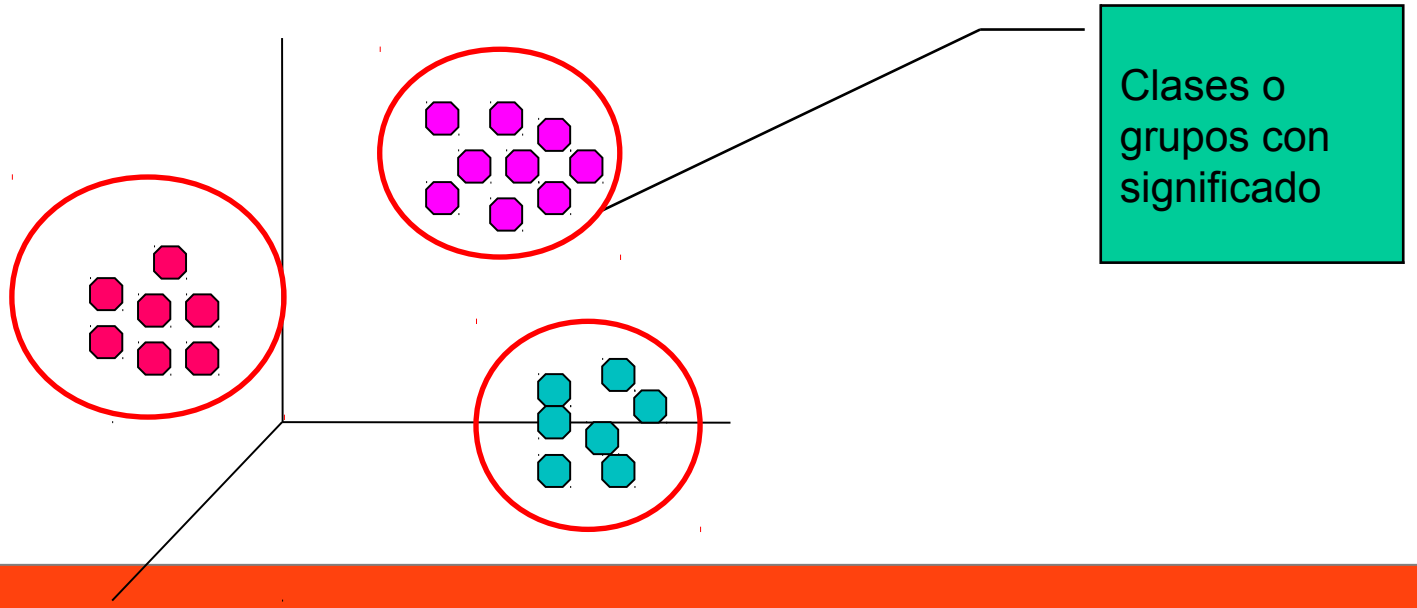
Módulo
Minería de Datos
Diplomado

Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS

Agrupamiento

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



Agrupamiento

- Las clases (grupos con significado) indican como las personas analizan y describen el mundo

- Los humanos tienen la habilidad de dividir los objetos en grupos (agrupamiento) y asignar objetos particulares a esos grupos (clasificación)
- Ej: los niños dividen objetos en fotografías: edificios, vehículos, gente, animales, plantas

- Cluster Análisis (clustering) es el estudio de técnicas para encontrar las clases automáticamente.

Aplicaciones de Agrupamiento

Biología: taxonomia (especies), analisis de información genética(grupos de genes que tienen funciones similares)

Recuperación de Información (Information retrieval): Agrupar resultados de búsquedas en la web (cada grupo contiene aspectos particularres dela consulta) Ej: cine (comentarios, estrellas, teatros)

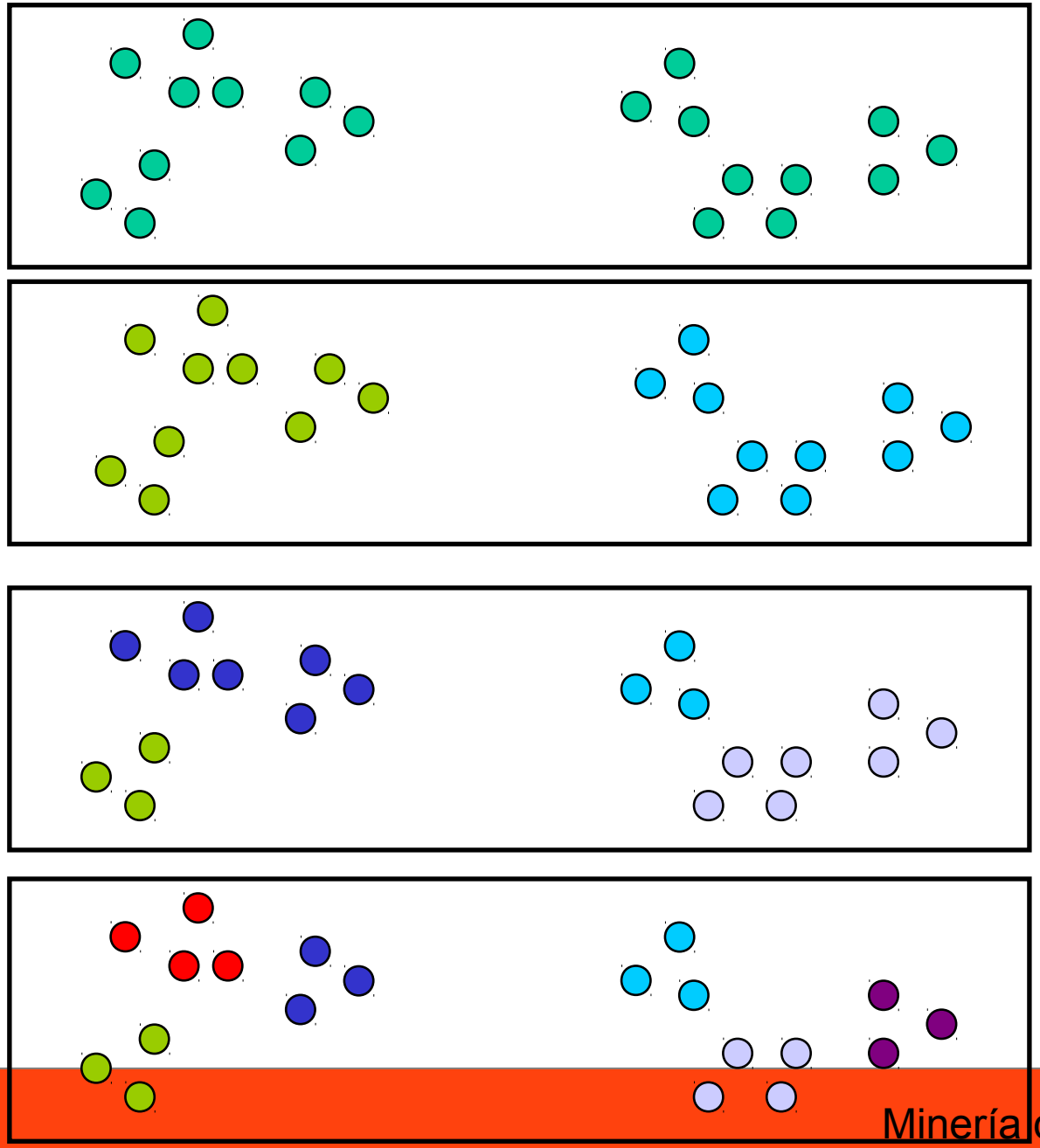
Psicología y Medicina: Agrupar diferentes tipos de depresión, detectar patrones en la distribución temporal de una enfermedad

Aplicaciones de Agrupamiento

Clima: Encontrando patrones en la atmosfera y oceano. Presión atmosférica de regiones polares y areas de el oceano que tienen un impacto significativo en el clima de la tierra.

Negocios: Segmentar los clientes en grupos para un analisis y actividades de mercadeo

Diferentes formas de agrupar el mismo conjunto de datos



Puntos originales

Dos clusters

Cuatro clusters

Seis clusters

Agrupamiento

- Sistema visual del humano (espacio Euclideano)
- La arbitrariedad en el número de clusters es el mayor problema en clustering.
- 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
- Clasificación NO supervisada (contraste con clasificación)

Medidas de similaridad/distancia

- La medida de similaridad es fundamental en la definición del cluster
- Debe ser escogida muy cuidadosamente, ya que la calidad de los resultados dependen de ella
- Se puede usar la disimilaridad (distancia)
- Dependen de los tipos de datos

Similitud y Disimilitud

- **Similitud**
  Medida numérica de semejanza entre objetos
  Valor alto para objetos parecidos
  A menudo definida en el intervalo [0,1]
- **Disimilitud**
  Medida numérica de diferencia entre objetos
  Valor bajo para objetos parecidos
  Varia entre [0, ∞)
  Usualmente es una distancia
- **Proximidad**
  Se refiere a similitud o disimilitud

Similitud y Disimilitud para atributos simples

p y q son los valores de los atributos para dos objetos de datos.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

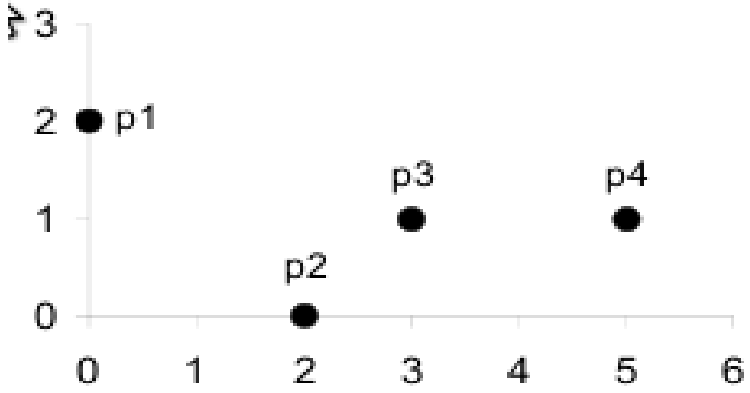
Distancia Euclideana

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

n es la dimensión (numero de atributos)
pk y qk son los k- ésimos atributos de los datos p y q.

- Se realiza normalización si las escalas de los atributos difieren.

Distancia Euclideana



punto	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Matriz de Distancias

Distancia Minskowski

Generalización de la distancia Euclidiana mediante el parámetro r

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

- r = 1. Distancia Manhattan
  Ejemplo típico: Distancia de Hamming: Numero de bits diferentes entre dos arreglos de bits
- r = 2. Distancia Euclidiana
- $r \rightarrow \infty$. Distancia "supremo" (norma Lmax o $L\infty$).
  La máxima diferencia entre los atributos

Distancia Minskowski

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

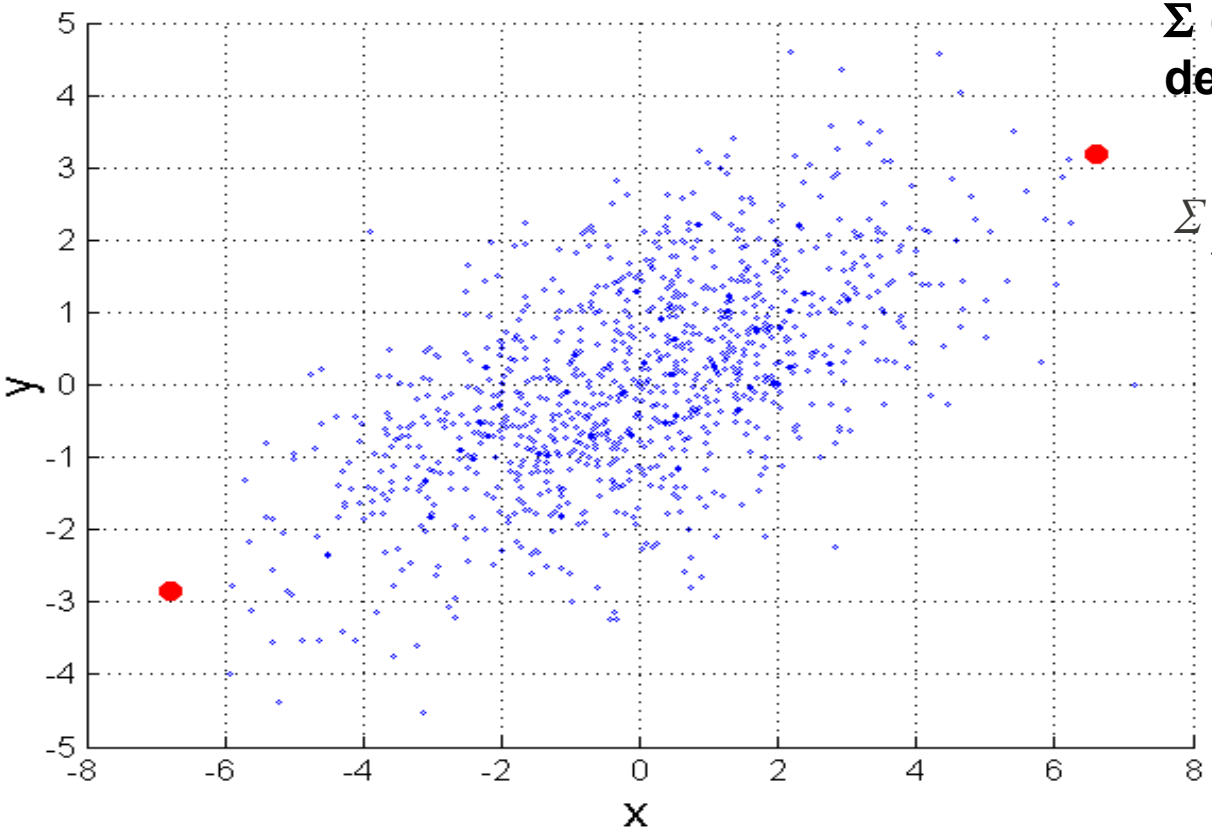
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Distancia Hamming: Número de bits que son diferentes entre dos objetos.

$$D_h(x,y) = b + c$$

Distancia Mahalanobis

$$mahalanobis(p,q) = (p-q)\sum(p-q)^T$$

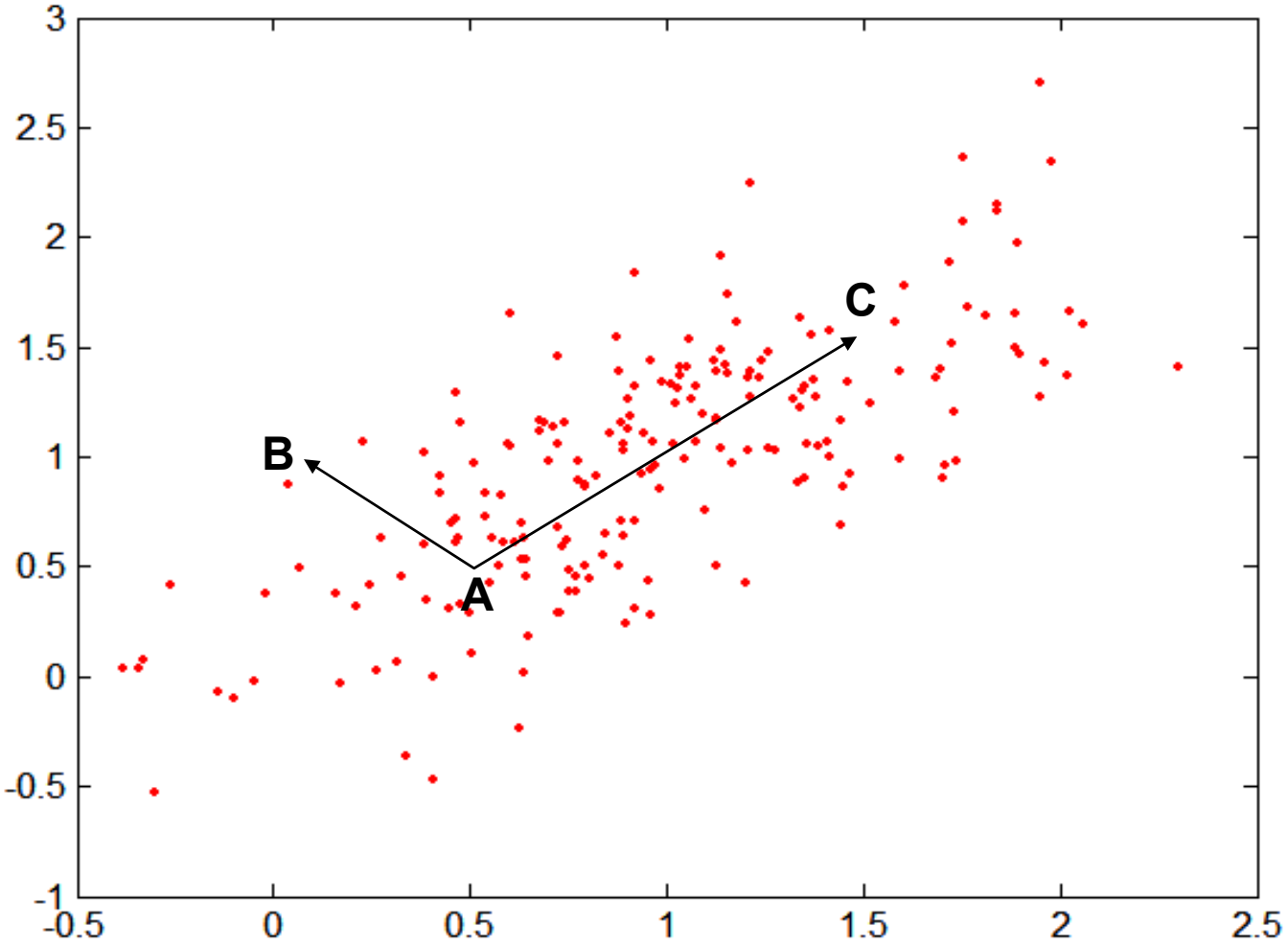


Σ es la matriz de covarianza del conjunto X

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

Para puntos rojos, la distancia Euclideana es 14.7, la distancia Mahalanobis es 6.

Distancia Mahalanobis



Matriz de covarianza

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Medidas de distancia

Propiedades de Medidas de distancia

1. Positiva

d(x,y)>=0

d(x,y)=0 solo si x=y

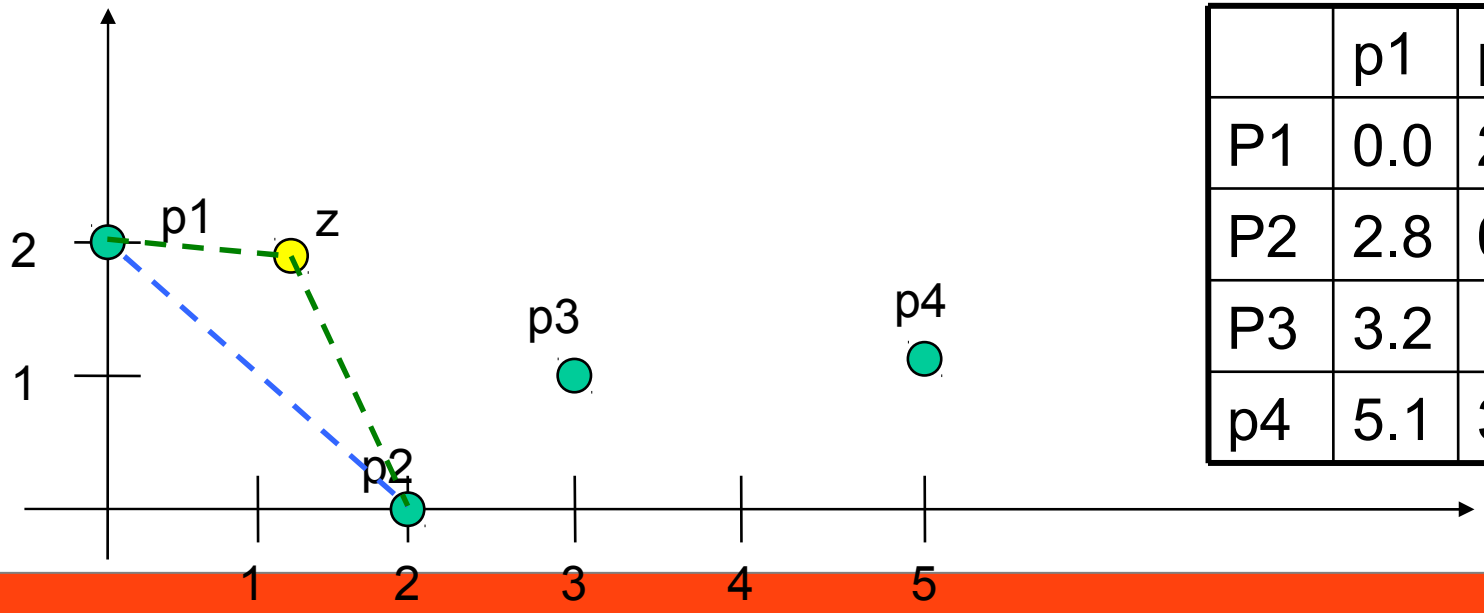
1. Simétrica

d(x,y) = d(y,x) para todo x y y

1. Desigualdad Triangular

d(x,z) <= d(x,y)+d(y,z) para todo punto x,y y z

Métricas



	p1	p2	p3	p4
P1	0.0	2.8	3.2	5.1
P2	2.8	0.0	1.4	3.2
P3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

euclidean

Medidas de Similaridad

Distancias/similaridades binarias

De ayuda construir tabla de contingencia

y \ x	1	0
1	a	b
0	c	d

SMC (Simple Matching Coeficient)

$$S_{smc}(x,y) = \frac{(a+d)}{(a+b+c+d)}$$

Coeficiente de Jaccard

$$S_{jc}(x,y) = \frac{a}{(a+b+c)}$$

Coeficiente de Rao

$$S_{rc}(x,y) = \frac{a}{(a+b+c+d)}$$

SMC vs Jaccard: Ejemplo

$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$M_{01} = 2$ (the number of attributes where $p$was 0 and $q$was 1)

$M_{10} = 1$ (the number of attributes where $p$was 1 and $q$was 0)

$M_{00} = 7$ (the number of attributes where $p$was 0 and $q$was 0)

$M_{11} = 0$ (the number of attributes where $p$was 1 and $q$was 1)

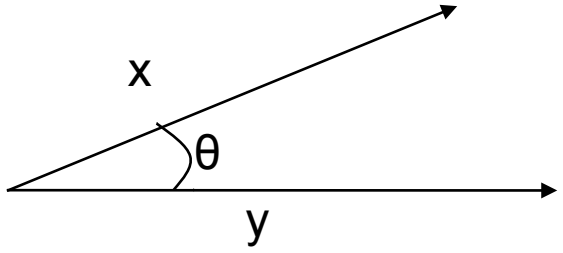
$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Medidas de Similaridad

Similaridad de Coseno

Los objetos se consideran vectores su similaridad se mide por el ángulo que los separa usando el coseno



$$S_{\cos}(x,y) = \frac{\sum_{i=1}^{m} (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{m} x_i^2 \cdot \sum_{i=1}^{m} y_i^2}}$$

Medida del ángulo entre x y y

Si la similaridad es 1 el ángulo es 0 grados, x y y son el mismo excepto por magnitud

Si la similaridad es 0 el ángulo es 90 grados

La medida mas común en calcular la similaridad entre documentos

Ejemplo: $\mathbf{d}_1$ y $\mathbf{d}_2$ son dos vectores de documentos:

$$\mathbf{d}_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$

$$\mathbf{d}_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

entonces

$$S_{cos}(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

donde: • indica producto punto de los vectores y

$\| d \|$ es la longitud del vector d.

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$S_{cos}(d_1, d_2) = 5/ (6.481*2.245) = 0.3150$

Las similitudes tienen algunas características bien conocidas:

1. *s(p, q) = 1 (o máxima similitud) solo si p = q.*

2. *s(p, q) = s(q, p) para todo p y q. (Simétrica)*

Donde *s(p, q) es la similitud entre puntos* (objetos de datos), *p y q.*

Ejercicio

$x=\{0,0,1,1,0,1,0,1\}$ $y=\{0,1,1,0,0,1,0,0\}$

S_{smc}, S_{jc}, $S_{rc,}$, S_{cosine} ?

CoeficienteJaccard Extendido(Tanimoto)

- Jaccard para valores continuos

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

Distinciones entre los conjuntos de Clusters

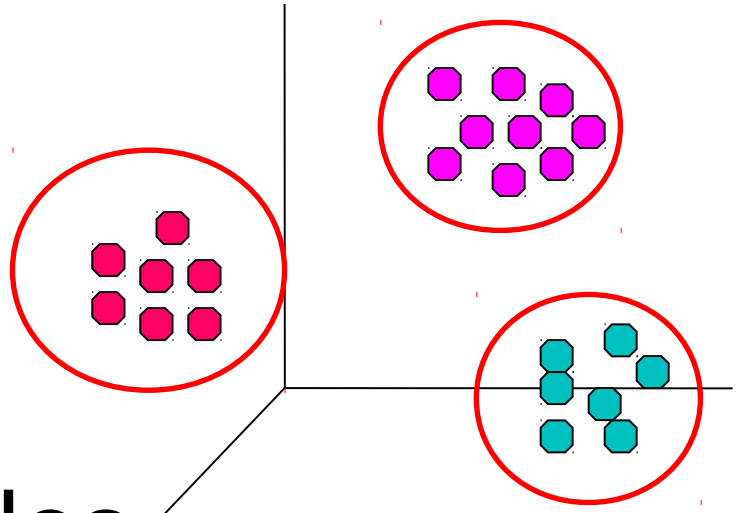
Exclusivo vs. no – exclusivo

- Agrupamientos no exclusivos: los puntos pueden pertenecer a múltiples clusters
- Se puede representar múltiples clases o puntos frontera.

Difuso vs. no - difuso

- En el agrupamiento difuso, un punto pertenece a todo cluster con algún peso entre 0 y 1.
- Los pesos deben sumar 1.

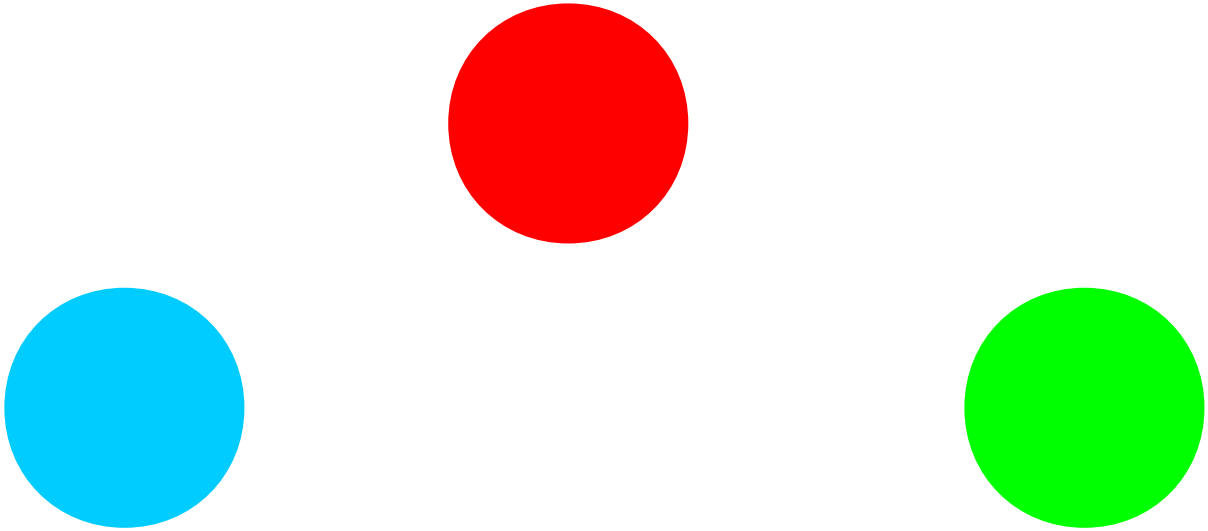
Tipos de Clusters



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

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

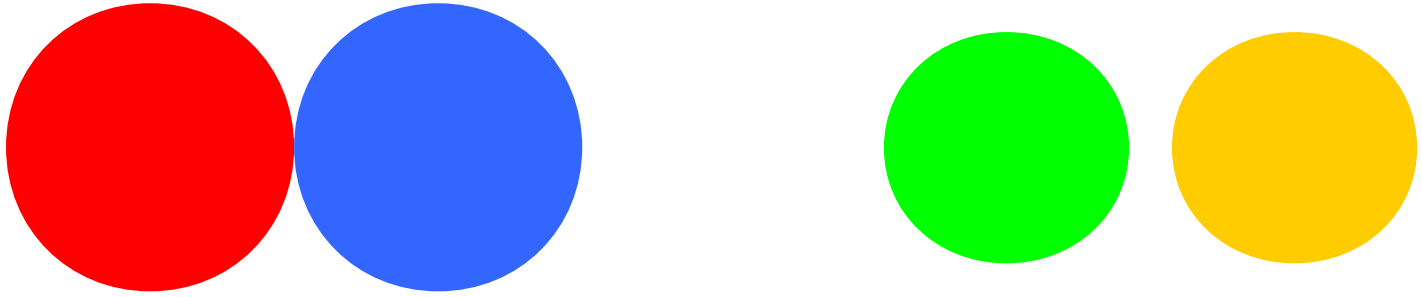


3 clusters bien separados

Basados en el centro

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

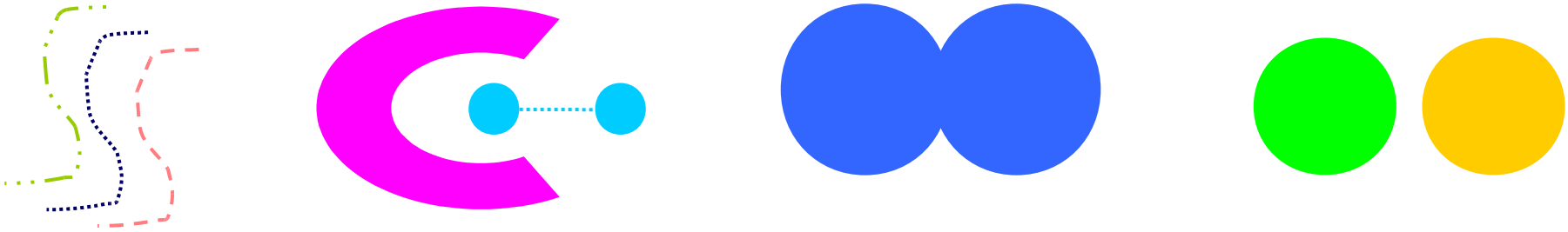
El centro de un cluster frecuentemente es llamado centroide, el promedio de todos los puntos en el cluster o el "medoid", el punto más representativo del cluster.



4 clusters basados en el centro

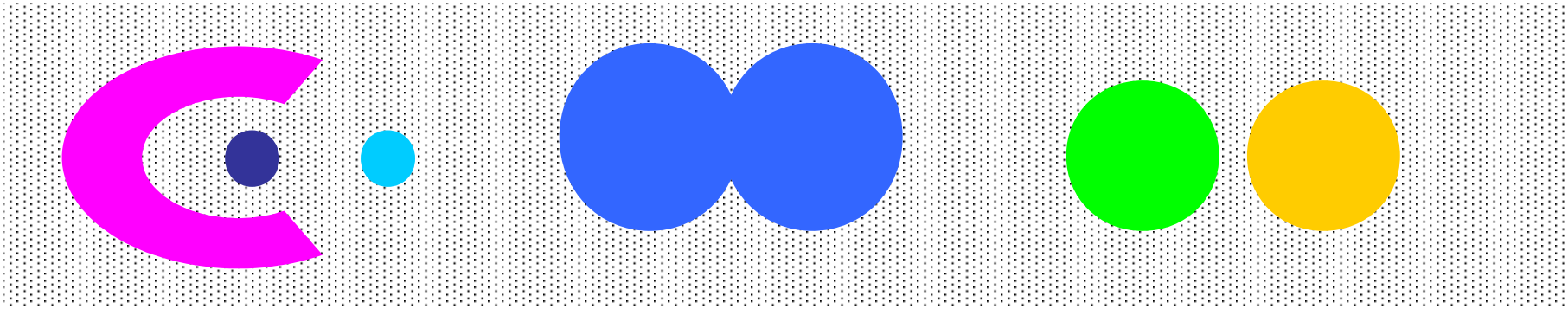
Contiguos

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



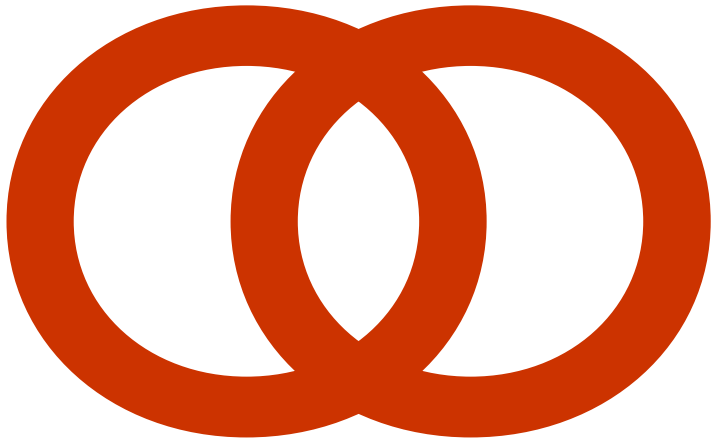
Basados en densidad

- Un cluster es una región densa de puntos, separados por regiones de baja densidad, de otras regiones de alta densidad.
- Se usan cuando los clusters son irregulares o entrelazados, y cuando se presenta ruido y datos atípicos



Conceptuales

- Son clusters que tienen alguna propiedad en comun o representan un concepto particular



Definidos por una función objetivo

- Son clusters que minimizan o maximizan una función objetivo
- Enumeran todas las posibles formas de dividir los puntos dentro de un cluster y evalúan la “bondad” de cada conjunto potencial de clusters usando una función objetivo dada (NP Hard)

Tipos de Agrupamiento

Agrupamiento Particional

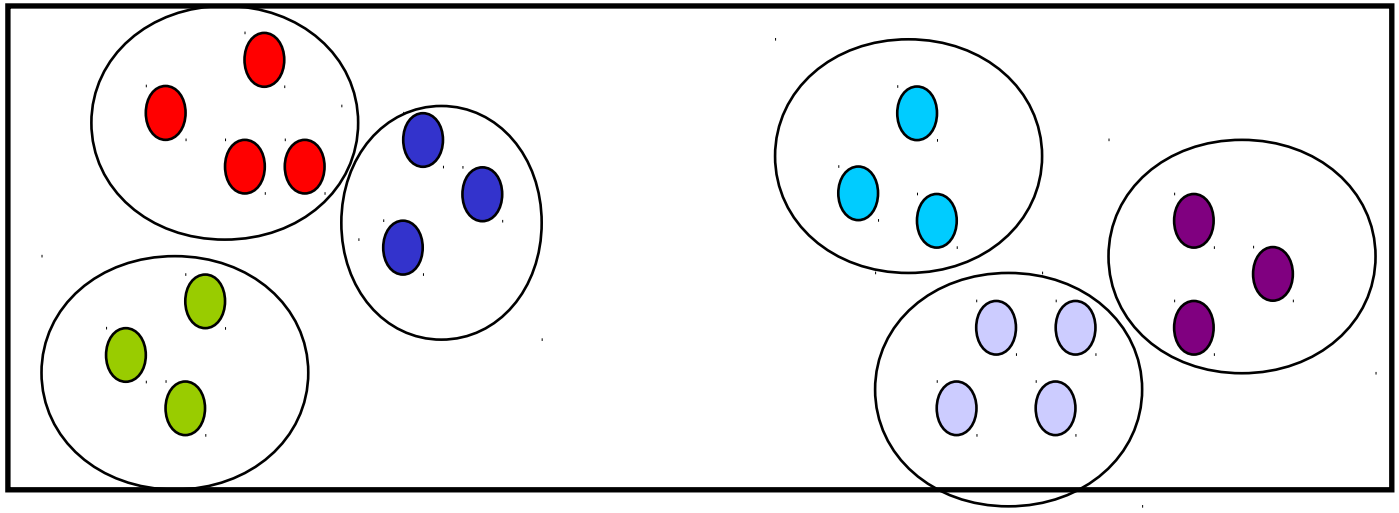
Dividir los datos (puntos, objetos, registros) en grupos no superpuestos, donde cada dato (punto, objeto, registro) pertenece a un único grupo.

Agrupamiento Jerárquico

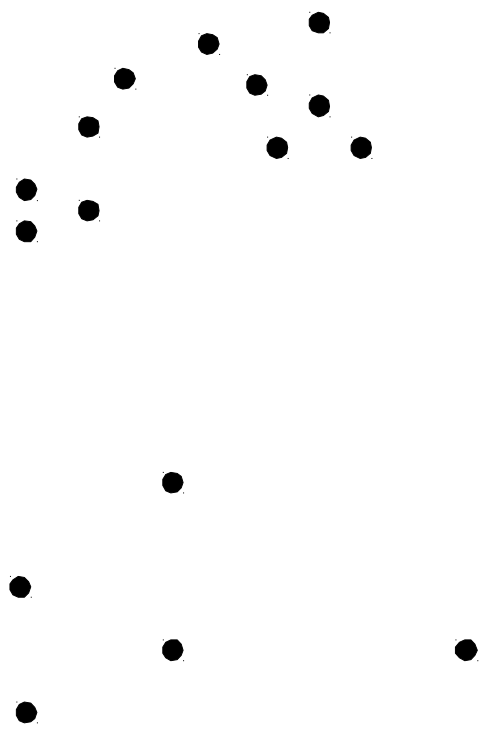
Organiza los datos (puntos, objetos, registros) en grupos sobrepuestos en forma de árbol. Usa estructura de árbol o dendograma

Agrupamiento Particional

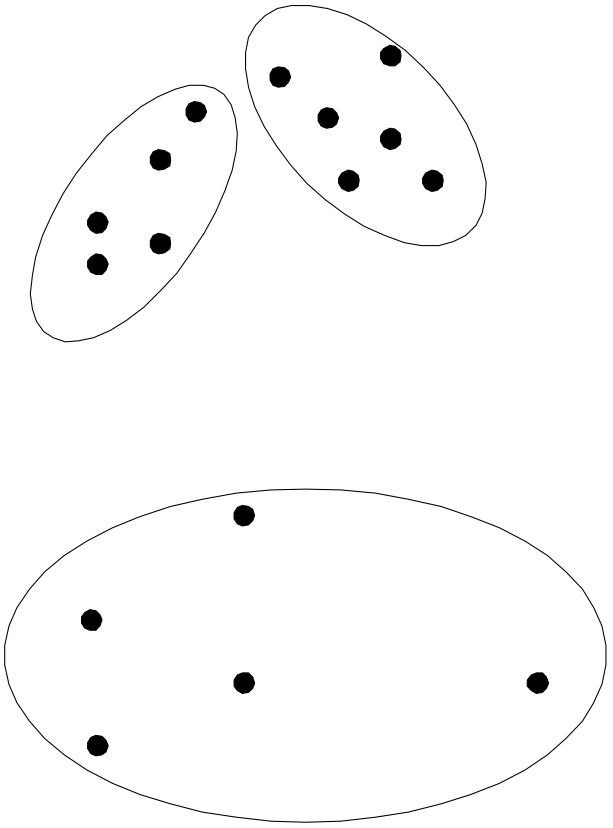
Minimizar la distancia en cada uno de los grupos
o maximizando la distancia entre los grupos



Agrupamiento Particional

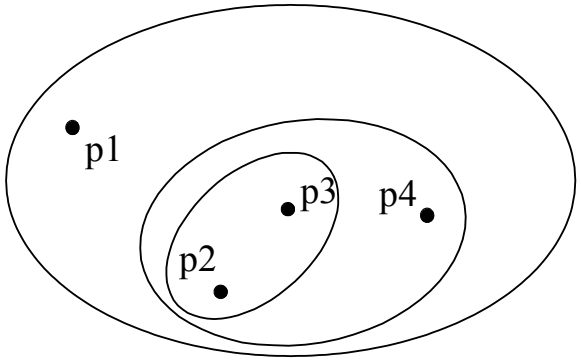


Puntos originales

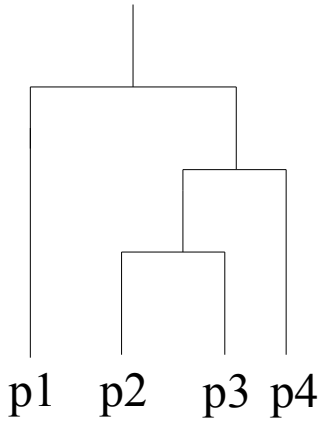


Agrupación particional

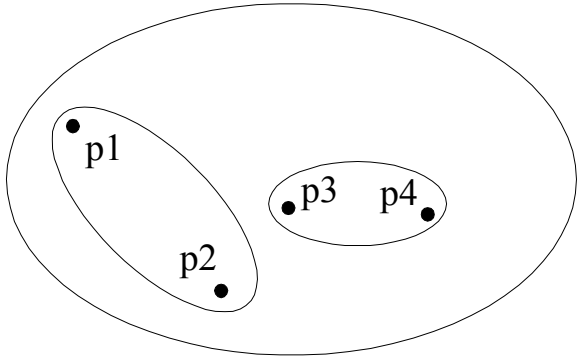
Agrupamiento Jerárquico



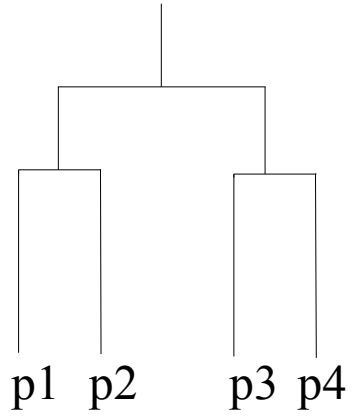
Agrupamiento jerárquico tradicional



Dendograma tradicional



Agrupamiento jerárquico No tradicional



Dendograma no tradicional

Definidos por una función objetivo

- Tipo de proximidad o medida de la densidad

  Medida derivada básica para el agrupamiento.
- Densidad (dispersión)

  tipo de similitud, eficiencia
- Tipo de atributo

  tipo de similitud
- Tipo de Datos

  tipo de similitud, Otra característica: auto-correlación
- Dimensionalidad
- Ruido y datos atípicos (Outliers)
- Tipo de Distribución

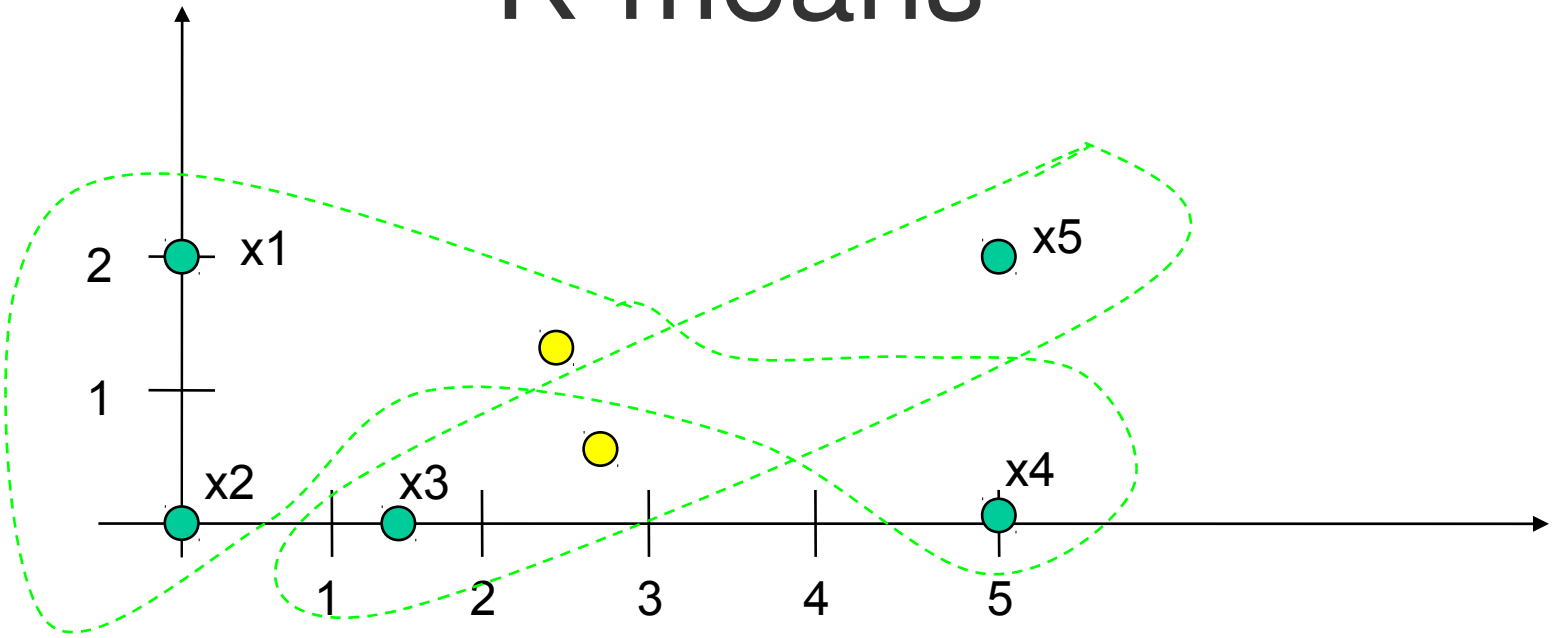
Algoritmos

K-means

- 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 clusters **"K"** debe ser especificado
- El algoritmo básico es muy simple

1: Seleccionar K puntos como los centroides iniciales
2: **Repetir**
3: Desde K clusters asignar todos los puntos al centroide más cercano
4: Recalcular el centroide de cada cluster
5: **Hasta** El centroide no cambia

K-means



K=2

C1={x1,x2,x4} y C2={x3,x5}

Centros M1= (0+0+5)/3,(2+0+0)/3 = (1.66, 0.66)

Centros M2= (1.5+5)/2,(0+2/2) = (3.25, 1.00)

Calcula variaciones en error:

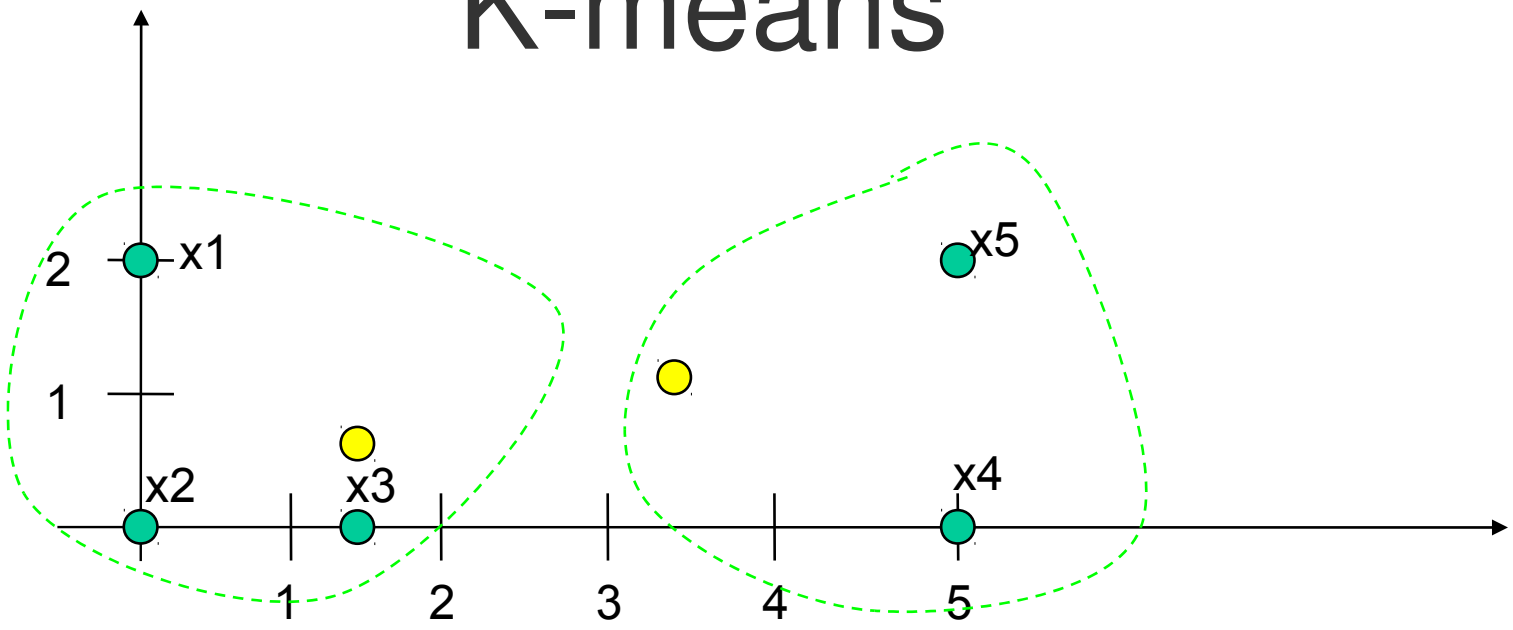
$e1^2 = [(0-1.66)^{2+}(2-0.66)^2]+[(0-1.66)^{2+}(0-0.66)^2]+[(5-1.66)^{2+}(0-0.66)^2]$

$= 19.36$

$e2^2 = [(1.5-3.25)^{2+}(0-1)^2]+[(5-3.25)^{2+}(2-1)^2] = 8.12$

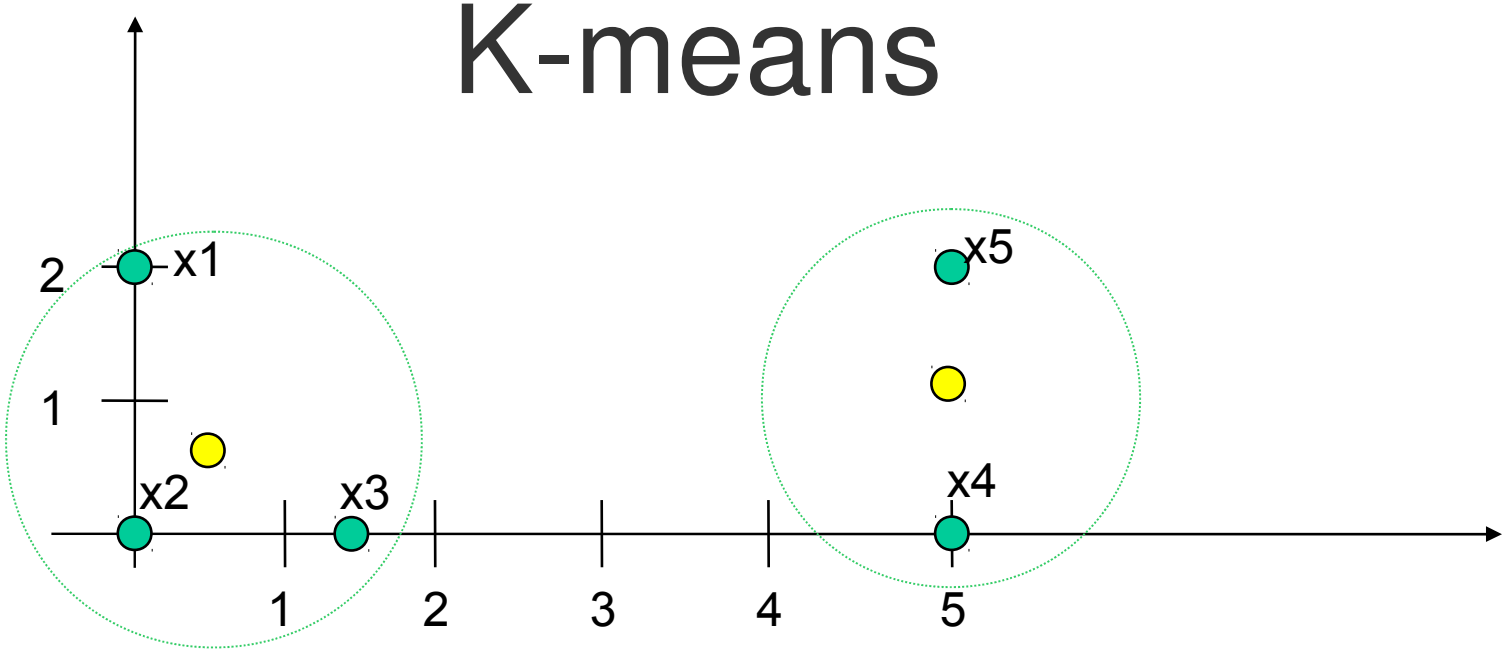
Total error = 19.36 + 8.12 = **27.48**

K-means



- Resignar ejemplos
  - d(M1,x1)=(1.66²+1.34²)=2.14 d(M2,x1)= 3.40 => x1 Є C1
  - d(M1,x2)= 1.79 d(M2,x2)= 3.40 => x2 Є C1
  - d(M1,x3)= 0.83 d(M2,x3)= 2.01 => x3 Є C1
  - d(M1,x4)= 3.41 d(M2,x4)= 2.01 => x3 Є C2
  - d(M1,x5)= 3.60 d(M2,x5)= 2.01 => x3 Є C2
- C1={x1,x2,x3} y C2={x4,x5}
- Centros M1= (0.5, 0.67)
- Centros M2= (5.0, 1.0)

K-means



- Calcula variaciones en error:
- $e1^2 = 4.17$
- $e2^2 = 2.0$
- Total error = 6.17 (se reduce de 27.48 a 6.17!)
- Se repite hasta que la diferencia de error sea mínima o los centros no cambien!

K-means

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)

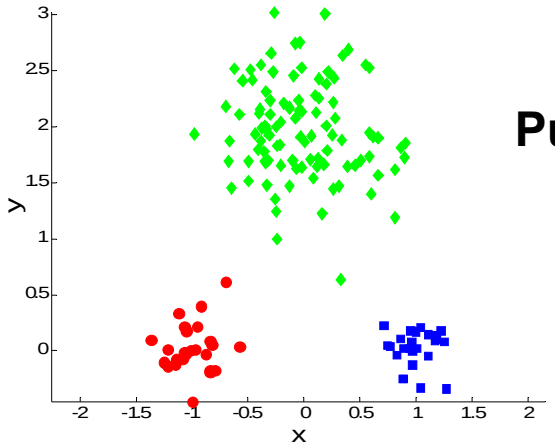
Menos sensible a ruido y outliers

K-means

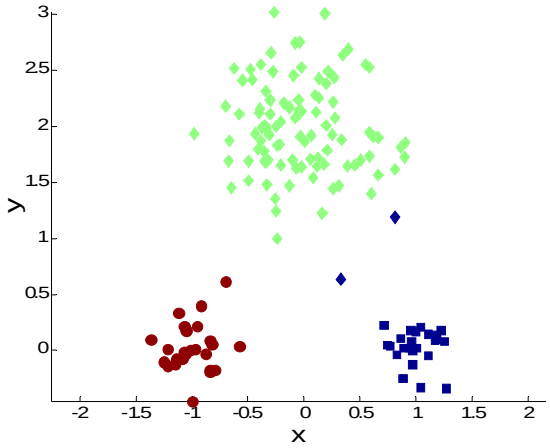
- Los centroides iniciales se escogen aleatoriamente.
- Los clusters generados varían de una ejecución a otra.
- La proximidad es medida por la distancia Euclidiana, la similitud por coseno, correlación, etc.
- K-means convergerá a una medida de similitud común mencionada anteriormente.
- La mayoría de la convergencia ocurre en las primeras iteraciones:

  Frecuentemente la condición para parar es cambiada por “Hasta que algunos puntos cambien de cluster”

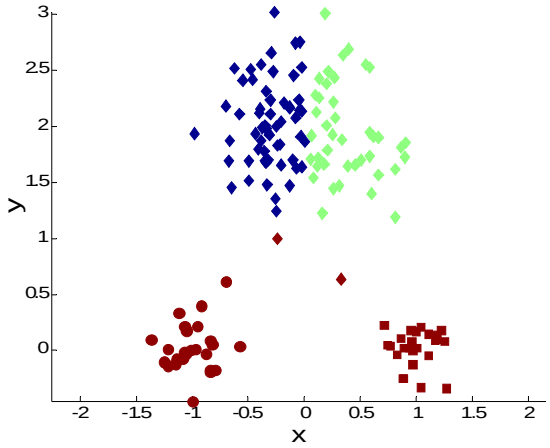
Dos agrupamientos diferentes con k-means



Puntos originales

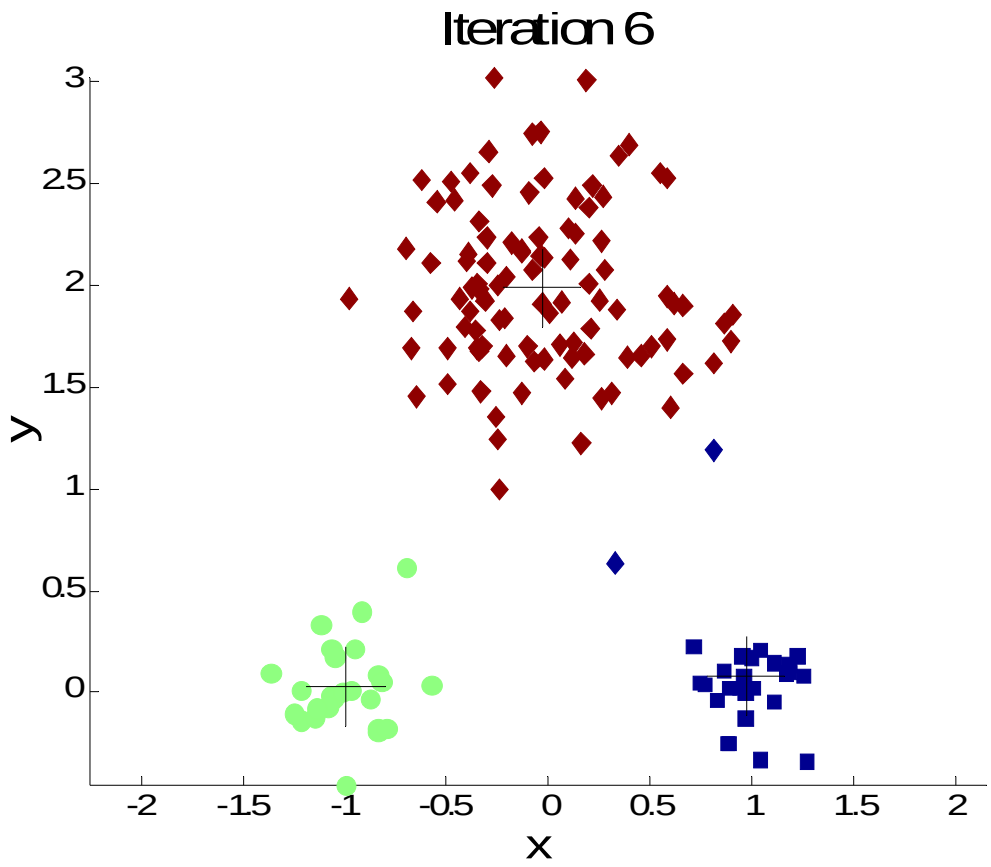


Agrupación Optima

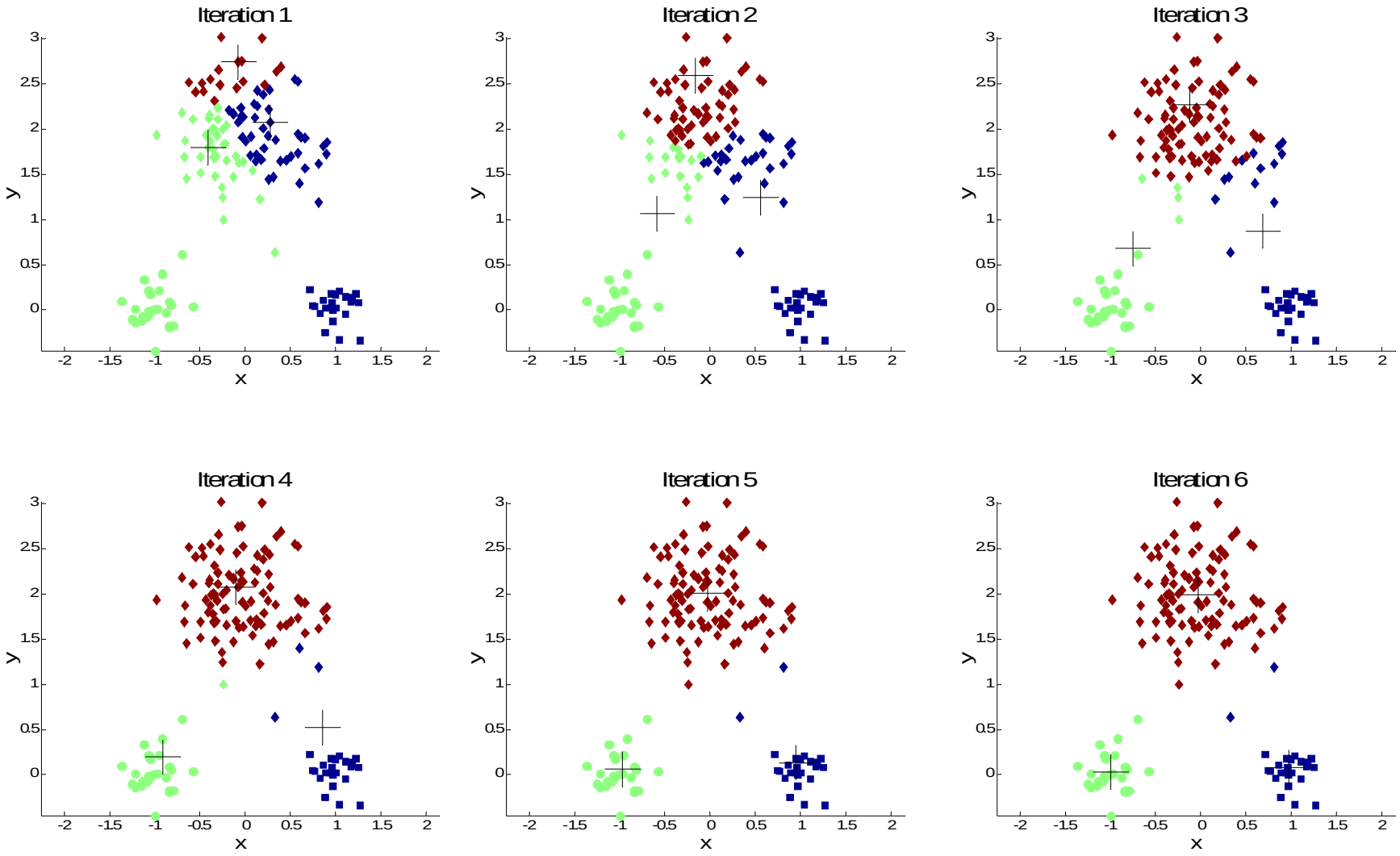


Agrupación subóptima

Inicialización: Importante!



Inicialización: Importante!



K-means: Evaluación de clusters

La medida más común es la **suma del Error Cuadrático (SSE)**

- Cada punto, el error es la distancia del cluster más cercano
- Para obtener el SSE, se elevan al cuadrado los errores y se suman

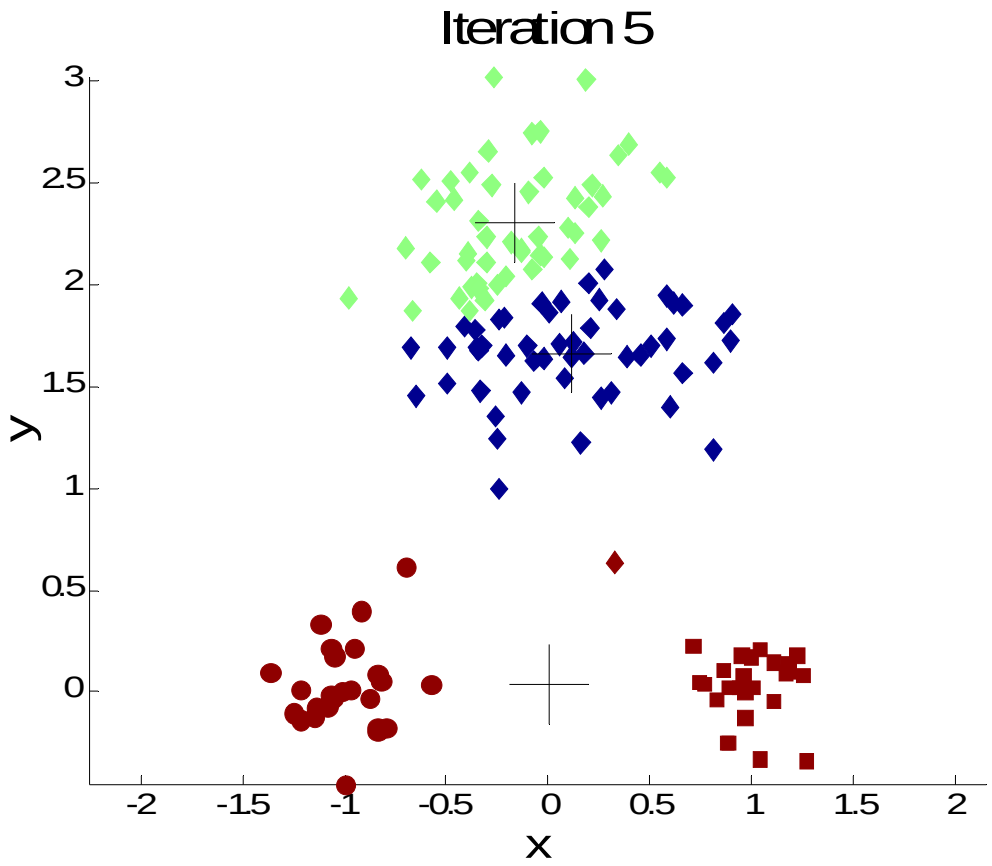
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- x es un punto en el cluster Ci y mi es el punto representativo para el cluster Ci

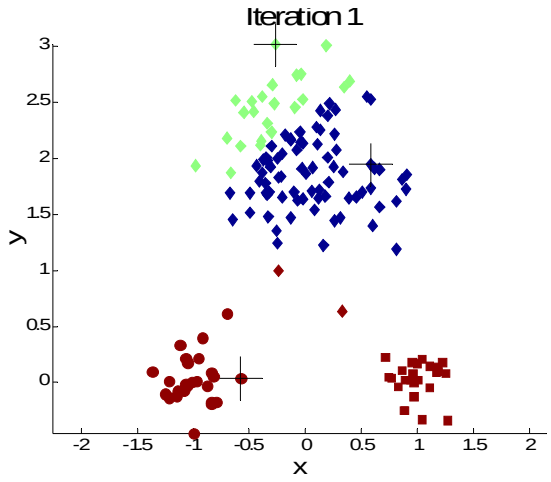
  Se puede mostrar que mi corresponde al centro (promedio) del cluster
- Dados dos clusters, se puede elegir uso con el menor error
- Una manera fácil de reducir el SSE es incrementar K, el número de clusters

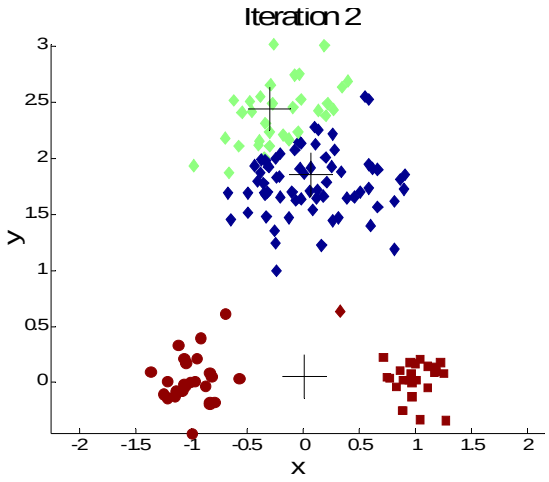
  Un buen agrupamiento con K pequeño, puede tener un SSE más bajo que un agrupamiento con un K grande

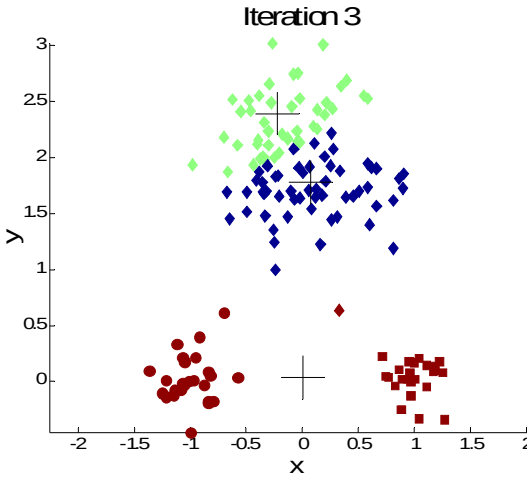
Inicialización: Importante!

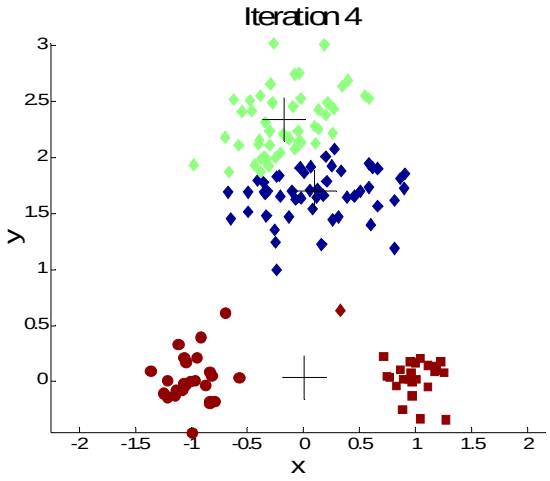


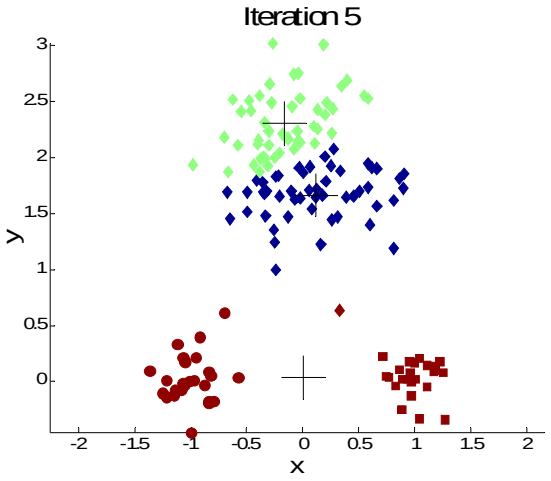
Inicialización: Importante!





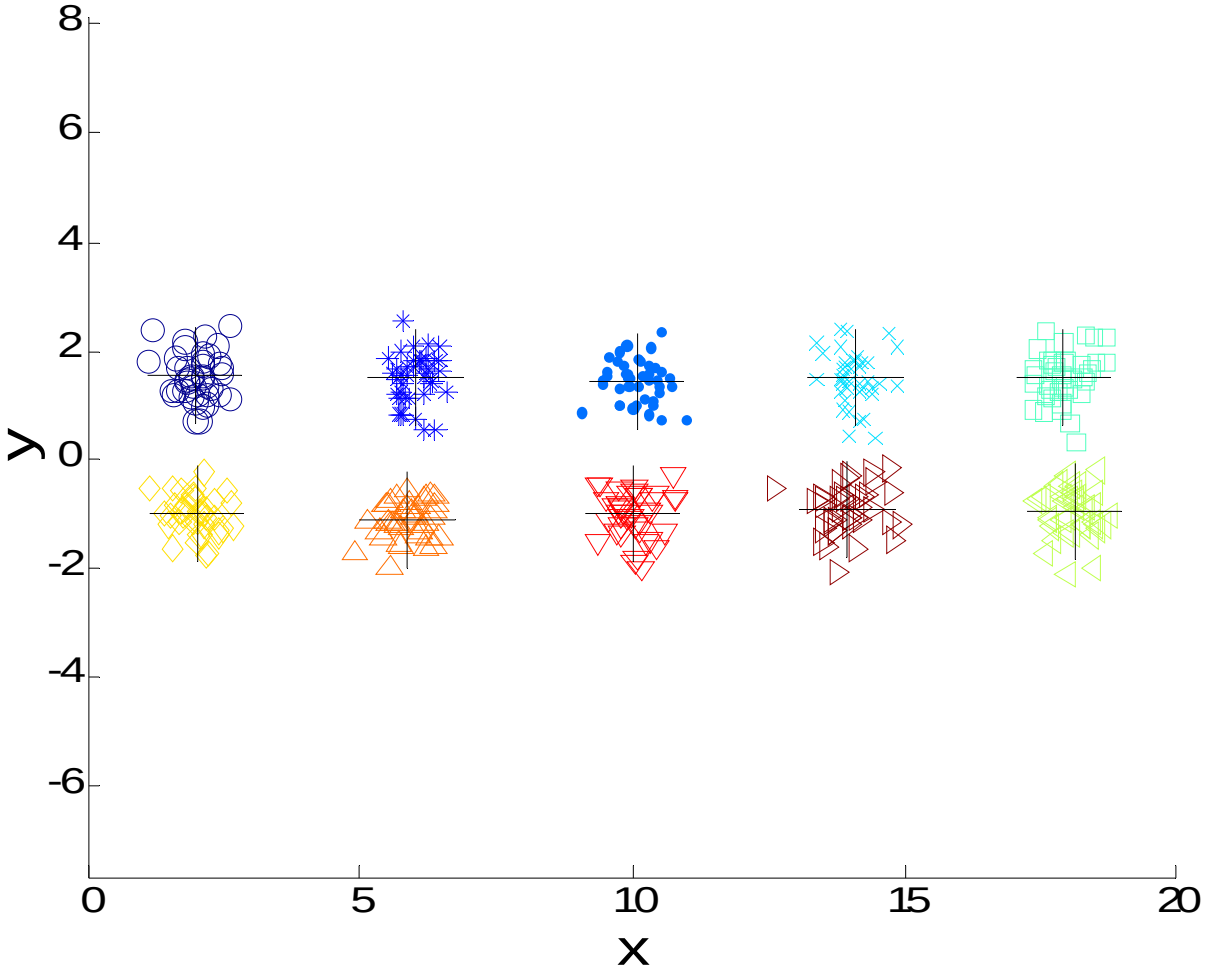




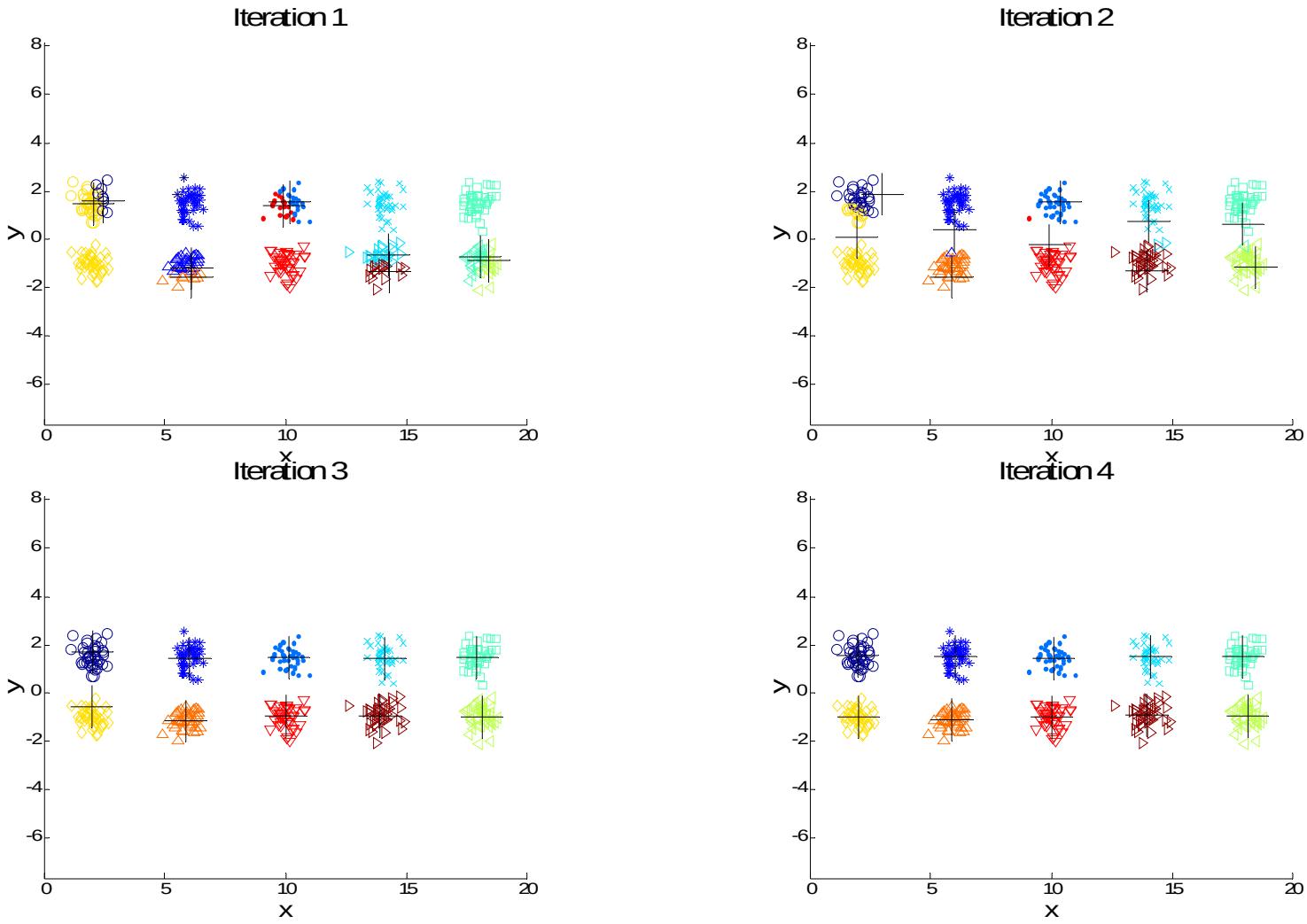


Ejemplo

10 clusters

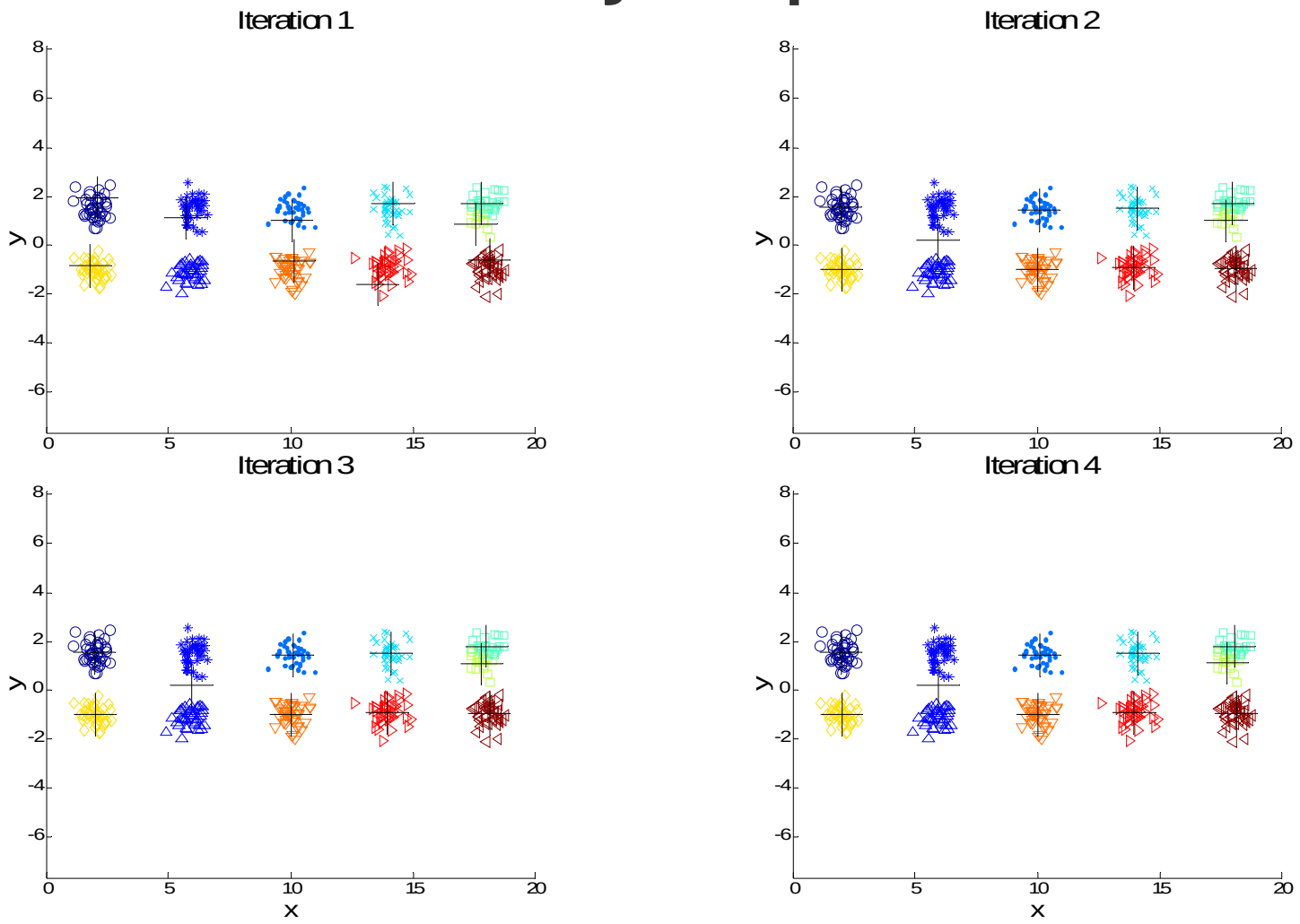


Ejemplo



Comenzando con dos centroides iniciales en un cluster para cada par de clusters

Ejemplo



Comenzando con algunos pares de clusters teniendo tres centroides iniciales, mientras los otros tienen solo uno

Soluciones para la inicialización

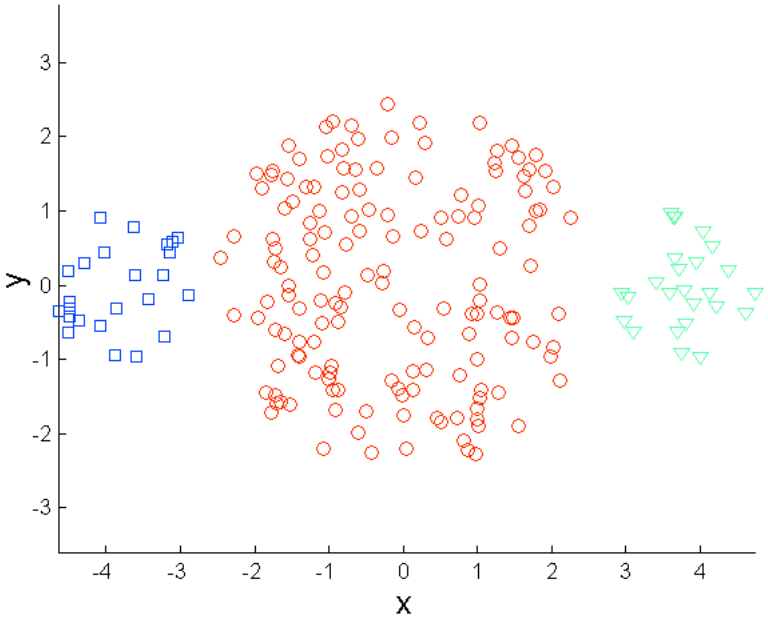
- Múltiples ejecuciones

  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

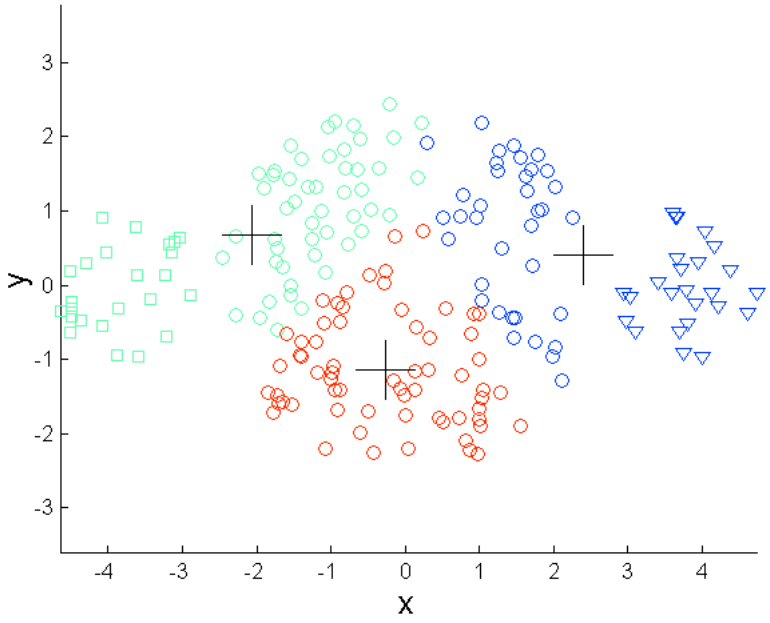
  Seleccionarlos ampliamente separados
- Postprocesamiento
- Bisectar K-means

  No se recomienda para inicialización

Limitaciones: Diferentes tamaños

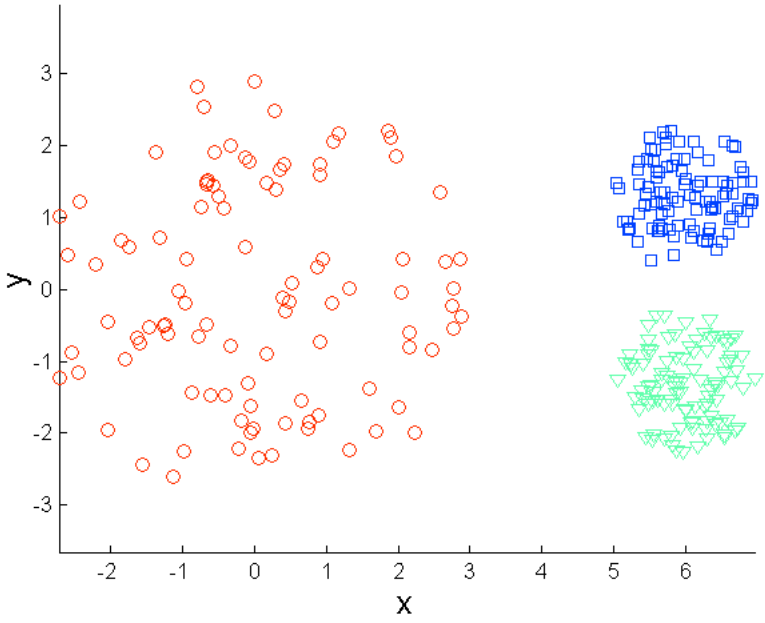


Puntos originales

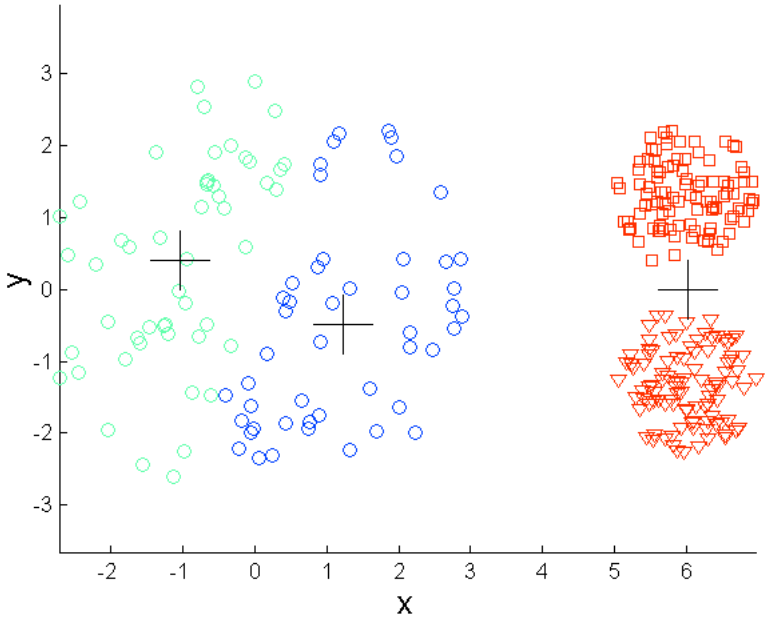


K-means (3 Clusters)

Limitaciones: Diferentes densidades

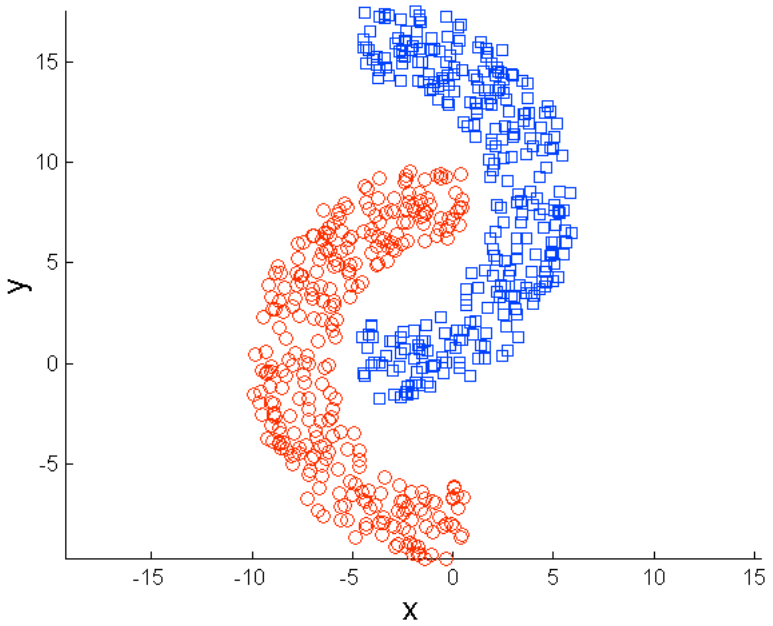


Puntos originales

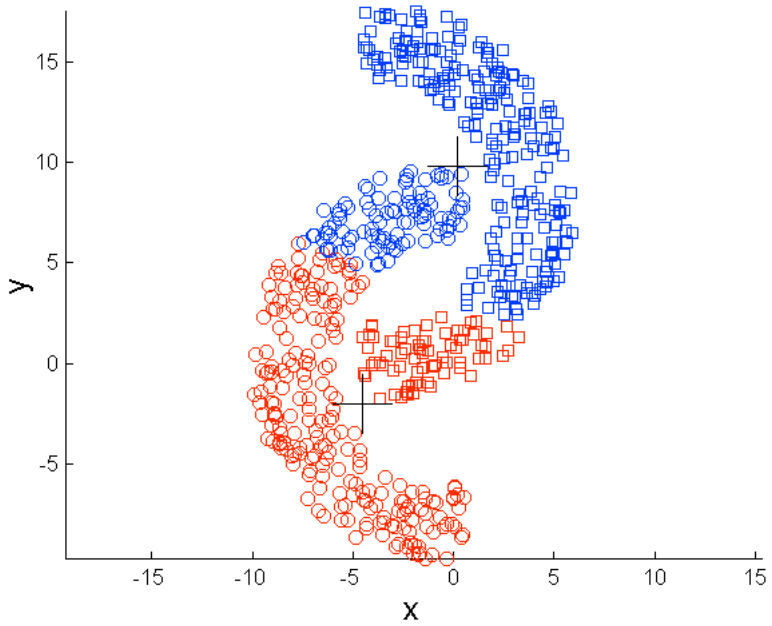


K-means (3 Clusters)

Limitaciones: Diferentes formas

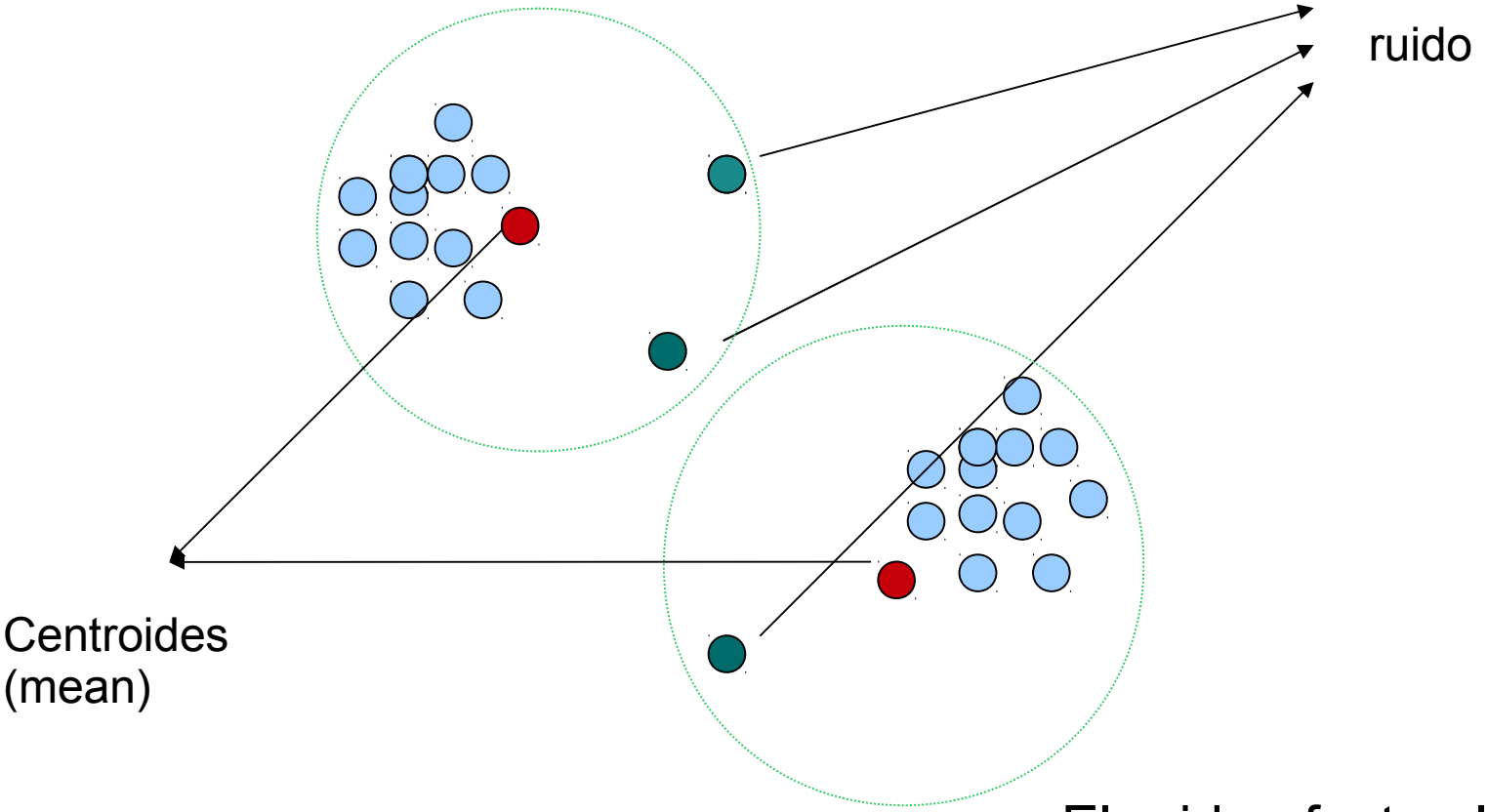


Puntos originales



K-means (2 Clusters)

Limitaciones: Ruido



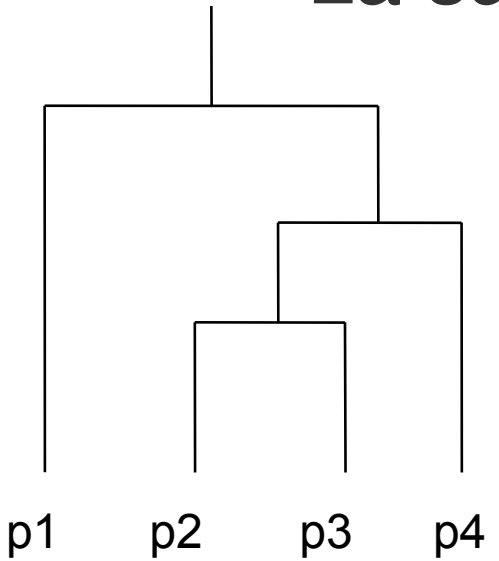
El ruido afecta el centroide del cluster

Ejercicio en Weka

- Abrir conjunto de datos Iris
- Quitar la clase
- Aplicar el algoritmo de k-means con tres clusters
- Visualizar
- Comparar con el conjunto de datos Iris original

Agrupamiento Jerárquico

NO se especifica el número de clusters
La salida es una jerarquía de clusters
Proceso iterativo



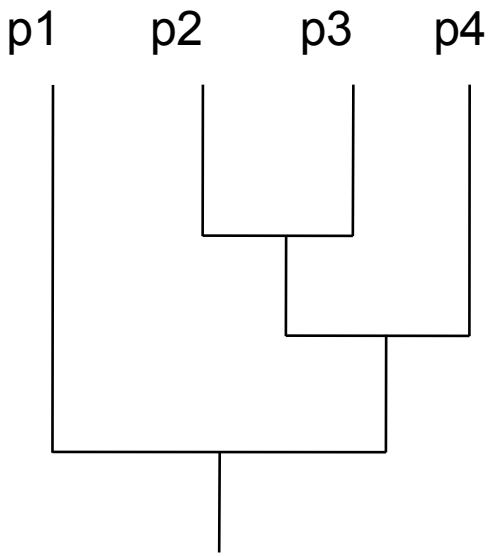
Dendograma

Los algoritmos de este tipo se dividen en dos clases:

- Aglomerativos
- Divisibles

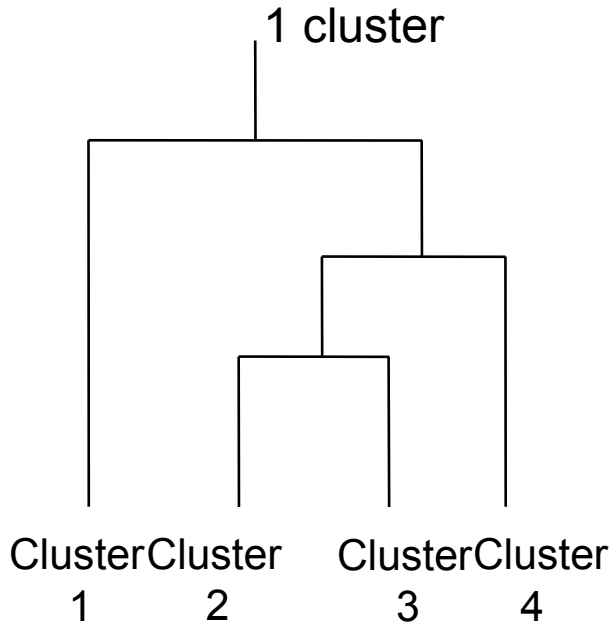
Agrupamiento Jerárquico
Aglomerativo

Al comienzo todos los puntos (objetos) son clusters individuales (tamaño 1) , en cada paso, los clusters mas cercanos se unen para formar un solo cluster, al final se tiene un solo cluster



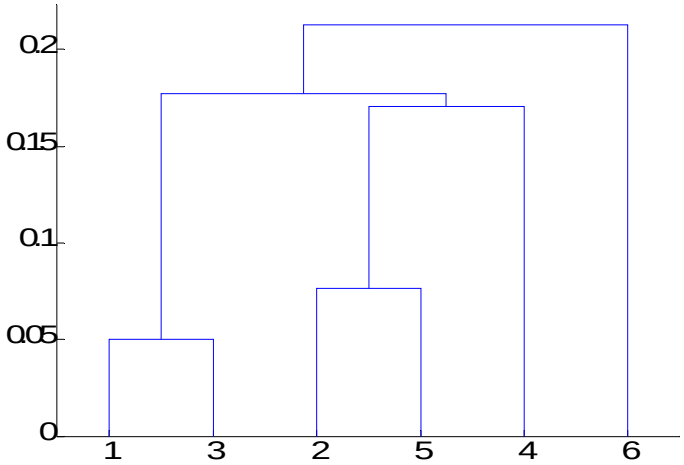
Agrupamiento Jerárquico
Divisible

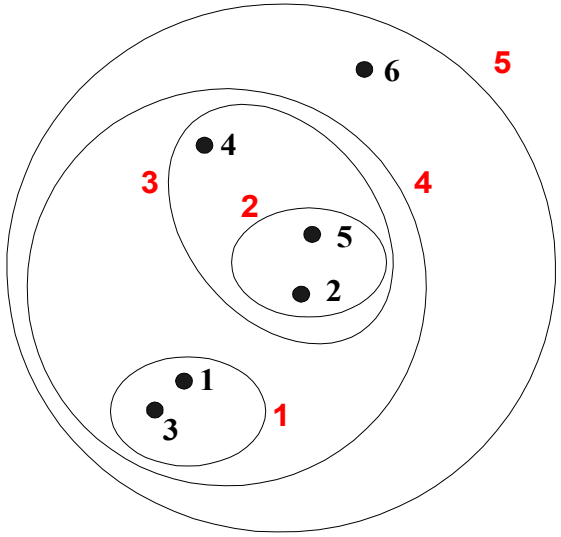
Al comienzo solo existe un cluster (contiene todos los objetos), en cada paso, se van dividiendo los clusters hasta que cada objeto es un solo cluster.



Agrupamiento Jerárquico

Los clusters se visualizan como un dendograma: árbol que registra las secuencias de las uniones y divisiones de los clusters





Fortalezas

- No se asume un número particular de clusters

  El número deseado de clusters se obtiene seleccionando el nivel adecuado del dendograma

- Puede corresponder a taxonomías significativas

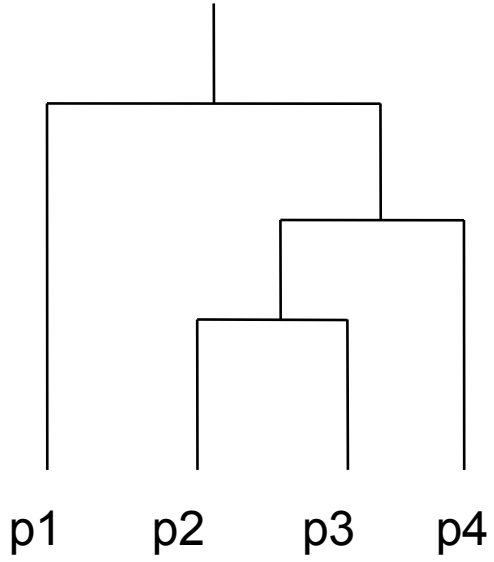
  Ejemplo: en biología

  reino animal, reconstrucción de filogenia

Agrupamiento Jerárquico
Aglomerativo

El más común

Usa árbol o diagrama llamado dendograma (despliega la relación entre el cluster y sub-cluster, y el orden en el cual los clusters fueron fusionados)



Dendograma

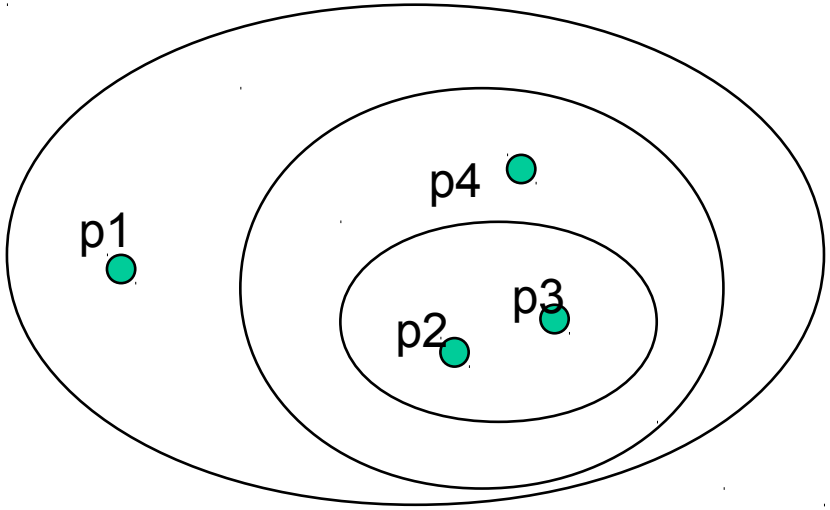


Diagrama de clusters anidados

Agrupamiento Jerárquico Aglomerativo

Algoritmo básico

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

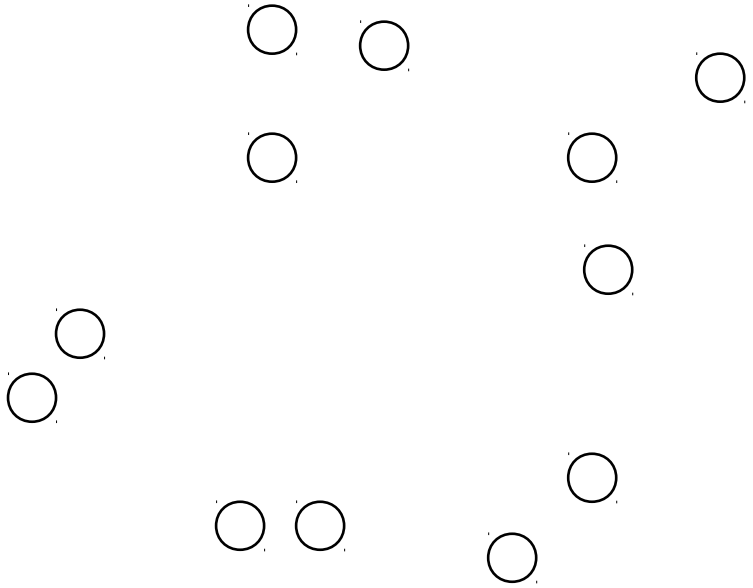
La operación clave es el cálculo de la proximidad entre dos clusters:

Las diversas formas de calcular la proximidad (distancia o similaridad) distinguen los diferentes algoritmos

Ejemplo

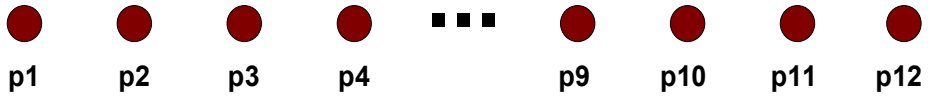
Situación inicial

Comenzar con clusters de puntos individules y la matriz de proximidad



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

Matriz de proximidad



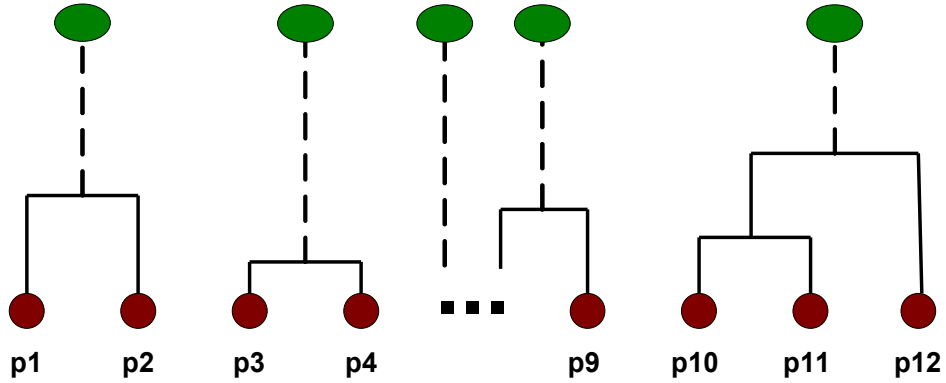
Ejemplo

Situación intermedia

Después de algunas uniones (merges), se tienen algunos clusters

	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

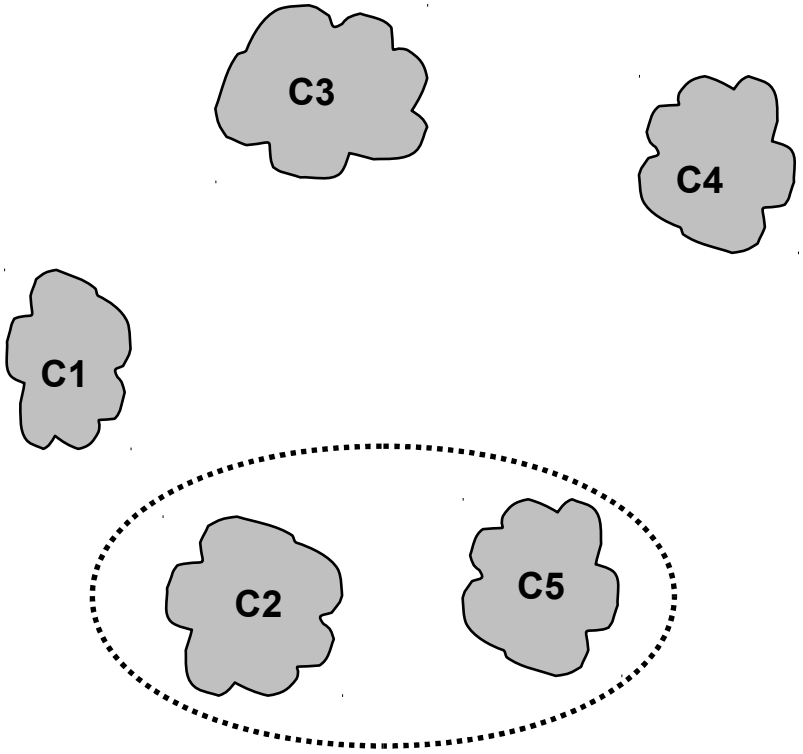
Matriz de proximidad



Ejemplo

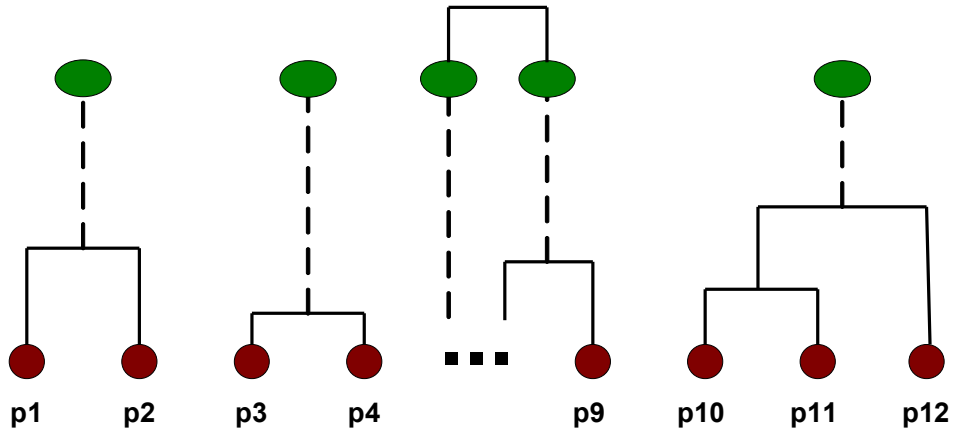
Situación intermedia

Se desea unir los dos clusters mas carcanos (C2 y C5) y actualizar la matriz de proximidad



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

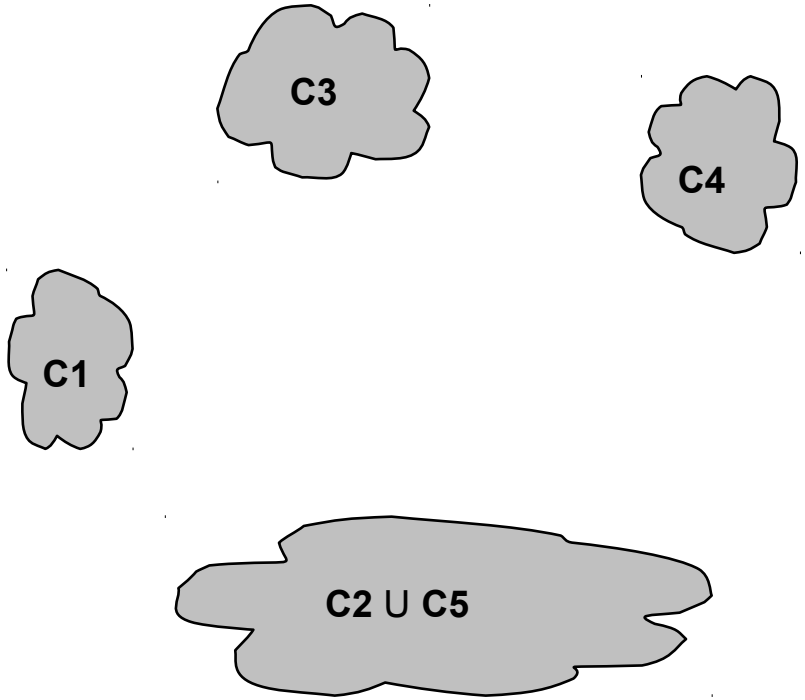
Matriz de proximidad



Ejemplo

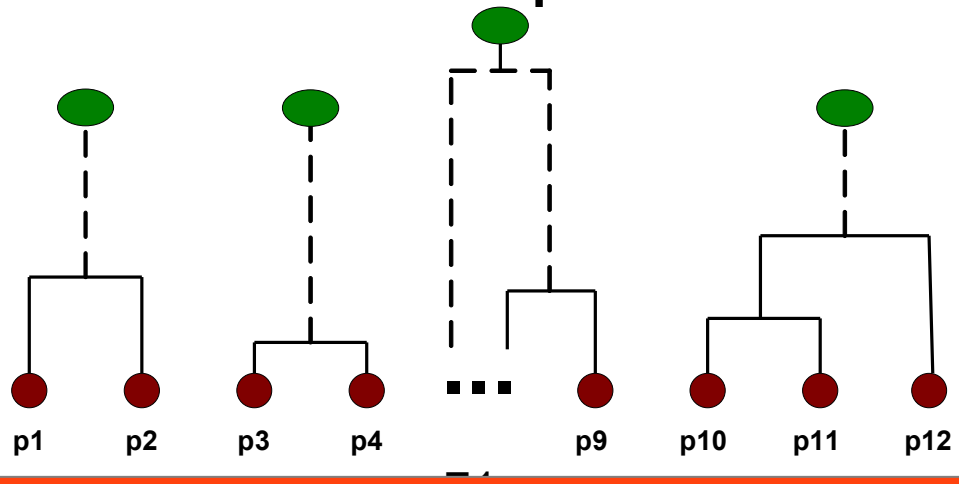
Despues de unir

La pregunta es ¿Cómo actualizar la matriz de proximidad?

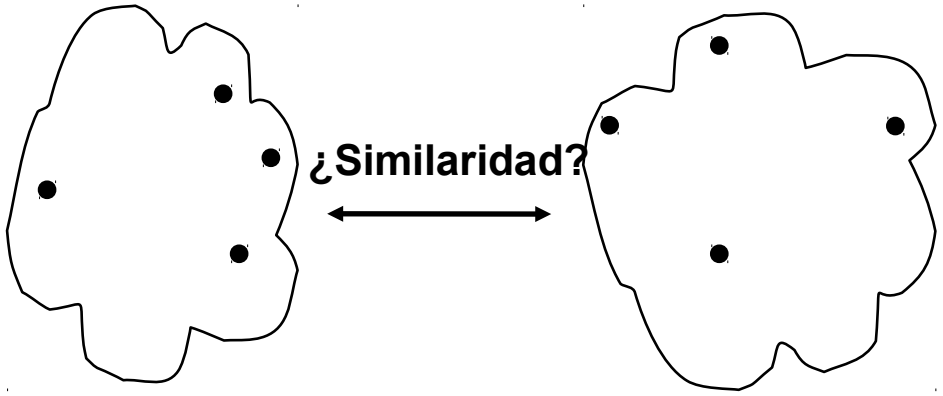


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de proximidad



Similaridad Inter-Cluster



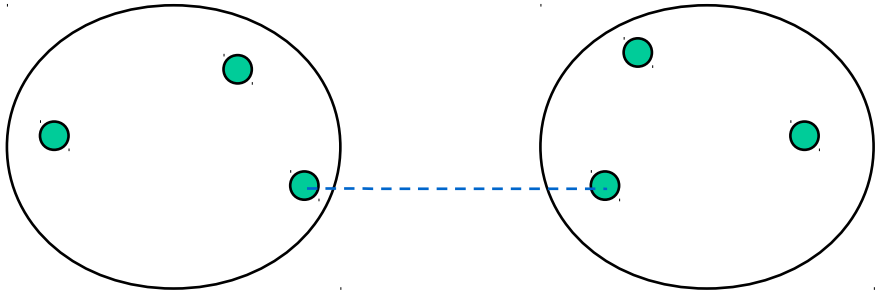
- MIN
- MAX
- Promedio grupo (average)
- Distancia entre centroides
- Usando función objetivo:
  - Metodo de Ward's usa el error cuadrático

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

Matriz de proximidad

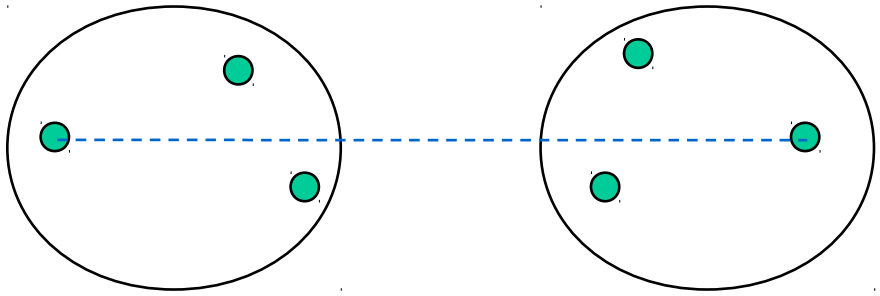
Similaridad Inter-Cluster
MIN

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



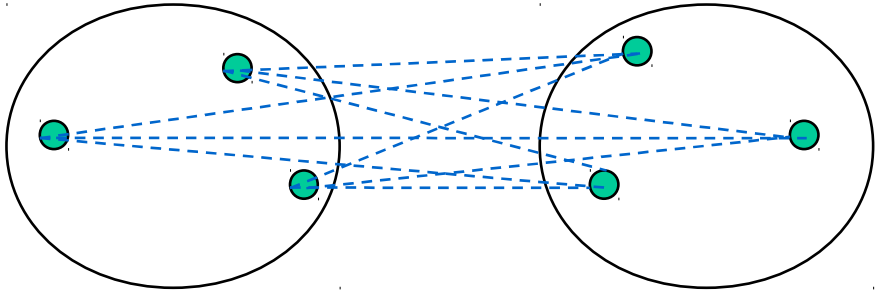
Similaridad Inter-Cluster
MAX

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



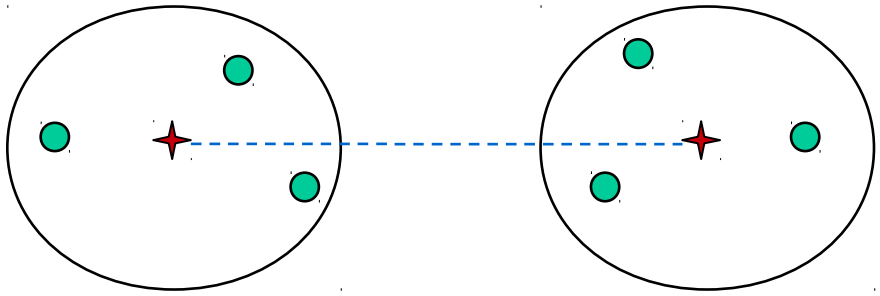
Similaridad Inter-Cluster
GROUP AVERAGE
promedio del grupo

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



Similaridad Inter-Cluster entre prototipos como centroides

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

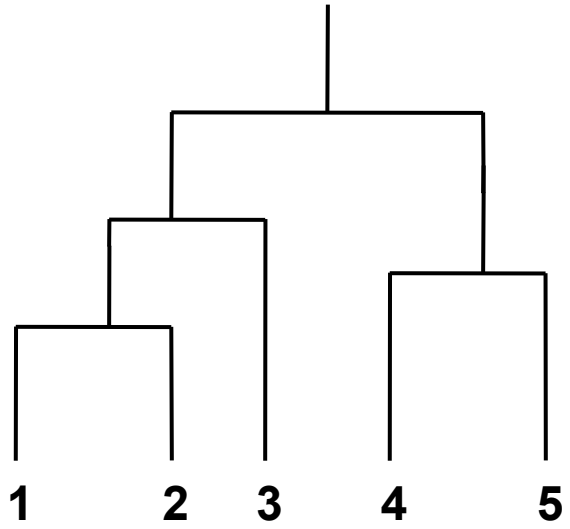


Similaridad Inter-Cluster
MIN o Enlace Simple

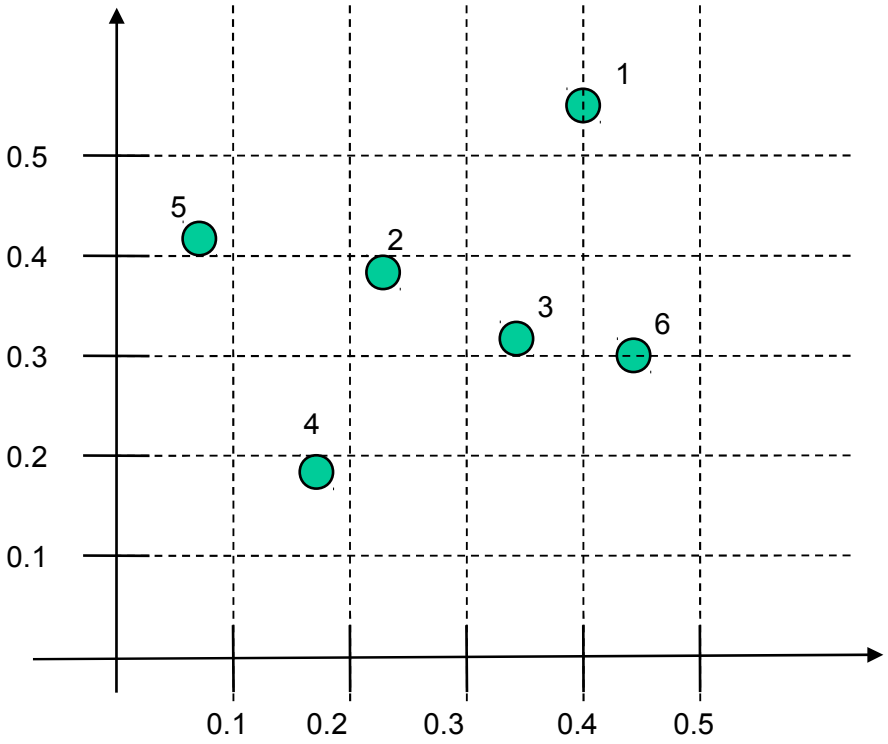
La similitud de dos cluster se basa en los dos puntos más similares (cercanos) en dos clusters

Es determinado por un par de puntos, es decir, un enlace en la gráfica de proximidad

	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



Ejercicio

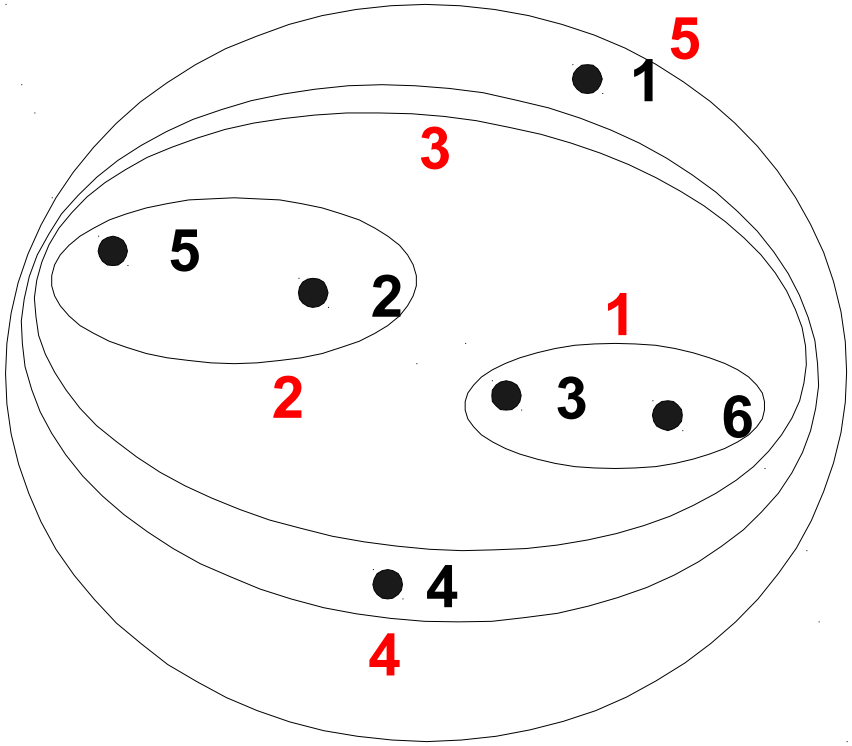


Coordenadas x,y

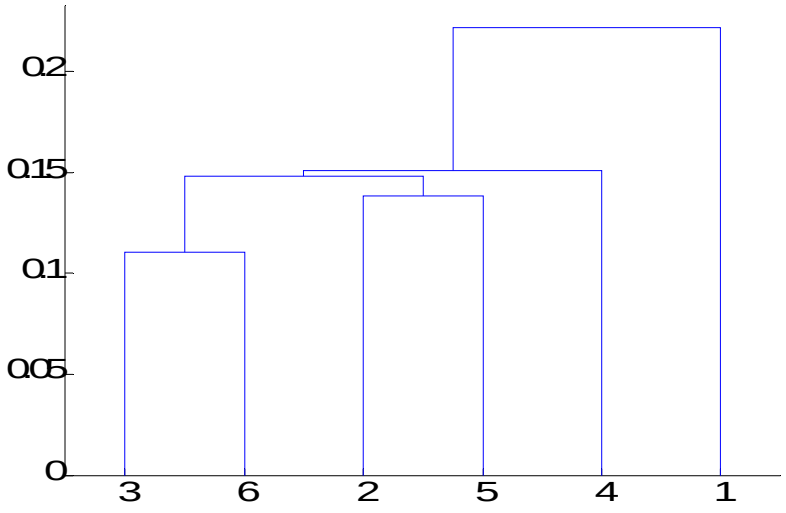
Punto	x	y
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

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

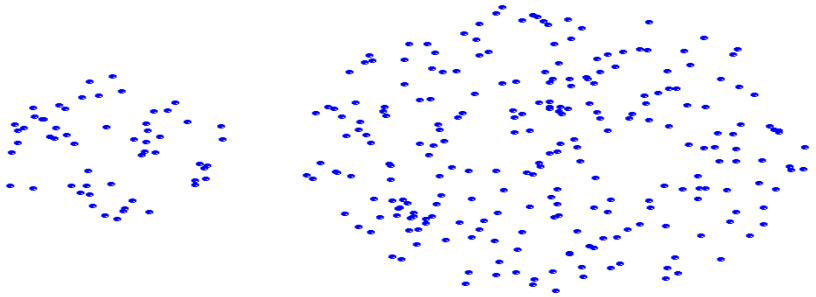


Clusters anidados

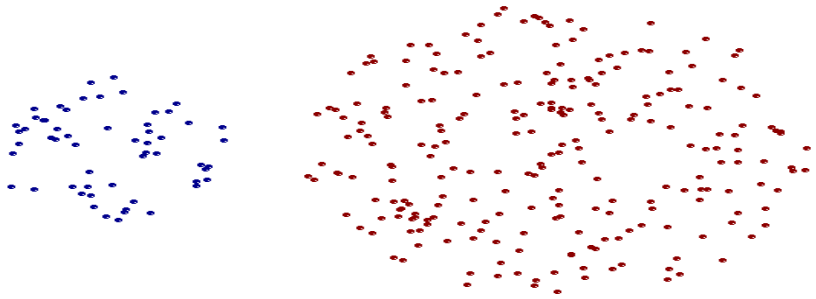


Dendrograma

Fortaleza de MIN o Enlace Simple



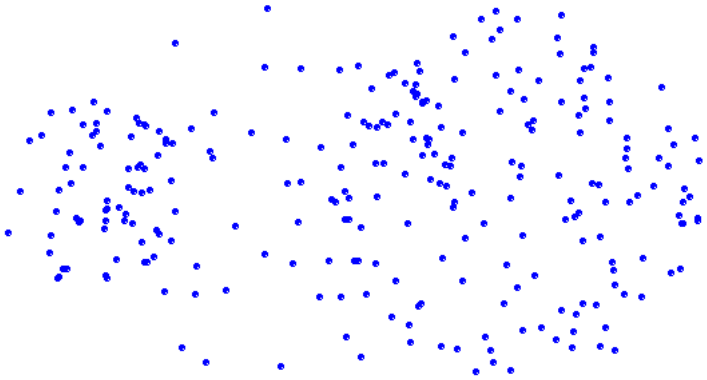
Puntos originales



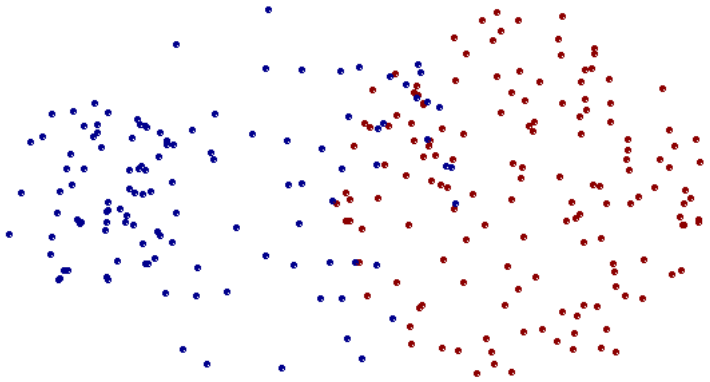
Dos Clusters

- **Puede manejar formas no elipticas**

Limitaciones de MIN o Enlace Simple



Puntos originales



Two Clusters

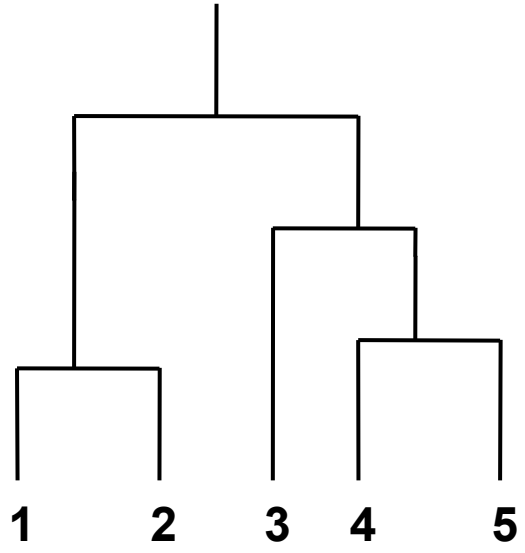
- **Sensible al ruido y valores atípicos**

Similaridad Inter-Cluster
MAX o Enlace Completo

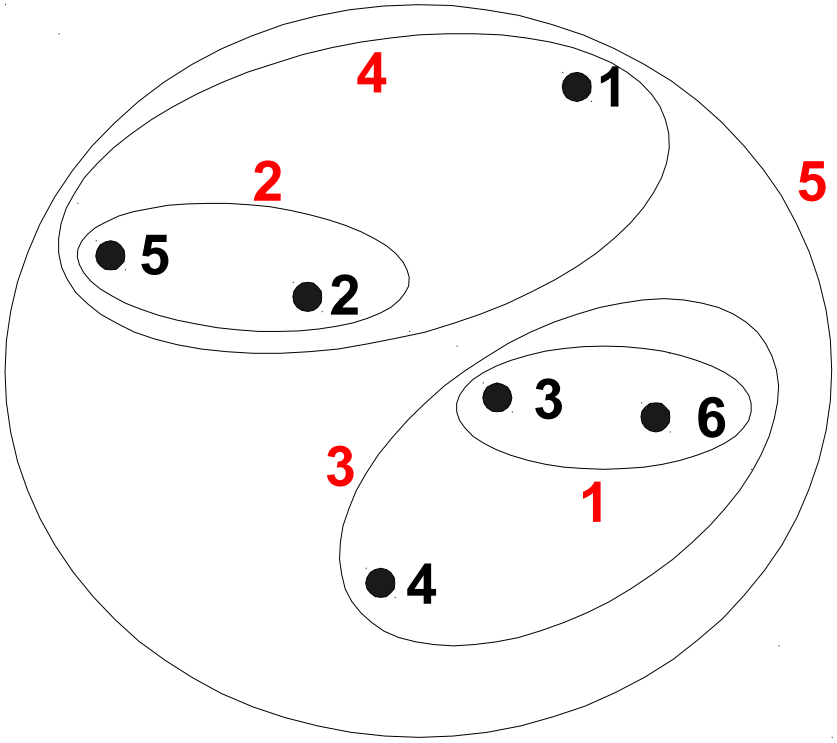
La similaridad de dos clusters esta basada en los dos mas distantes (mas diferentes) puntos de los clusters

Determinado por todos los pares de puntos de los clusters

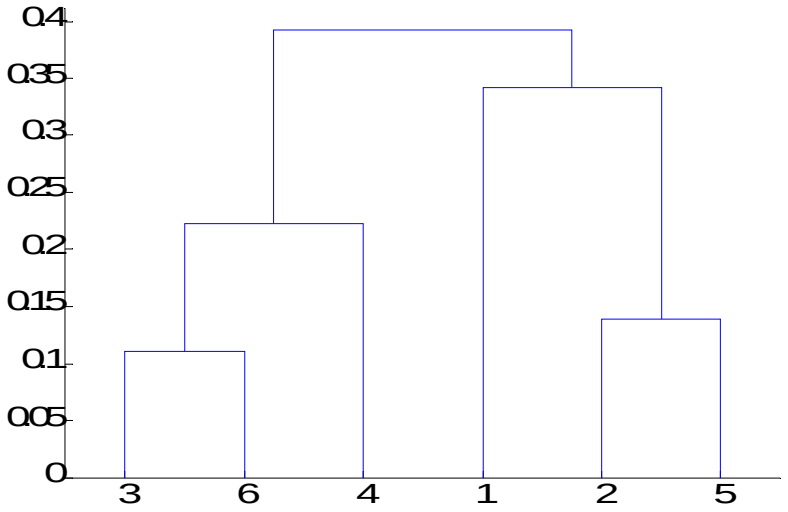
	I1	I2	I3	I4	I5
I1	1,00	0,90	0,10	0,65	0,20
I2	0,90	1,00	0,70	0,60	0,50
I3	0,10	0,70	1,00	0,40	0,30
I4	0,65	0,60	0,40	1,00	0,80
I5	0,20	0,50	0,30	0,80	1,00



Similaridad Inter-Cluster
MAX o Enlace Completo

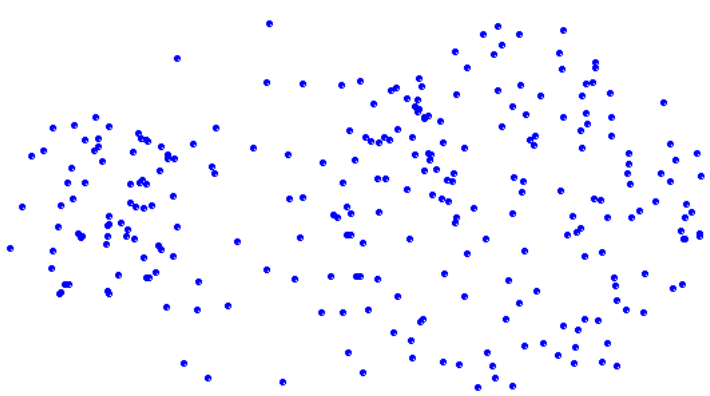


Clusters anidados

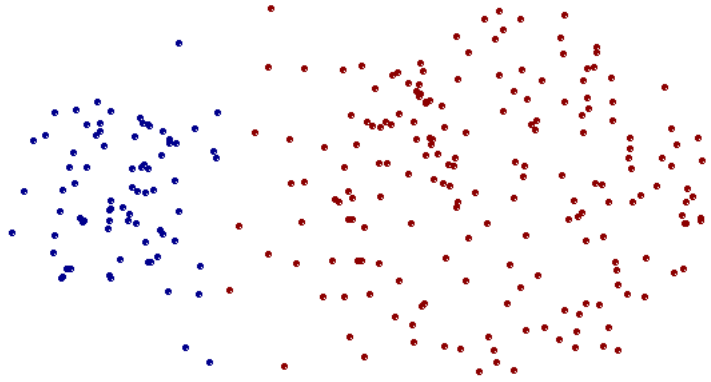


Dendrograma

Fortaleza de MAX o Enlace Completo



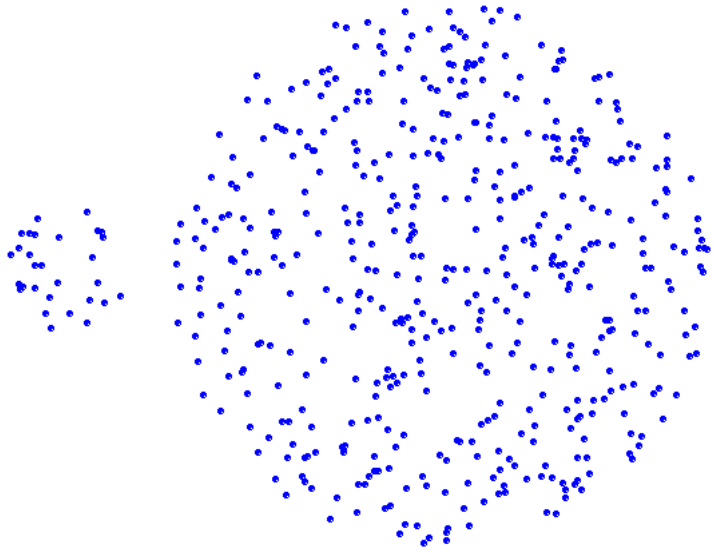
Puntos originales



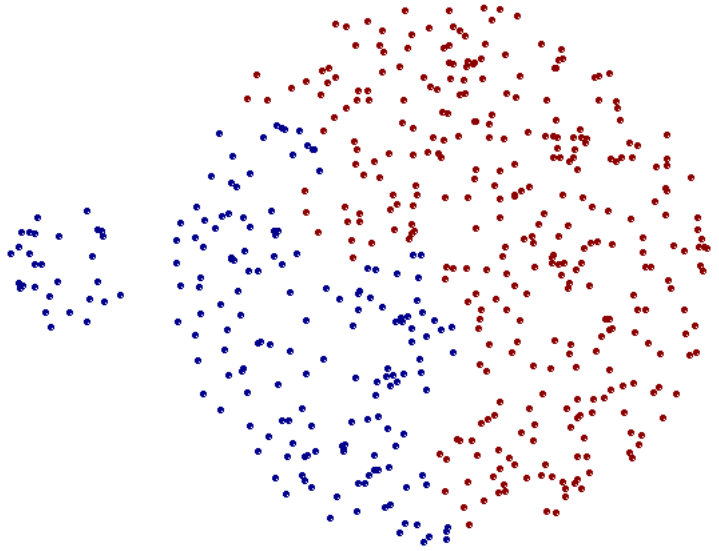
Dos Clusters

- **Menos suceptible a rudio y datos atípicos**

Limitación de MAX o Enlace Completo



Puntos Originales



Dos Clusters

- **Tiende a dividir grandes grupos**
- **Predispuesto para grupos globulares**

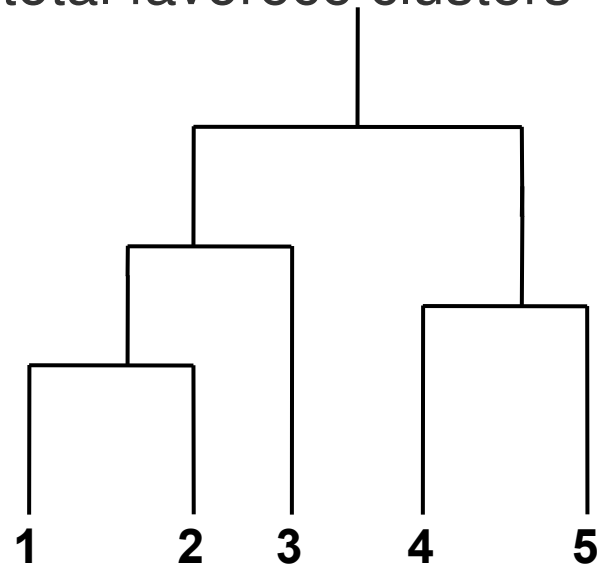
Similaridad Inter-Cluster
Promedio del cluster

Proximidad de dos clusters es el promedio de la proximidad de las parejas entre los puntos de los dos clusters

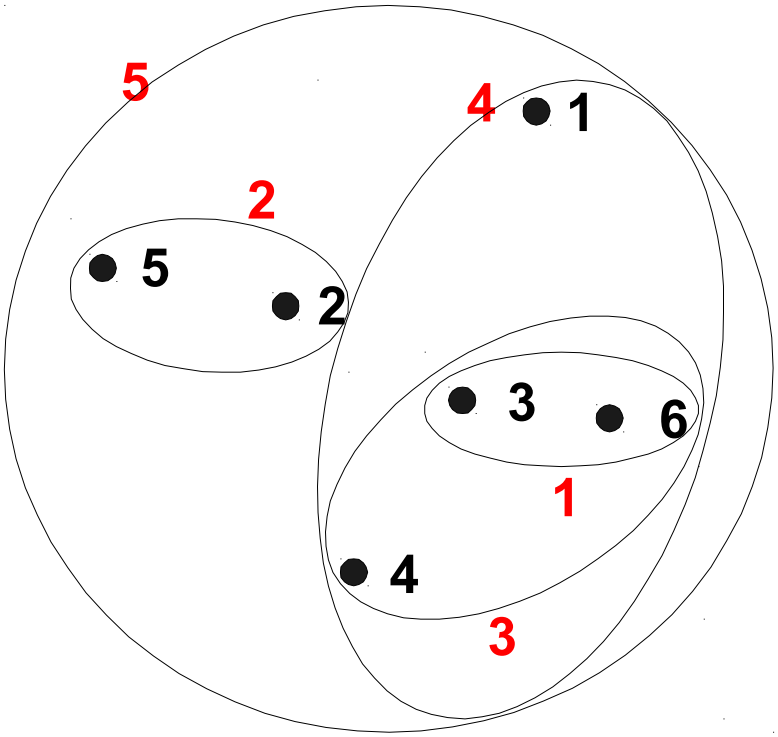
$$proximity(Cluster_i, Cluster_j) = \frac{\sum_{\substack{p_i \in Cluster_i \\ p_j \in Cluster_j}} proximity(p_i, p_j)}{|Cluster_i| * |Cluster_j|}$$

Es necesario usar el promedio para la conectividad y escalabilidad dado que la proximidad total favorece clusters grandes

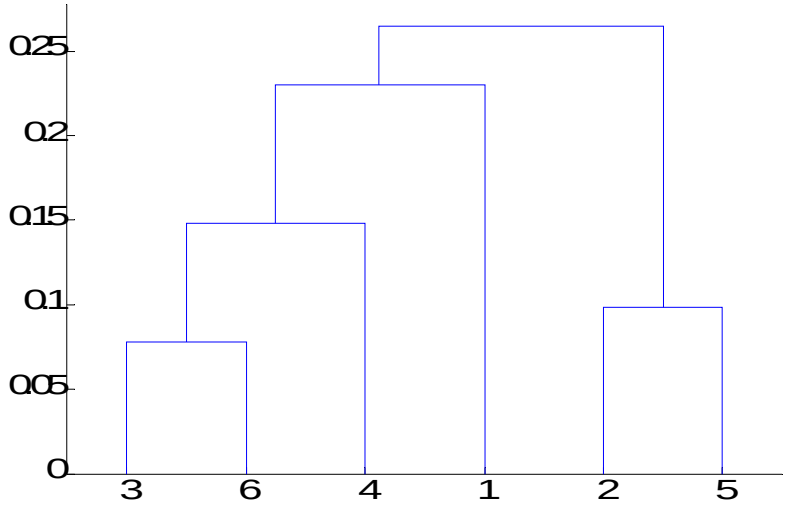
	I1	I2	I3	I4	I5
I1	1,00	0,90	0,10	0,65	0,20
I2	0,90	1,00	0,70	0,60	0,50
I3	0,10	0,70	1,00	0,40	0,30
I4	0,65	0,60	0,40	1,00	0,80
I5	0,20	0,50	0,30	0,80	1,00



Similaridad Inter-Cluster
Promedio del cluster



Clusters anidados



Dendrograma

Similaridad Inter-Cluster
Promedio del cluster

Compromete tanto el enlace simple como el completo

Fortalezas

- Menos susceptible al ruido y a los datos atípicos

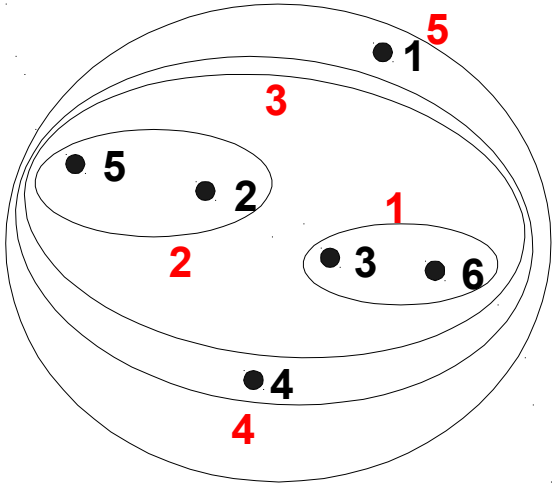
Limitaciones

- Predispuesto para clusters globulares

Similaridad Inter-Cluster
Metodo de Ward

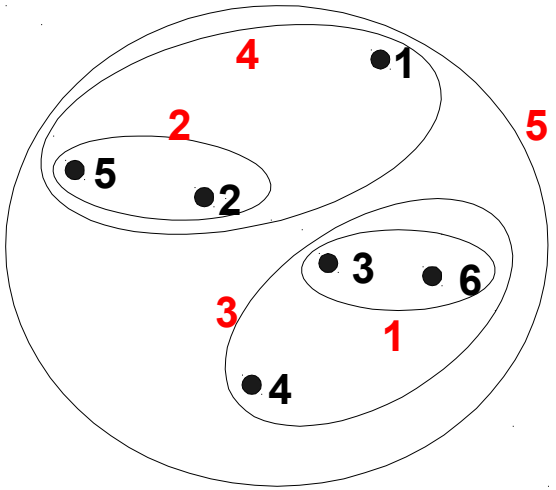
- La similitud de dos clusters se basa en el incremento del error cuadrático cuando dos clusters se combinan
  - Similar al promedio de grupo, si la distancia entre puntos es la distancia cuadrática
- Menos susceptible al ruido y a los datos atípicos
  Predispuesto para clusters globulares
- Jerarquía análoga al K - means
  Puede ser usada para inicializar K- means

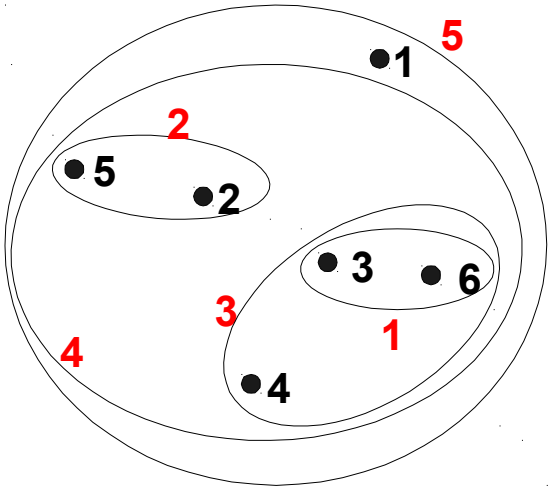
Similaridad Inter-Cluster Comparación



MIN

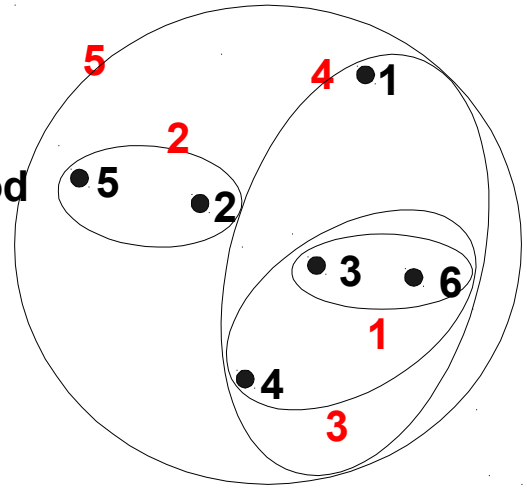
MAX





Group Average

Ward's Method



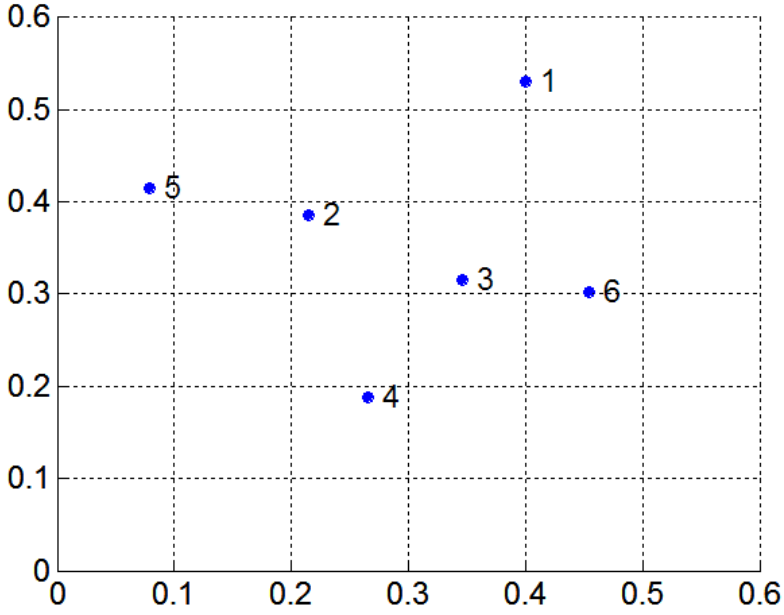
Agrupamiento jerárquico
Problemas y limitaciones

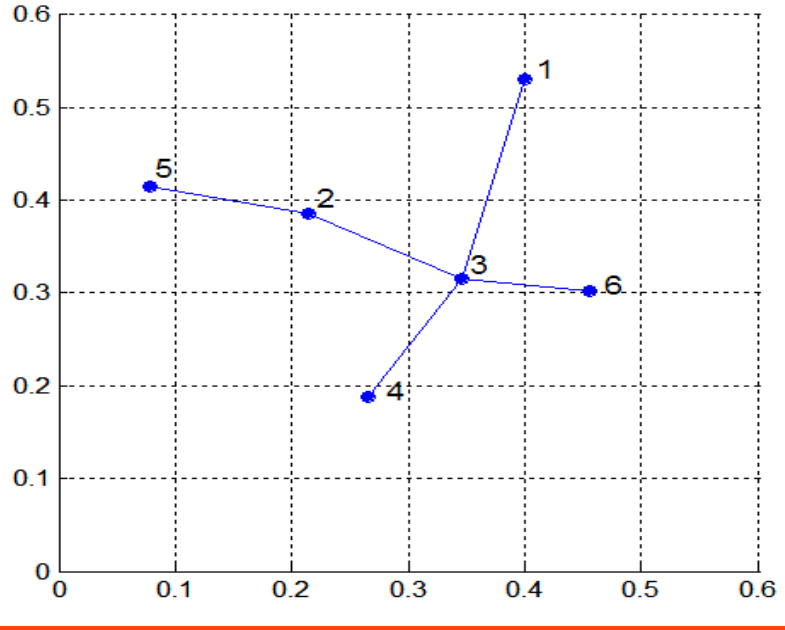
- Una vez se toma una decisión para combinar dos clusters, no puede deshacerse
- La función objetivo no es directamente minimizada
- Los esquemas diferentes tienen problemas con uno o más de los siguientes factores:
  - Sensibilidad al ruido y a los datos atípicos
  - Dificultad para manejar clusters de diferente
  - tamaño y formas convexas
  - Rompimiento de clusters grandes

MST: Agrupamiento Jerárquico Divisivo

Construir MST (Minimum Spanning Tree, Árbol de Mínima cobertura)

- Comenzar con un árbol que consiste en cualquier punto
- En los pasos sucesivos, buscar el par de puntos más cercanos (p, q), el punto “p” está en el árbol actual pero el otro “q” no lo esta
- Adicionar q al árbol y colocar un enlace entre p y q.





Bibliografía

"Introduction to Data Mining" by Tan, Steinbach, Kumar. Chapter 8

"Data Mining: Cluster Analysis: Basic Concepts and Algorithms"