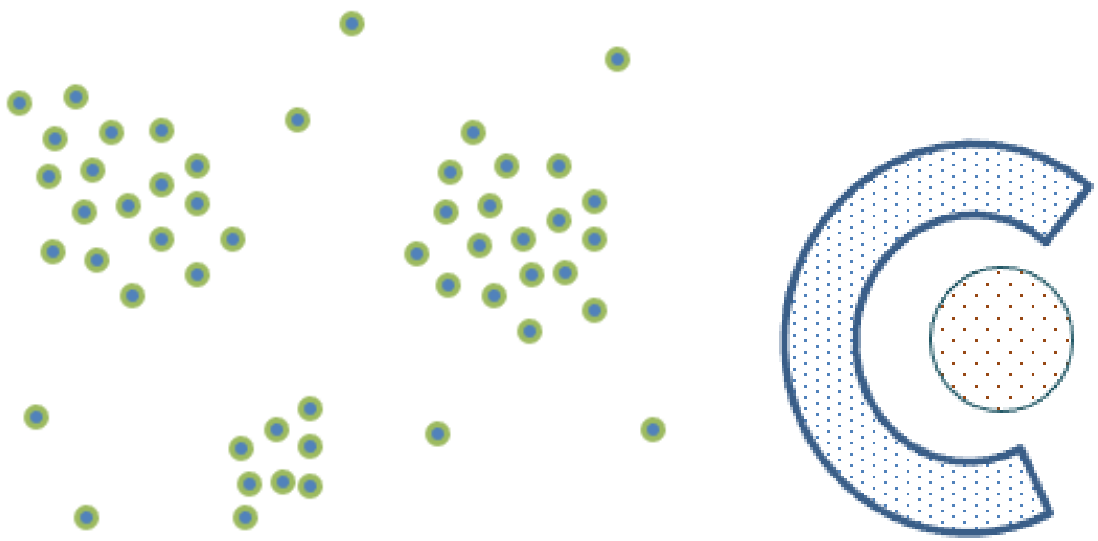
- Puede haber problemas cuando hay ruido.
- Los cluster pueden ser irregulares.
- Clique: conjunto de nodos en un grafo que estan completamente conectados con cada uno. Adicionar conecciones entre los objetos en order con su *distancia*, un cluster es formado cuando un conjunto de objetos que formen un *clique*.



Clusters basados en densidad

Un cluster es una región **densa** de puntos, separados por regiones de baja densidad, de otras regiones de alta densidad.

- Se usan algoritmos basados en densidad para identificar clusters irregulares o entrelazados, y
- cuando el conjunto de datos presenta ruido y datos atípicos.



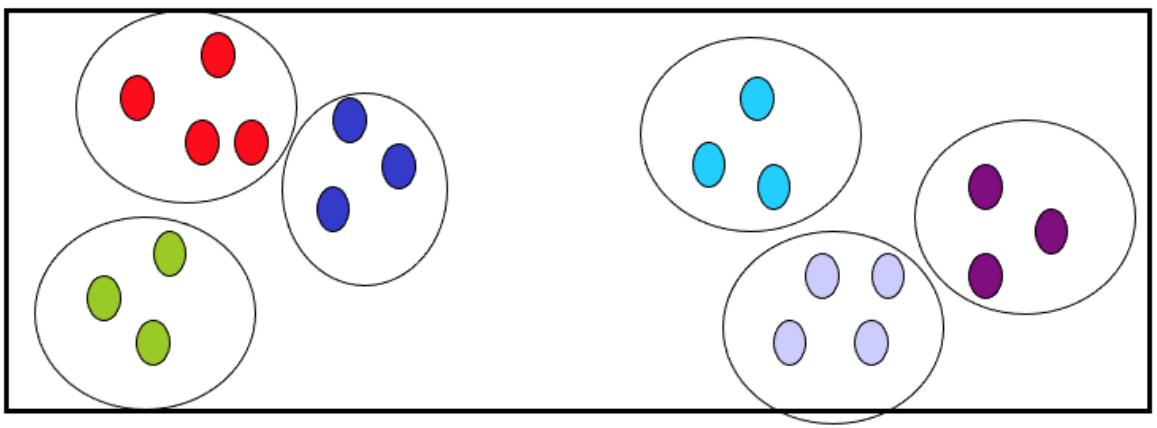
Tipos de Agrupación

- **Agrupación Particional:** Separa los datos (puntos, objetos, registros) en grupos no superpuestos, donde cada dato (punto, objeto, registro) pertenece a un único grupo.
- **Agrupación Jerarquica:** Organiza los datos (puntos, objetos, registros) en grupos sobrepuestos en forma de árbol. Usa estructura de árbol o dendograma.

- **Exclusivo:** Cada objeto es asignado a un solo grupo.
- **Sobrelape o No exclusivo:** Un onjeto puede ser asignado a más de un grupo.
- **Difuso (fuzzy)** Cada objeto pertenece a todos los grupos con un peso o grado de pertenencia (de 0 a1), 1 si pertencece absolutamente al grupo y 0 si absolutamente no pertenece.

Agrupación Particional

Asigna los datos en grupos donde cada dato pertenece a un único grupo, minimizando la distancia en cada uno de los grupos y/o maximizando la distancia entre los grupos.



Kmeans

- Agrupamiento particional
- Cada cluster está asociado con un centroide (valor de la media del cluster)
- Cada punto es asignado al cluster más cercano al centroide
- El número de grupos "K" debe ser especificado como parámetro

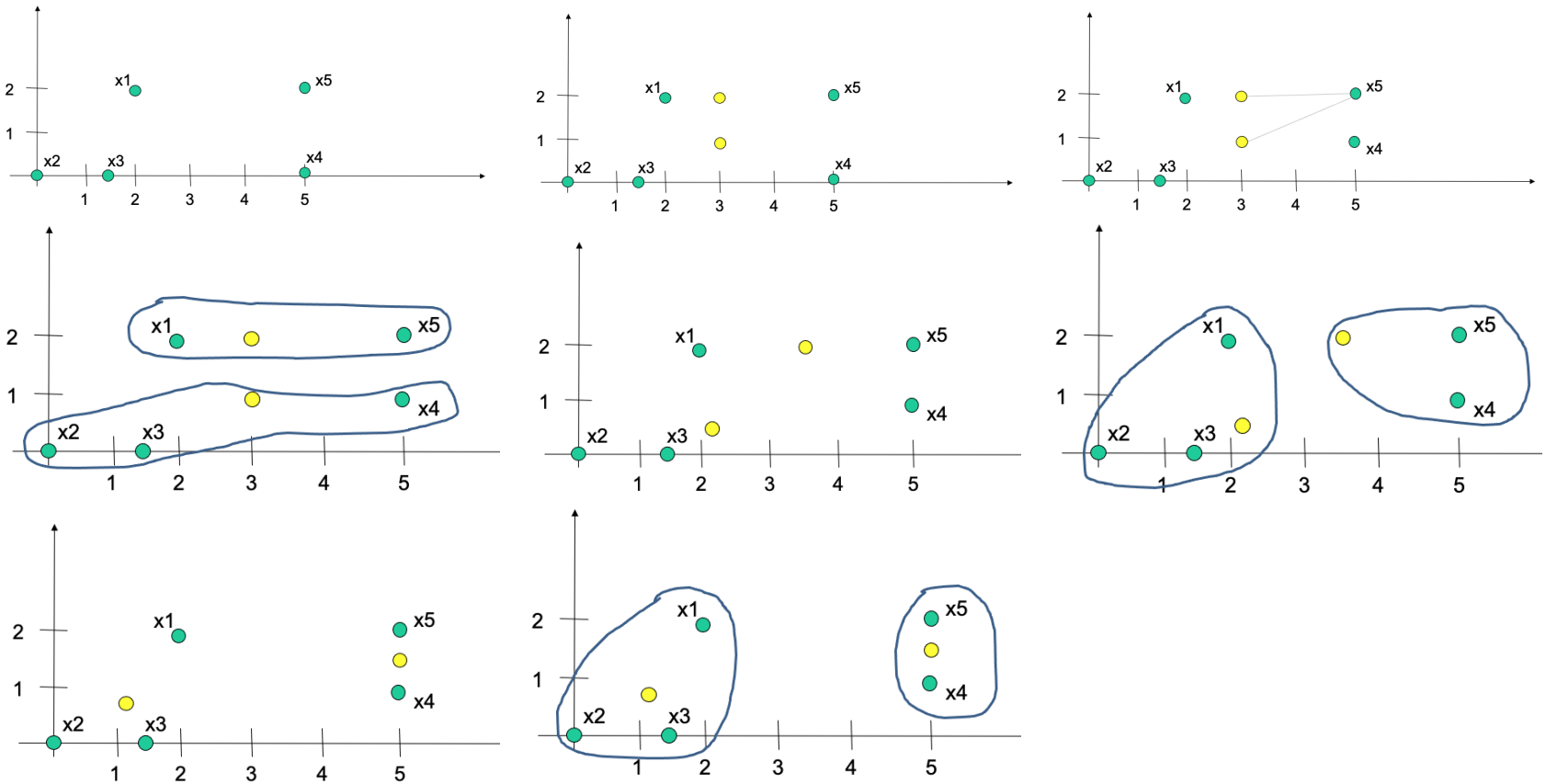
El algoritmo básico es muy simple:

Algoritmo KMeans

1: Seleccionar K puntos como centroides iniciales
2: **Repetir**
3: Asignar los puntos del dataset al centroide más cercano
4: Recalcular el centroide de cada grupo
5: **Hasta** que el centroide no cambie

Kmeans

Ejemplo: Algoritmo KMeans



Kmeans

- Los centroides iniciales se escogen aleatoriamente.
- La proximidad es medida por la distancia Euclidiana, la similitud por coseno, correlación, etc.
- La función objetivo (que mide la calidad del grupo) es la *Suma del error cuadrático SSE (**S**um of the **S**quared **E**rror)*

  $$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

  El centroide que minimiza el SSE es el *centroide (mean)* definido como: $c_i = \frac{1}{m_i} \sum_{x \in C_i} x$
- La mayoría de la convergencia ocurre en las primeras iteraciones (mínimo local).

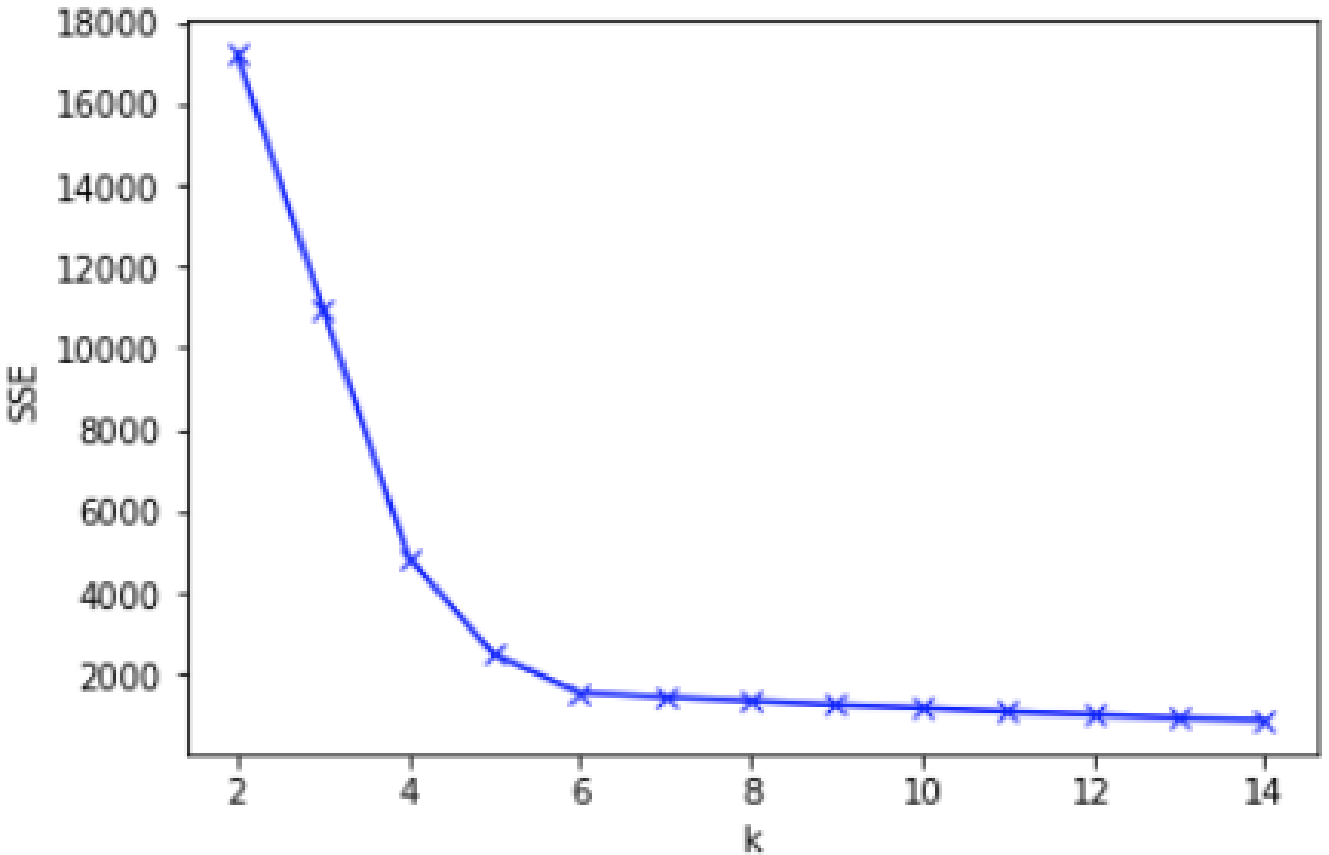
Kmeans

- Número K?? (número de clusters)
- Sensible a inicialización
- Sensible a ruido y outliers (afectan la media (mean))
- Problema de optimización: minimizar el error cuadrático
- Variación: k-mediods:No usa la media, usa el objeto mas centrado (mediod). Menos sensible a ruido y outliers

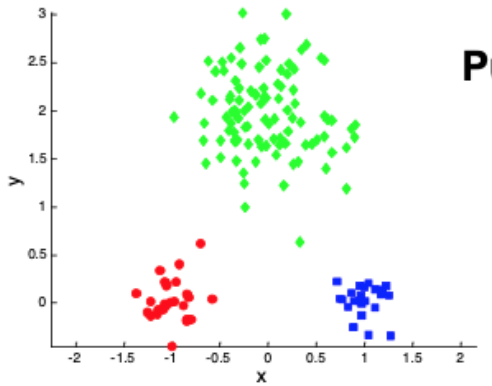
Kmeans

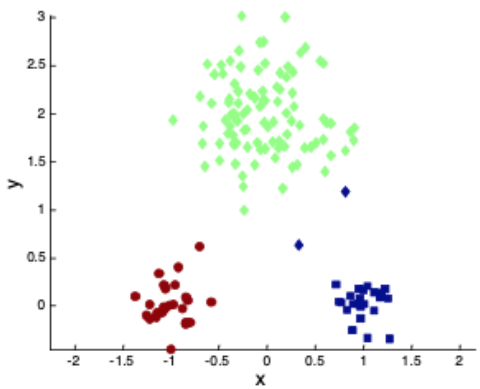
Determinar el número de grupos K

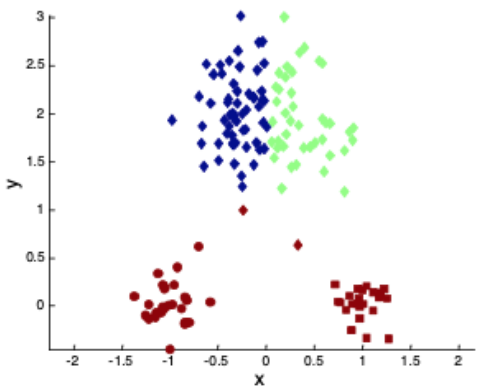
Gráfica "del codo"



Kmeans



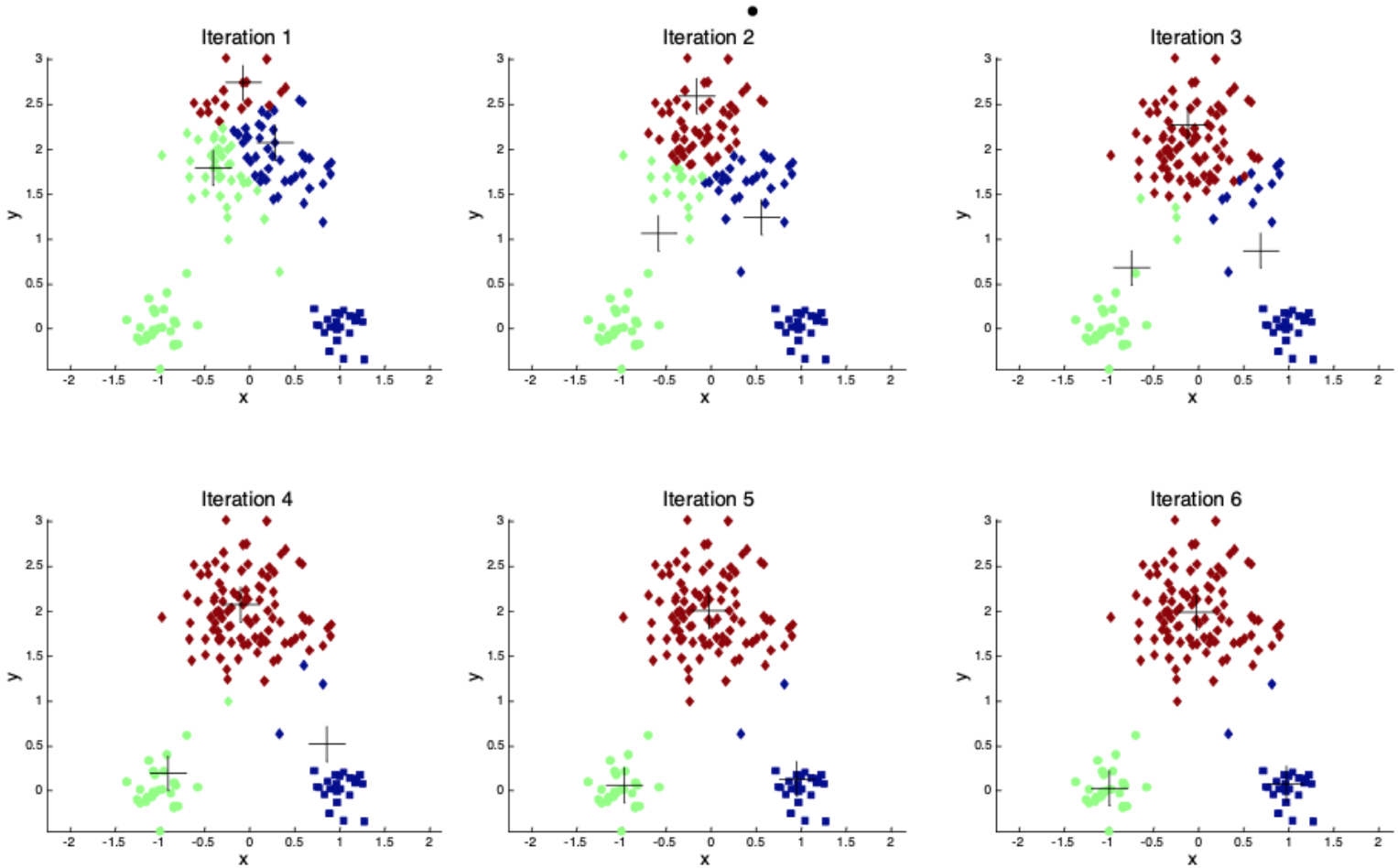




Source: Tan et al. (2005)

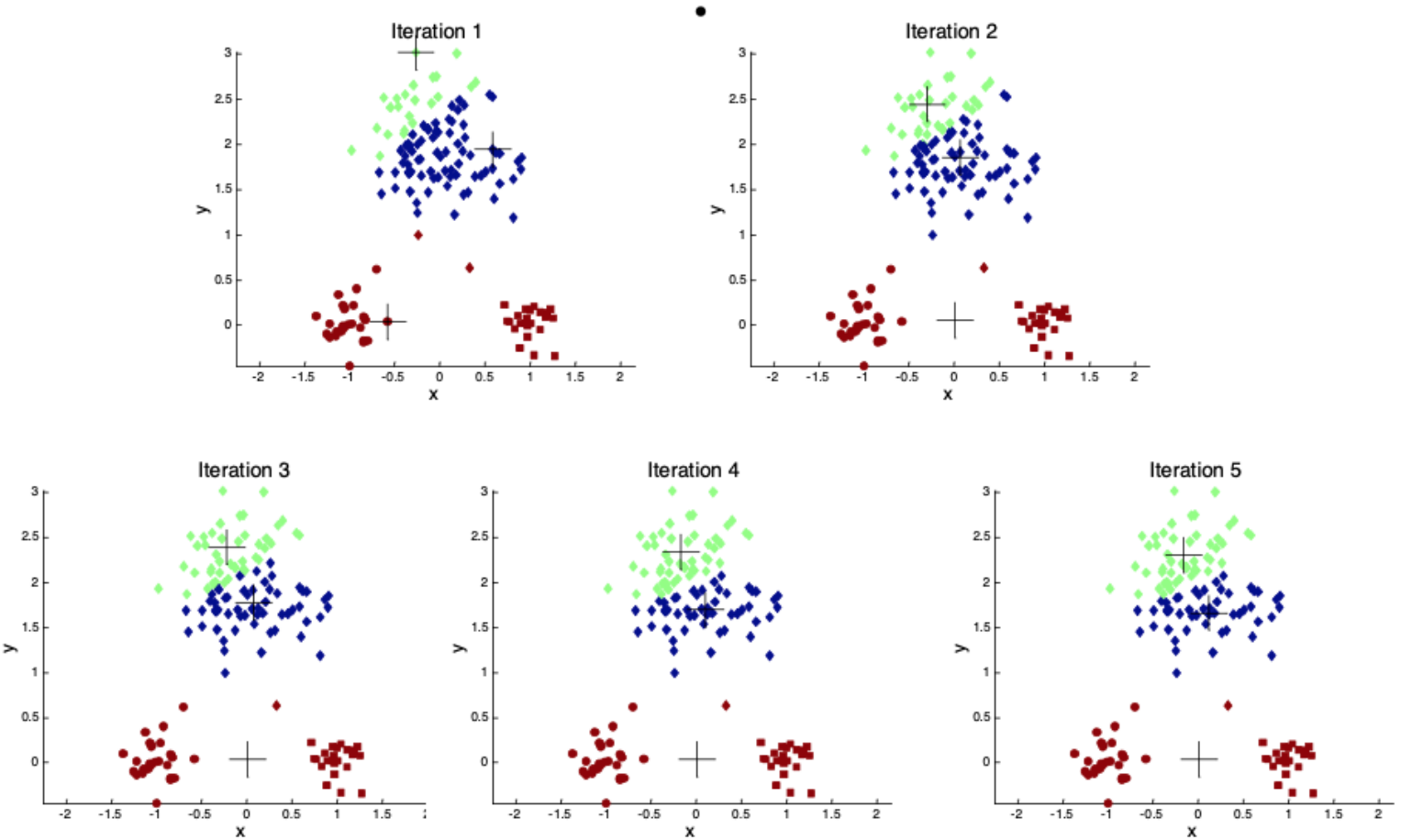
Kmeans

¡La inicialización es importante!



Kmeans

¡La inicialización es importante!



Kmeans

Soluciones para el problema de inicialización

- Múltiples ejecuciones ayuda, pero la probabilidad no está de su lado
- Agrupamiento de prueba y agrupamiento jerárquico para determinar los centroides iniciales
- Seleccionar mas de un K inicial de centroides y luego seleccionar entre estos los centroides iniciales (Seleccionarlos ampliamente separados)
- Bisectar K-means

Kmeans

Bisecting Kmeans

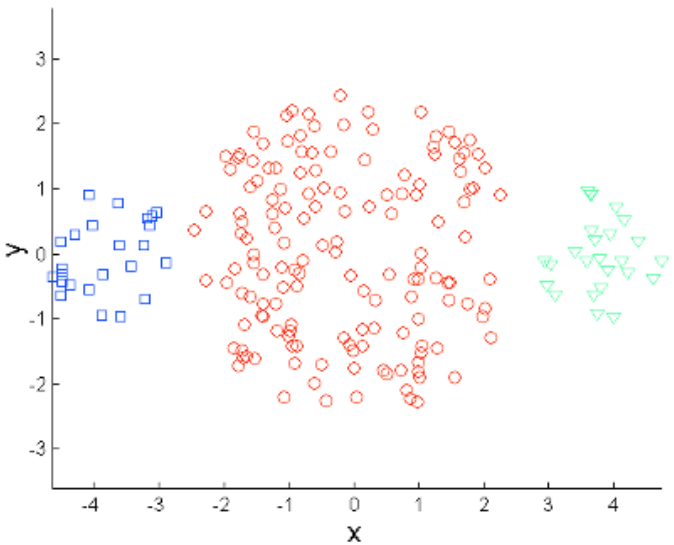
1:Inicializar lista de grupos solucion de tamaño K
2:**repeat**
3: Encontrar dos grupos usando kmeans
4: Seleccionar uno de los grupos para dividir
5: Incluir el otro grupo a la lista de grupos solución
6:**until** La lista de grupos contiene K grupos

- Kmeans (línea 3) puede repetirse varias veces hasta encontrar una solución con bajo SSE.
- Diferentes formas de seleccionar el grupo a dividir (línea 4): más grande, mayor SSE, o basado en los dos (tamaño y SSE).

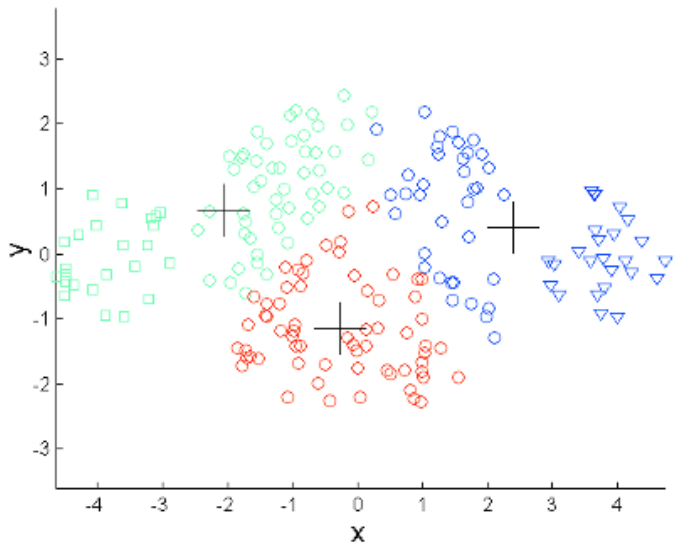
Se utiliza como inicialización para el Kmeans, ya que está usando el k-means localmente (divide grupos individuales).

Kmeans

Diferentes tamaños



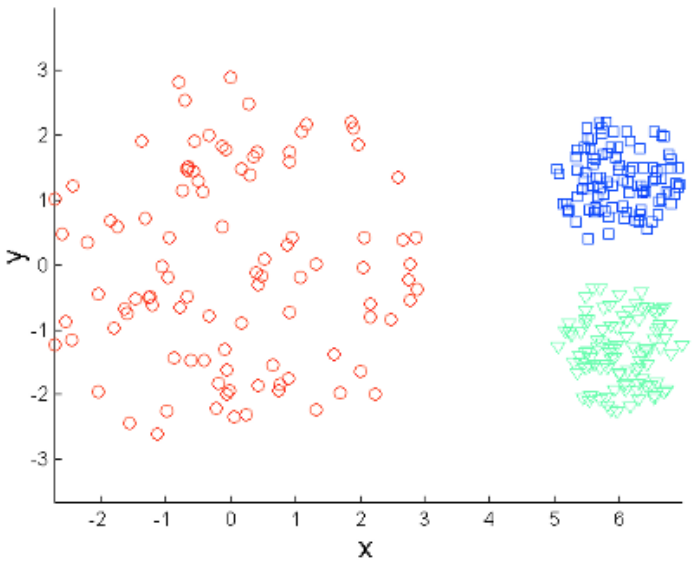
Puntos originales



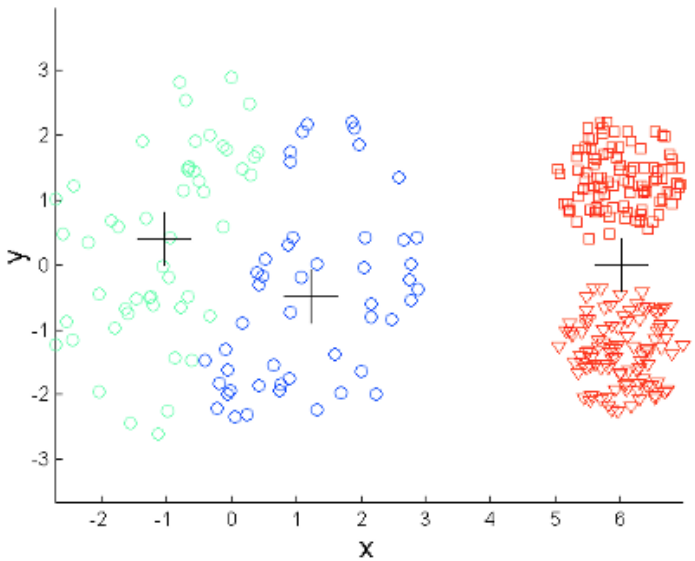
K-means (3 Clusters)

Kmeans

Diferentes densidades



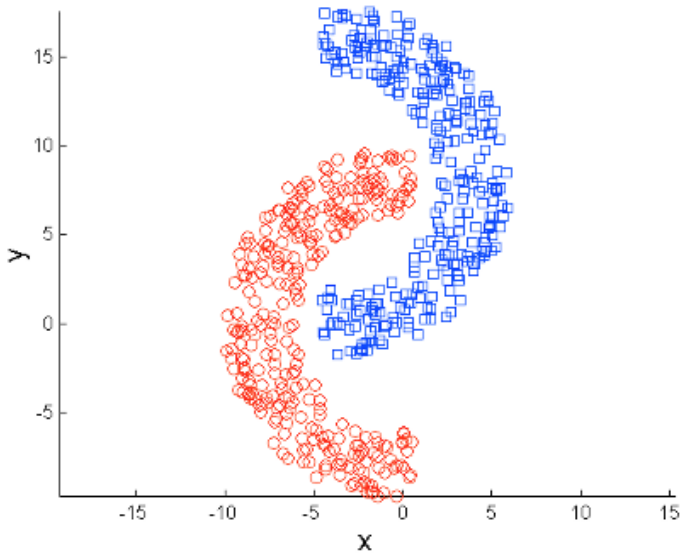
Puntos originales



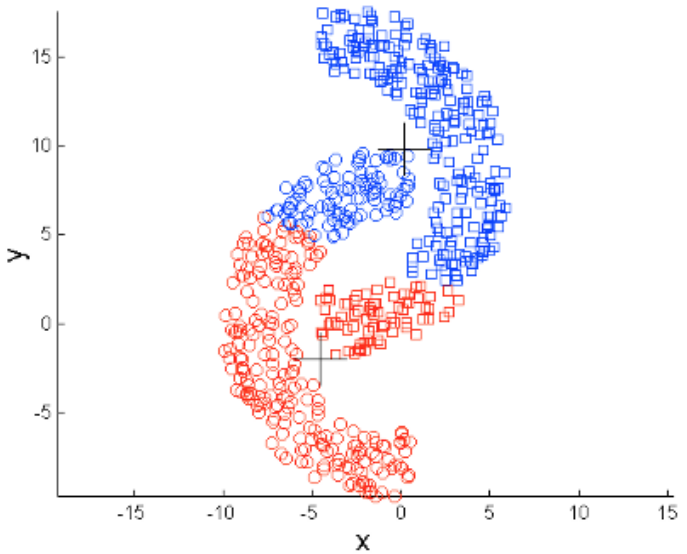
K-means (3 Clusters)

Kmeans

Diferentes formas



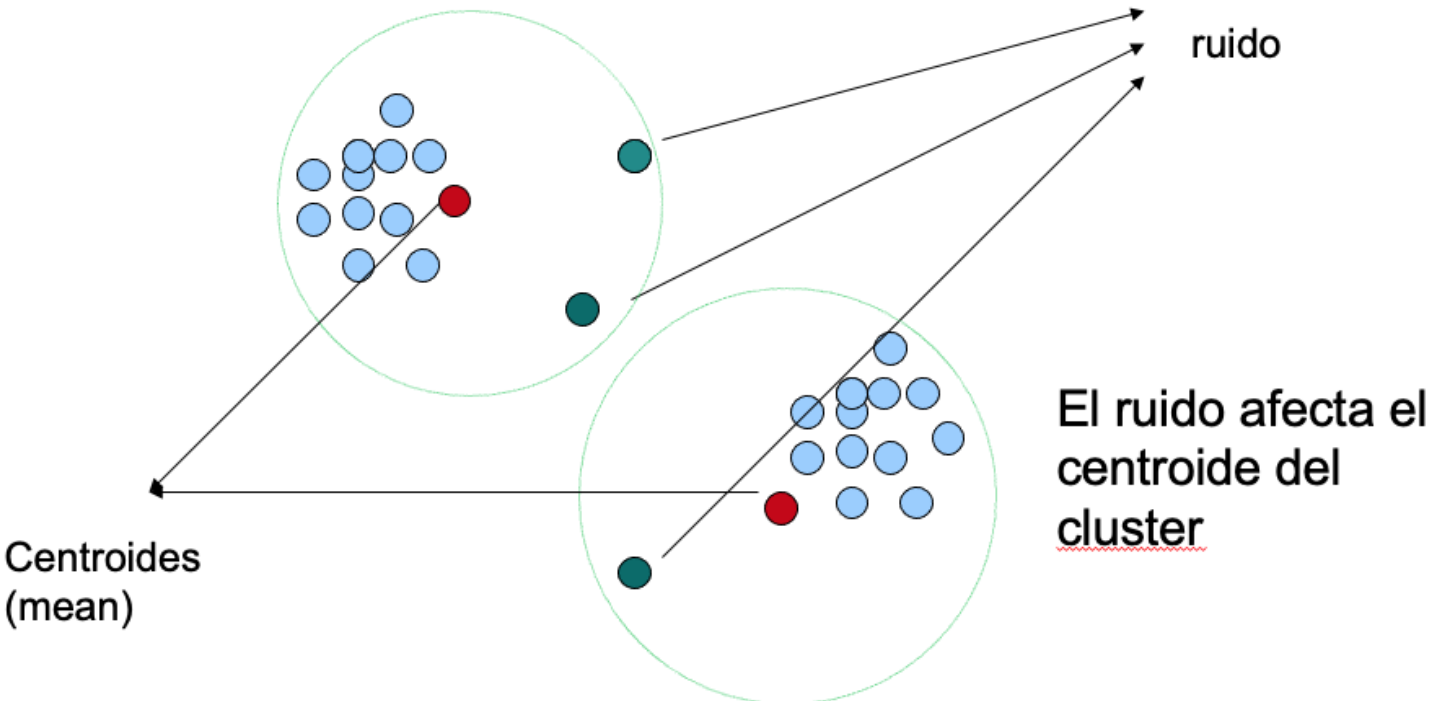
Puntos originales



K-means (2 Clusters)

Kmeans

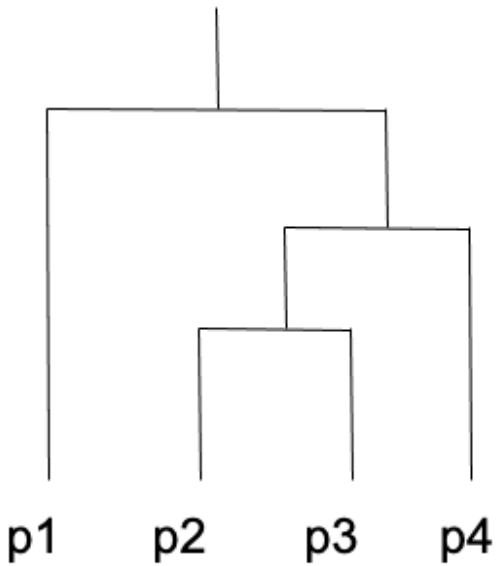
Ruido, outliers



Agrupación Jerárquica

Junto con el K-means son algoritmos de "clustering" pioneros que aún siguen siendo utilizados ampliamente.

- NO se especifica el número de clusters
- Proceso iterativo que une o divide los grupos
- La salida es una jerarquía de clusters. Generalmente desplegada en un dendograma



Agrupación Jerárquica

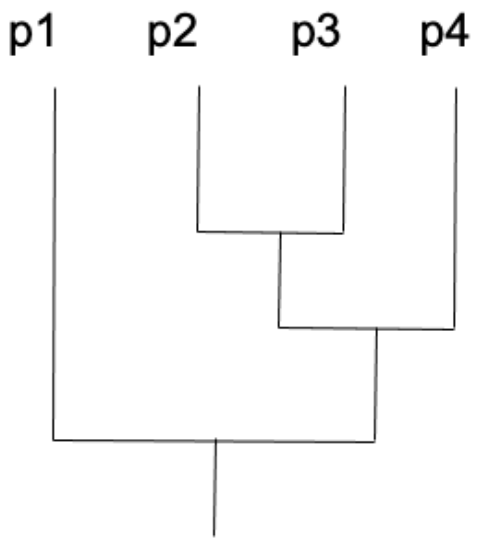
Los algoritmos de este tipo se dividen en dos clases:

- **Aglomerativos**
- **Divisibles**

Algortimos Jerárquicos

Aglomerativo

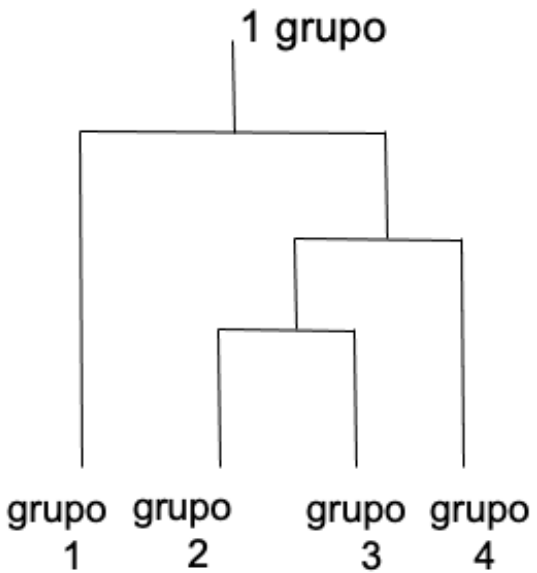
Inicialmente todos los puntos (objetos) son grupos individuales (tamaño 1), en cada iteración, los grupos más cercanos se unen para formar un solo grupo, al final de todas las iteraciones se tiene un solo cluster.



Algortimos Jerárquicos

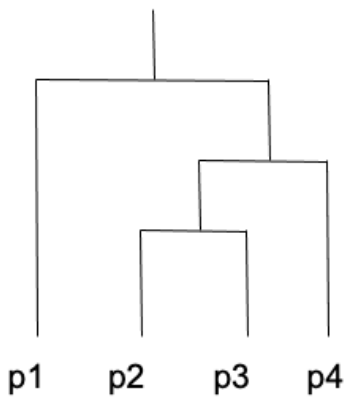
Divisible

Inicialmente solo existe un cluster (contiene todos los objetos), en cada paso, se van dividiendo los clusters hasta que cada objeto es un solo cluster.



Algortimos Jerárquicos Aglomerativos

- El más común
- El dendograma despliega la relación entre el cluster y sub-cluster, y el orden en el cual los clusters fueron fusionados



Dendograma

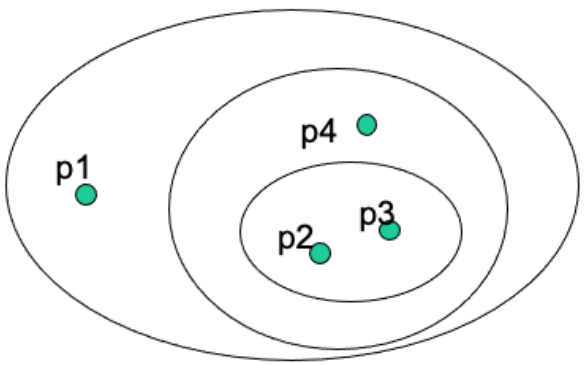


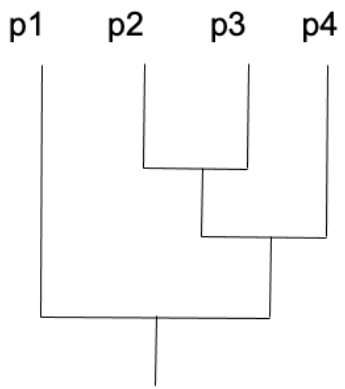
Diagrama de clusters anidados

Algortimos Jerárquicos Aglomerativos

Algoritmo Básico

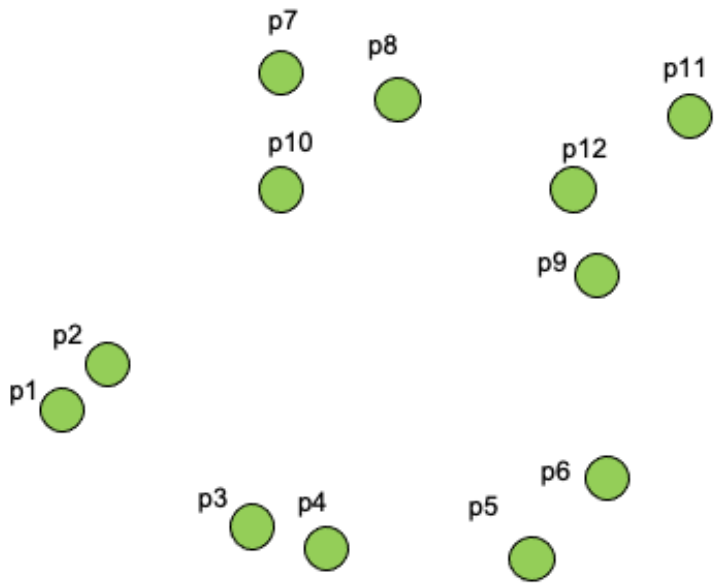
1: Cada punto es un grupo
2: Calcular matriz de proximidad
3: **Repetir**
4: Unir los dos grupos más cercanos
5: Actualizar la matriz de proximidad
6: Hasta que solo quede un grupo

El cálculo de la proximidad entre dos clusters es clave en el proceso de fusión o unión de dos grupos.



Algortimos Jerárquicos Aglomerativos

Ejemplo

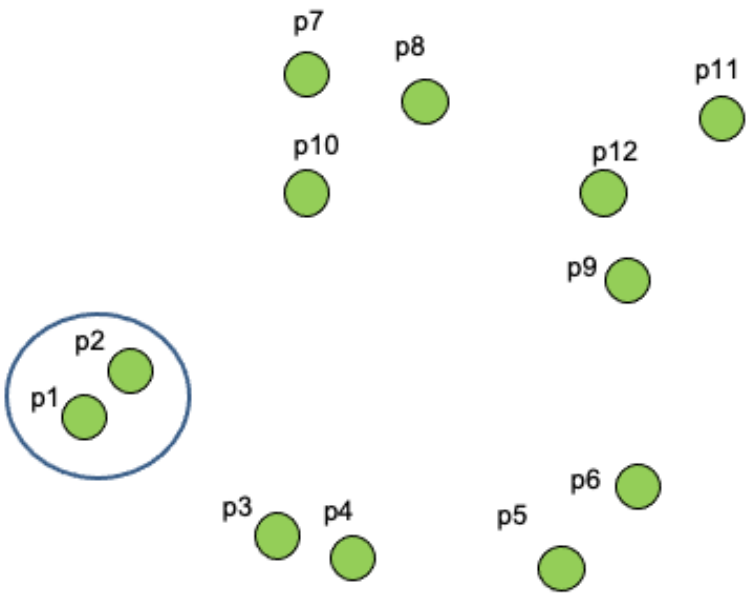


	p1	p2	p3	p4	p5	. . .	p12
p1							
p2							
p3							
p4							
p5							
. . .							
p12							

Matriz de proximidad

Algortimos Jerárquicos Aglomerativos

Ejemplo

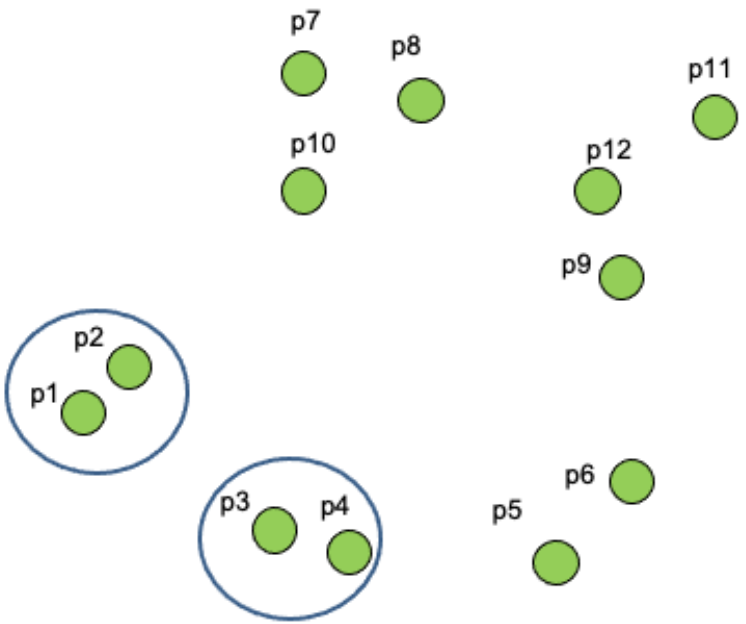


	c1	p3	p4	p5	. . .	p12
c1						
p3						
p4						
p5						
p6						
. . .						
p12						

Matriz de proximidad

Algortimos Jerárquicos Aglomerativos

Ejemplo

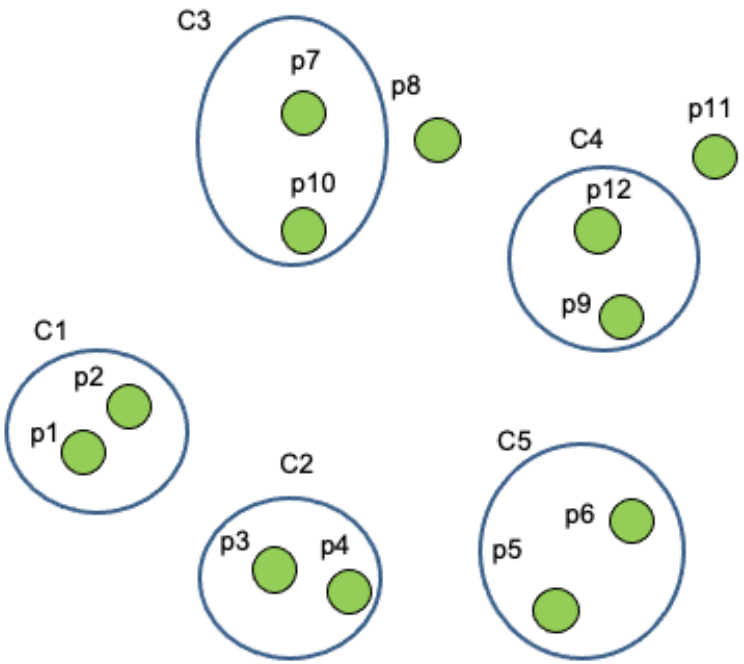


	c1	c2	p5	p6	. . .	p12
c1						
c2						
p5						
p6						
p7						
.						
.						
p12						

Matriz de proximidad

Algortimos Jerárquicos Aglomerativos

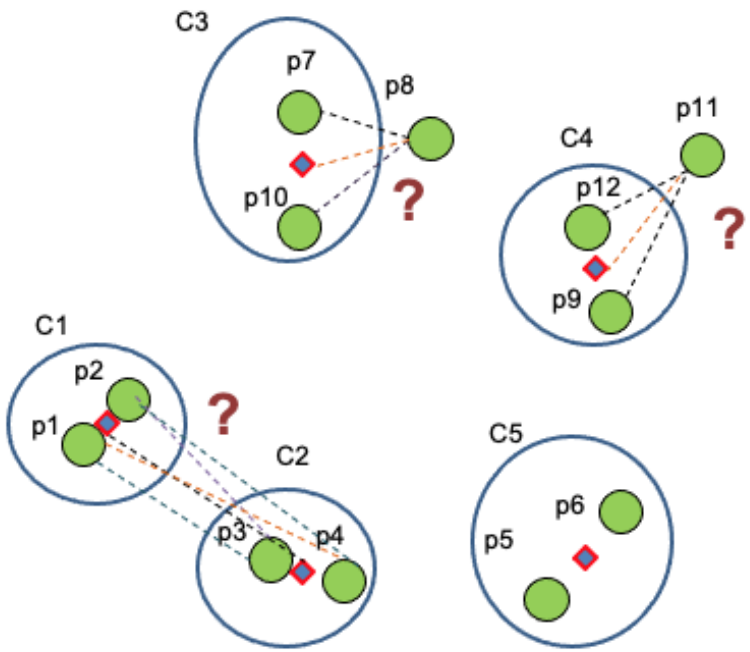
Ejemplo



	c1	c2	c3	c4	c5	p8	p12
c1							
c2							
c3							
c4							
c5							
p8							
p12							

Algortimos Jerárquicos Aglomerativos

Ejemplo



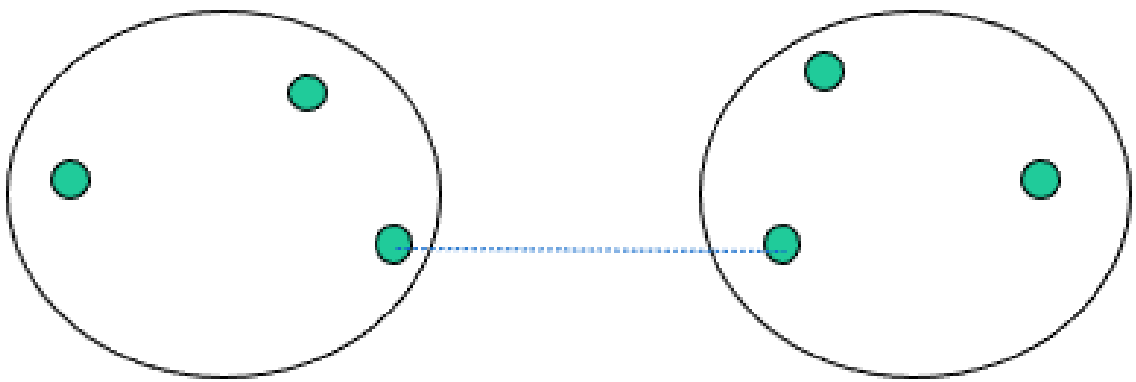
	c1	c2	c3	c4	c5	p8	p12
c1							
c2							
c3							
c4							
c5							
p8							
p12							

Medida intercluster

Proximidad Intercluster

Single Link(Enlace Simple) - MIN

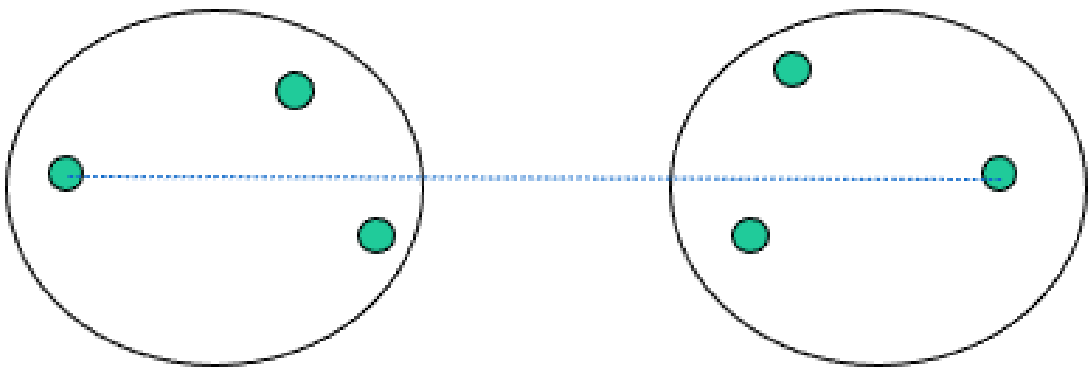
La proximidad de los clusters esta definida como la proximidad entre los **puntos más cercanos** que están en diferentes clusters.



Proximidad Intercluster

Complete Link(Enlace Completo) - MAX

La proximidad de los clusters esta definida como la proximidad entre los **puntos más lejanos** que están en diferentes clusters.

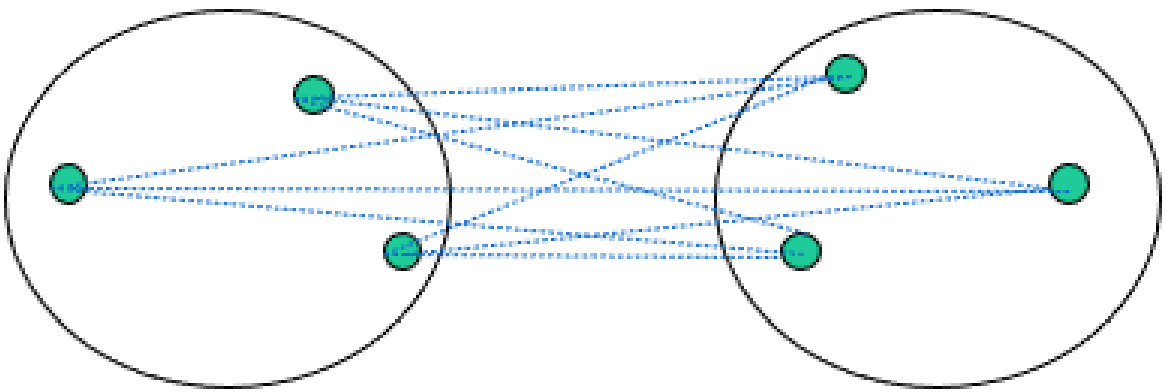


Proximidad Intercluster

Group Average(Promedio)

La proximidad de los clusters esta definida como el **promedio** de todas las proximidades de cada uno de los pares de puntos

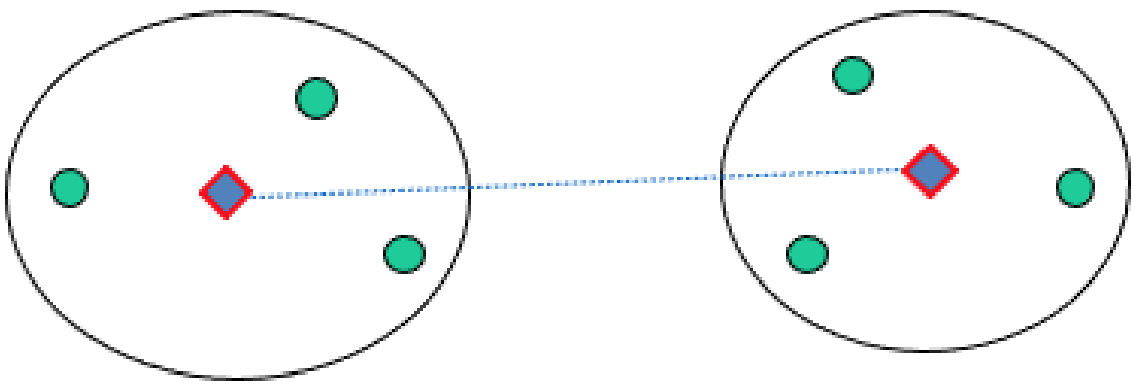
$$Prox(C_a, C_b) = \frac{\sum_{i=1}^{|C_a|} \sum_{j=1}^{|C_b|} prox(p_i, p_j)}{p_i \in C_a \; p_j \in C_b}$$
$$|C_a| * |C_b|$$



Proximidad Intercluster

Prototype(Centroides como prototipos)

Cuando se usan **prototipos** , como el centro, la proximidad de los clusters es la **proximidad entre los centros** de los clusters.



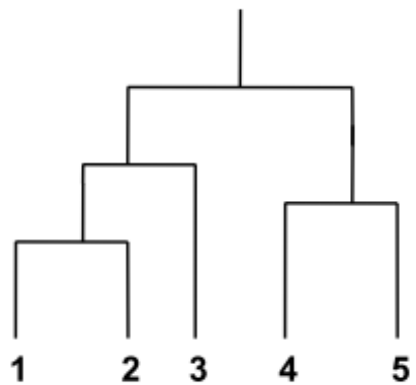
Proximidad Intercluster

Ejercicio

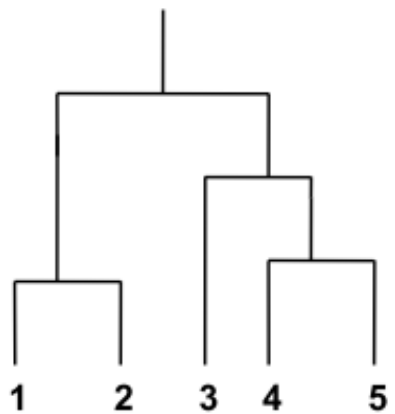
Dada la matrix de similaridad construir el dendograma usando Single Link (MIN) como proximidad intracluster.

	I1	I2	I3	I4	I5
I1	1,00	0,90	0,10	0,65	0,20
I2	0,90	1,00	0,70	0,60	0,50
I3	0,10	0,70	1,00	0,40	0,30
I4	0,65	0,60	0,40	1,00	0,80
I5	0,20	0,50	0,30	0,80	1,00

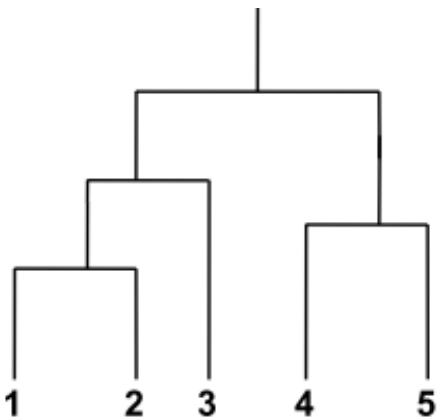
Proximidad Intercluster



MIN



MAX

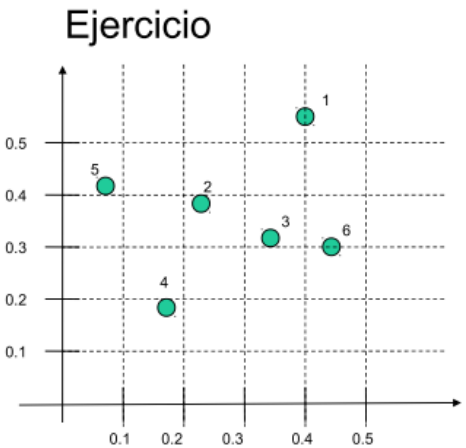


AVG

Proximidad Intercluster

Ejercicio 2

Dada la matrix de distancias construir el dendograma usando Single Link (MIN) como proximidad intracluster.



Coordenadas x,y

Punto	x	y
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Matriz de proximidad usando distancia euclideana

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

eleonguz@unal.edu.co

www.midas.unal.edu.co

References I

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining, (first edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.