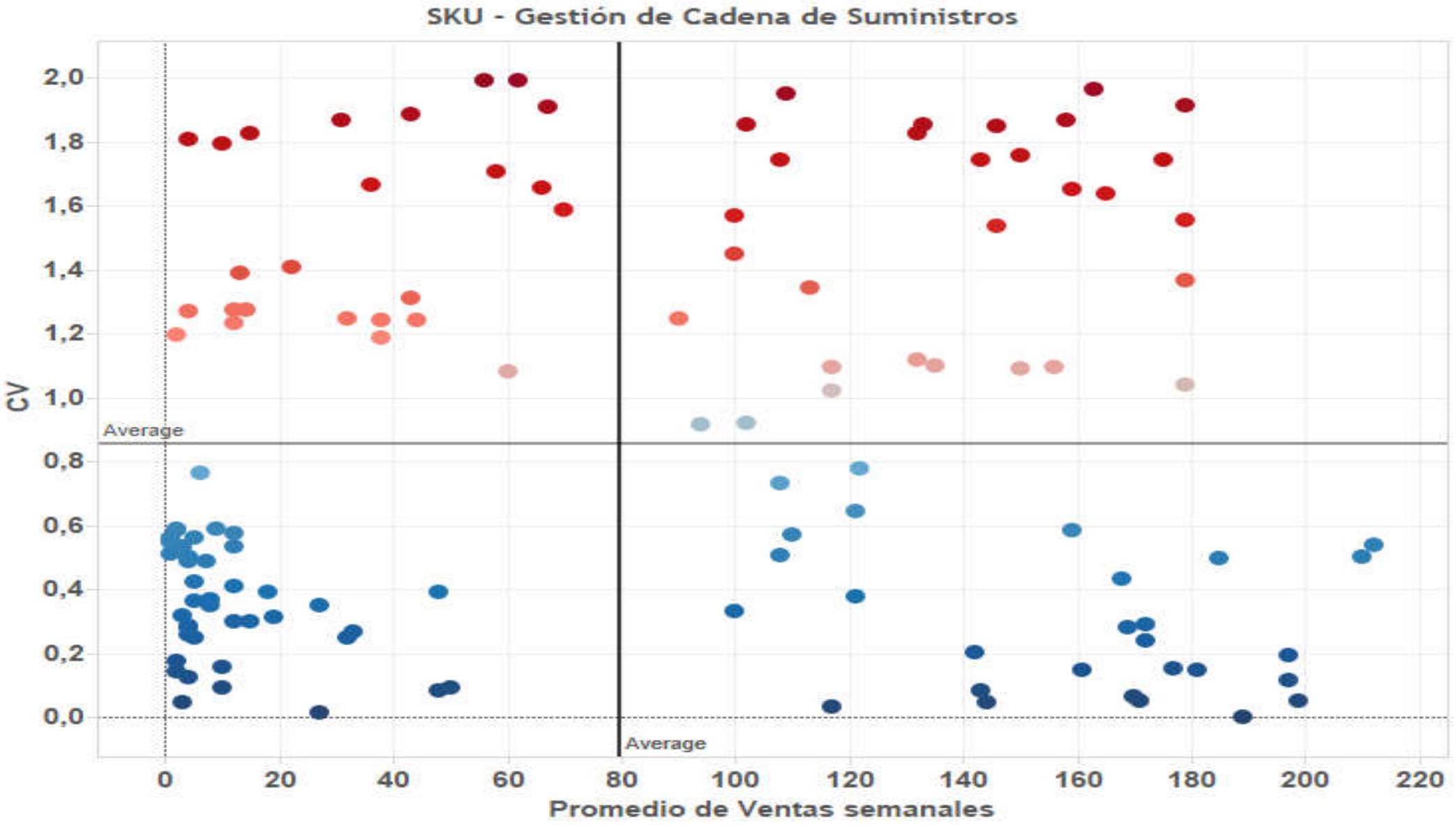
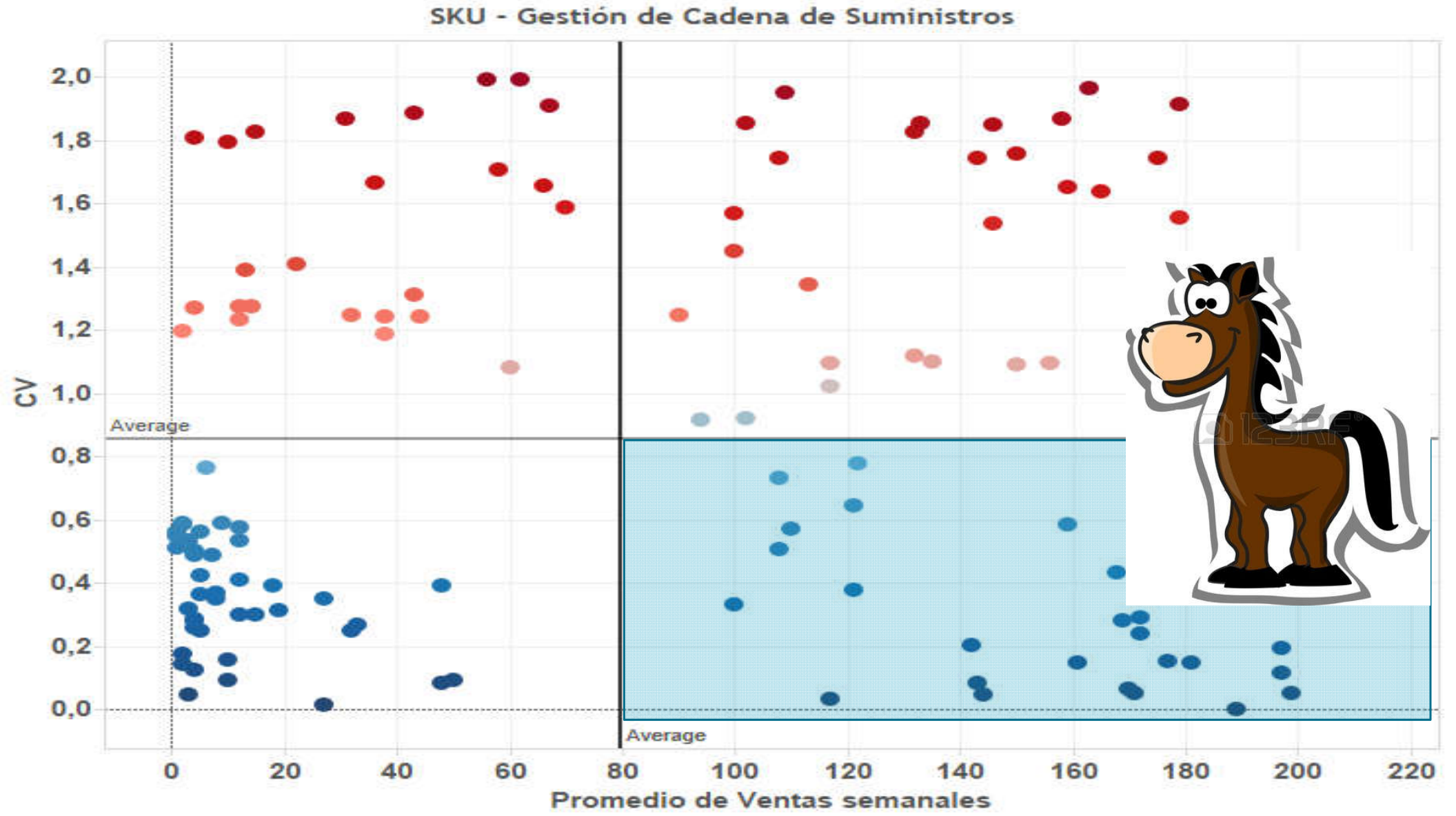
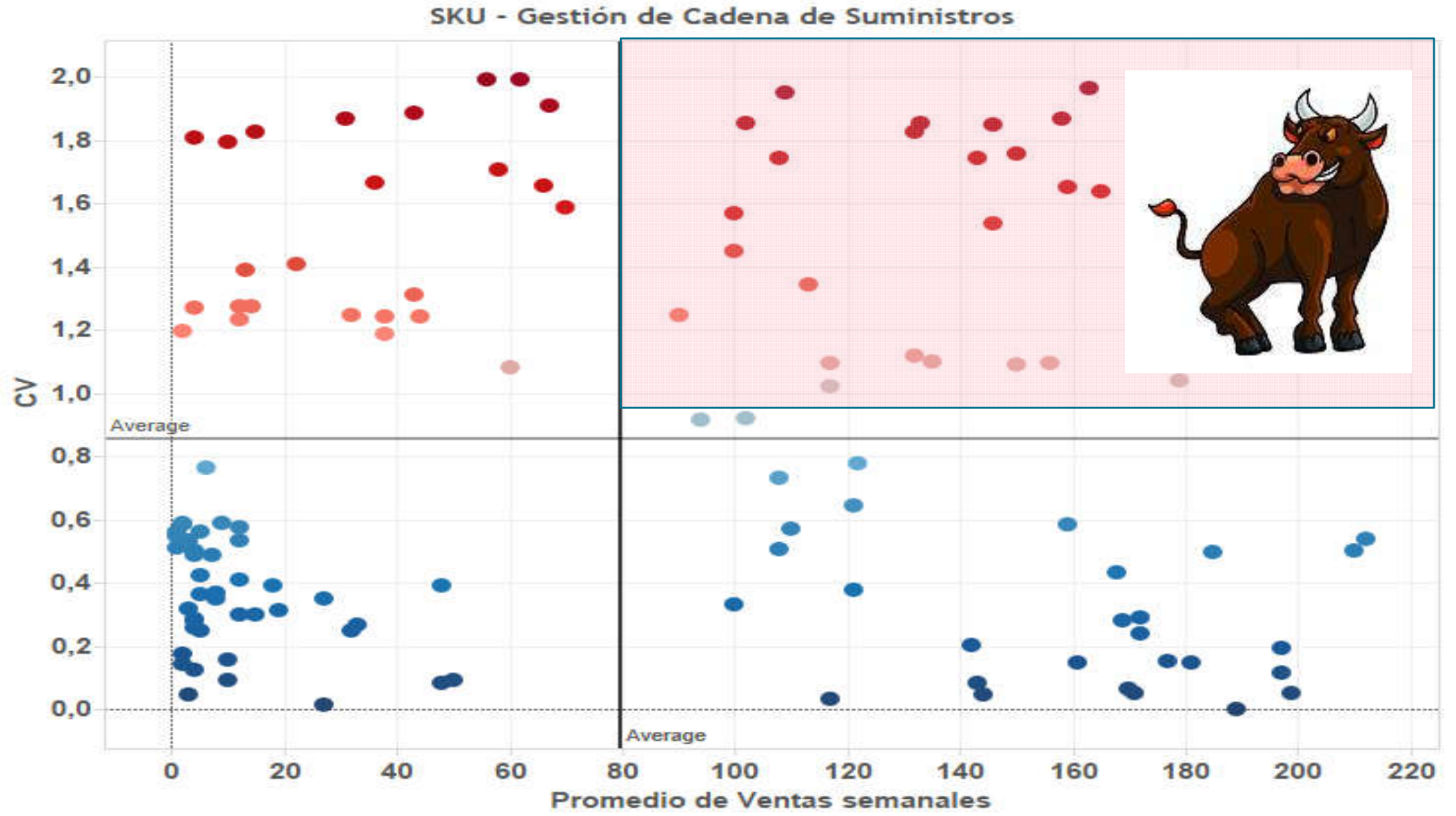
Minería de datos aplicada

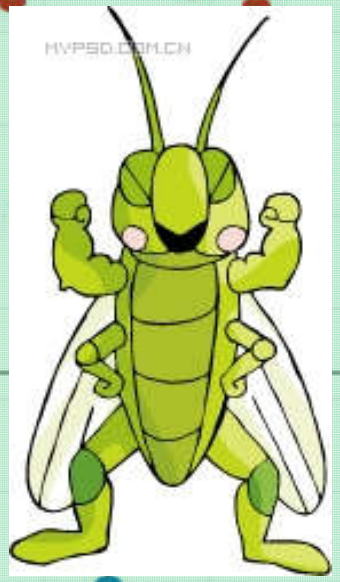
Clusters


SKU - Gestión de Cadena de Suministros
CV
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
Average
Average
0
20
40
60
80
100
120
140
160
180
200
220
Promedio de Ventas semanales


SKU - Gestión de Cadena de Suministros
CV
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
Average
Average
0
20
40
60
80
100
120
140
160
180
200
220
Promedio de Ventas semanales


SKU - Gestión de Cadena de Suministros
CV
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
Average
Average
0
20
40
60
80
100
120
140
160
180
200
220
Promedio de Ventas semanales

SKU - Gestión de Cadena de Suministros



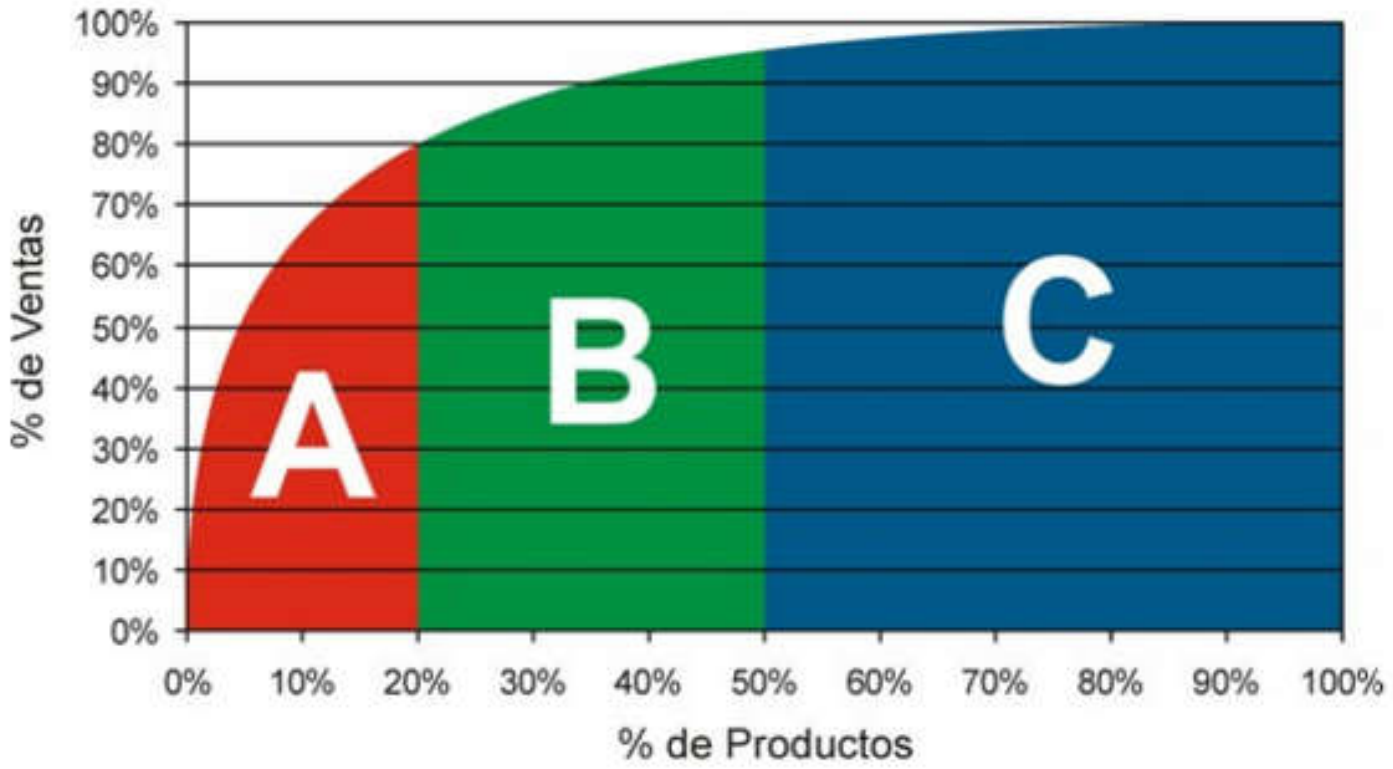
CV

Average

Average

Promedio de Ventas semanales

Segmentación ABC



Extraído de : http://slideplayer.es/slide/158616/

Preguntas ...

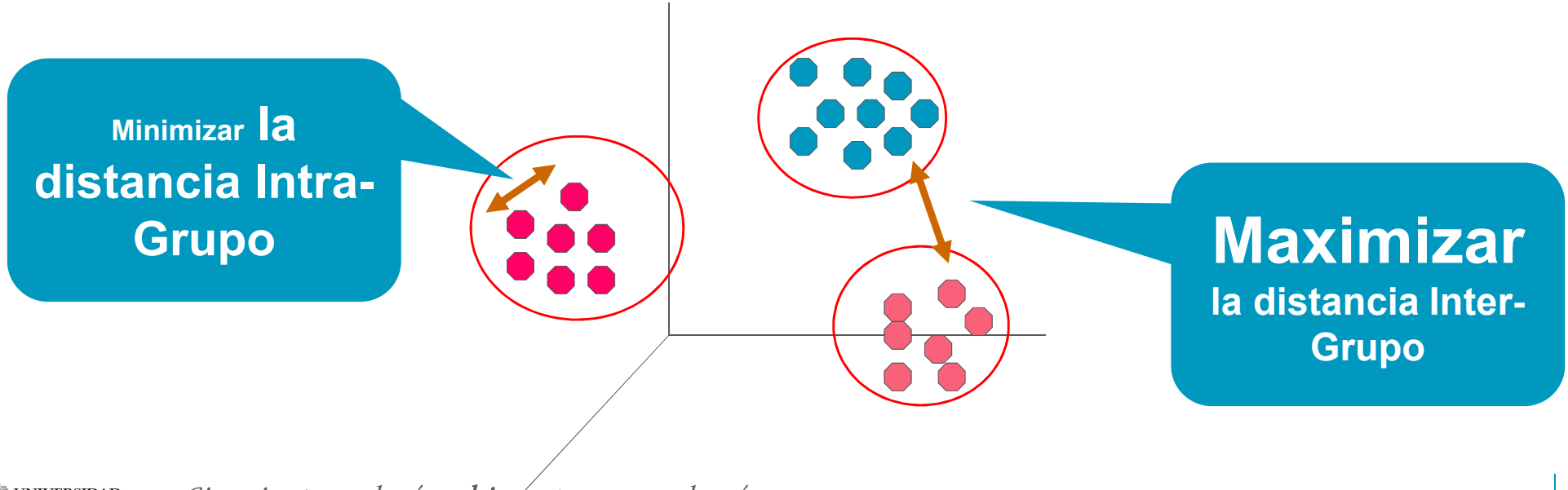
¿Existe otras formas de conglomerar?

¿Se podrá automatizar estos procesos?

¿En otros dominios diferentes a el de gestión de cadena de suministros puedo aplicar conceptos parecidos?

¿ Que es el análisis de Clúster?

Es hallar grupos tales que los objetos de un grupo sean similares unos a otros (Relacionados) y diferentes a otros miembros de otro grupo (No relacionados)



¿ Que es el análisis de Clúster?

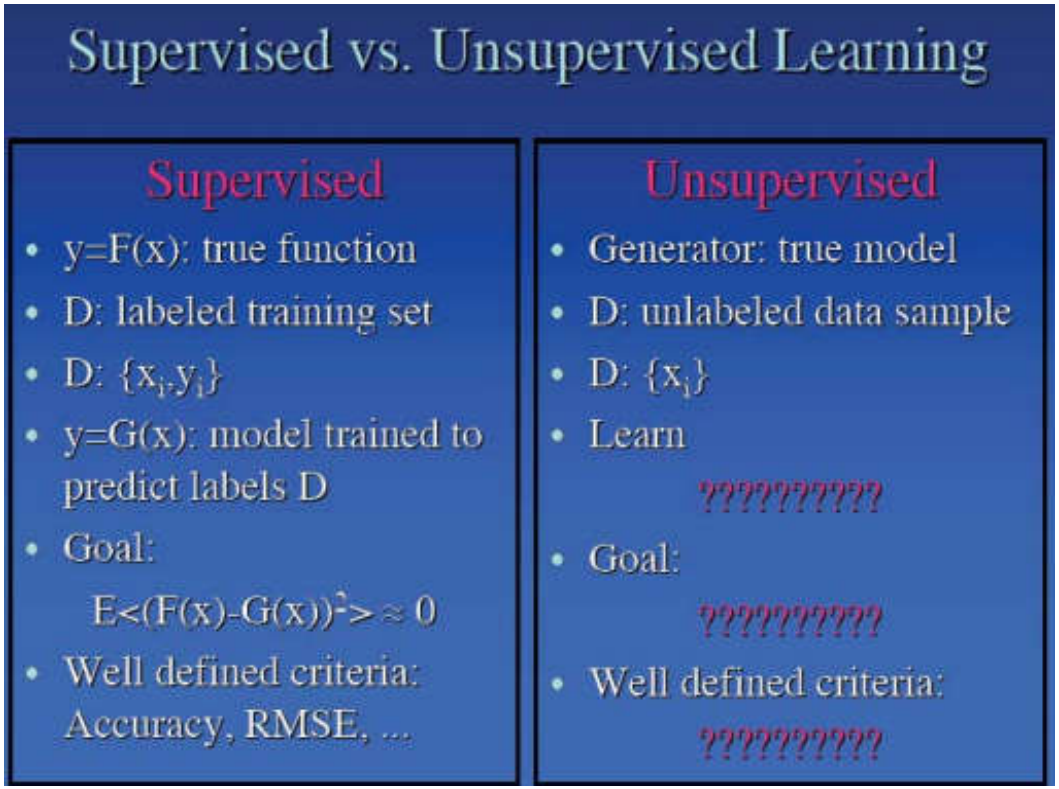
- Dado un conjunto de archivos (Observaciones, instancias , objetos, ejemplos, …) organizarlos dentro de grupos (clases, segmentos, clúster)

- Un Clúster es un subconjunto de elementos que son “similares”

- Una región conectada de un espacio multidimensional que contiene una densidad relativamente alta de los objetos.

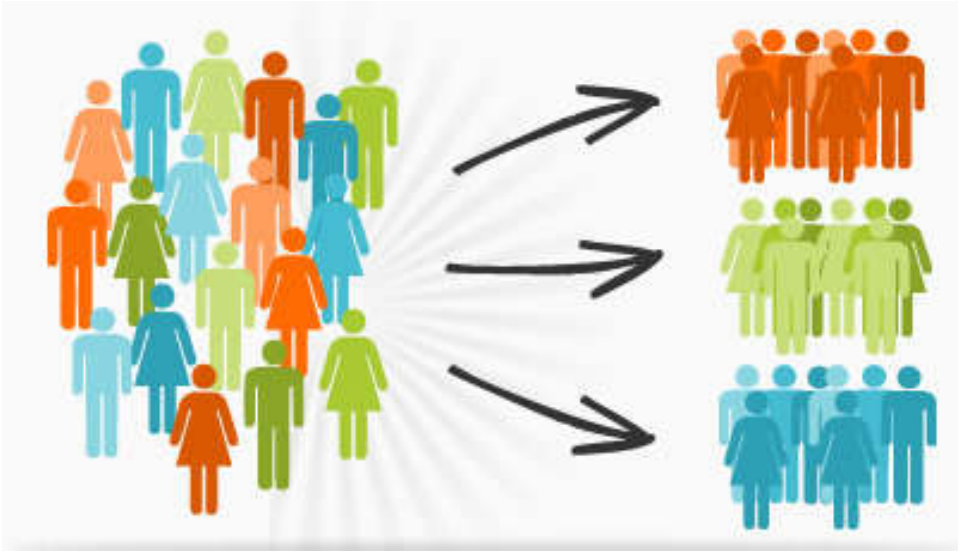
¿ Que es NO el análisis de Clúster?

- Clasificación supervisada, Recordemos →



¿ Que es NO el análisis de Clúster?

- **Segmentación simple**
  - La división de los alumnos en los diferentes grupos de registro por orden alfabético, por el apellido, etc.



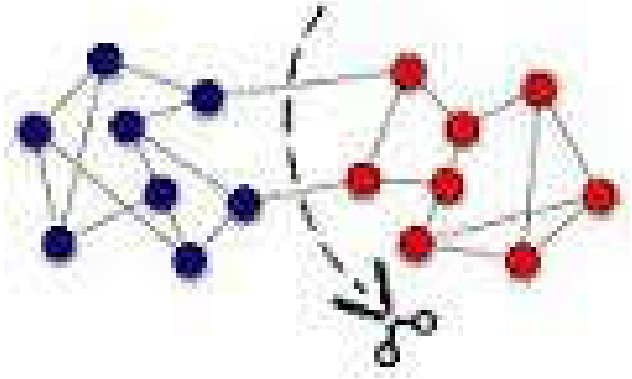
¿ Que es NO el análisis de Clúster?

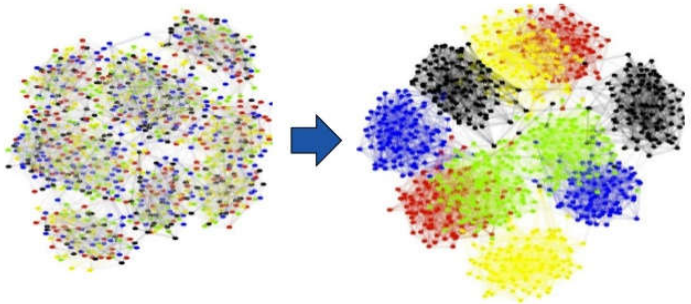
- **Los resultados de una consulta**
  - Agrupaciones son el resultado de una especificación externa

Quantity Sum (% of each column total)		ACCESSORIES	DIAGNOSTIC KITS	REPAIR KITS	All Products B1
AMERICA	CANADA	21.97%	21.49%	21.31%	**21.59%**
	USA	24.44%	26.02%	18.26%	**22.47%**
	AMERICA	**46.41%**	**47.51%**	**39.57%**	**44.06%**
ASIA	CHINA	8.13%	14.42%	12.08%	**11.31%**
	JAPAN	11.32%	6.62%	14.10%	**11.16%**
	ASIA	**19.45%**	**21.04%**	**26.18%**	**22.47%**
EUROPE	FRANCE	16.37%	4.01%	11.31%	**11.16%**
	GERMANY	7.85%	16.73%	12.23%	**11.88%**
	ITALY	9.92%	10.71%	10.71%	**10.43%**
	EUROPE	**34.14%**	**31.45%**	**34.25%**	**33.47%**
All		**100.00%**	**100.00%**	**100.00%**	**100.00%**

¿ Que es NO el análisis de Clúster?

- **Simple particionamiento de un grafo**
  - Algunos relevancia mutua y la sinergia, pero las áreas no son idénticos





Tomato de : http://www.slideshare.net/anisaadi/gossip-based-partitioning-and-replication-for-online-social-networks

En esencia lo que buscamos es ayuda en entender la agrupación "natural " o estructura de un conjunto de datos.

Aun así no es toda la historia, nos enfrentamos algunos retos y objetivos adicionales que estas técnicas no permiten ……

Clúster como herramienta de Pre procesamiento

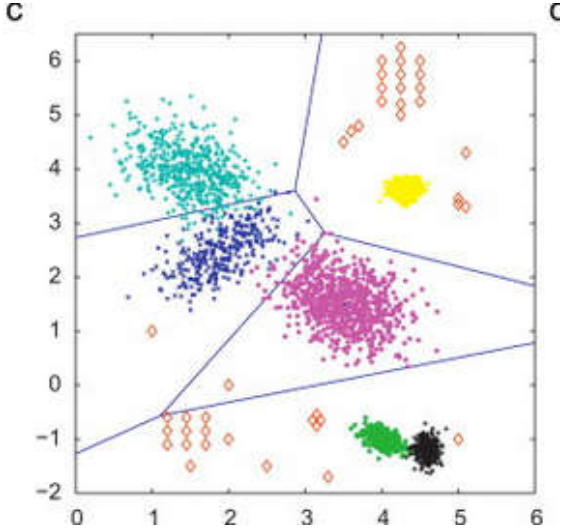
- **Sumarización:**
  - Reprocesamiento de regresión, PCA, clasificación y análisis de asociación
- **Compresión:**
  - Procesamiento de imágenes: cuantificación vectorial
- **Encontrar K-vecinos más cercanos**
  - La localización de búsqueda a uno o un pequeño número de conglomeradosla detección de valores atípicos
- **Los valores atípicos**
  - son a menudo vistos como los "muy lejos" de cualquier grupo

Requerimientos y Retos

- **Escalabilidad**
  - La agrupación de todos los datos en lugar de sólo en muestras
- **Capacidad para hacer frente a diferentes tipos de atributos**
  - Numérico, binario, categórica, ordinal, vinculado, y mezclas de los mismos
- **Clustering basado en restricciones**
  - El usuario puede dar aportaciones sobre restricciones
  - Utilizar el conocimiento de dominio para determinar los parámetros de entrada
- **Interpretabilidad y usabilidad**
- **Otros**
  - Descubrimiento de clusters con forma arbitraria
  - Capacidad para hacer frente a los datos de ruido
  - Agrupación incremental y la insensibilidad a la orden de entrada
  - alta dimensionalidad

Enfoques principales de clusterización

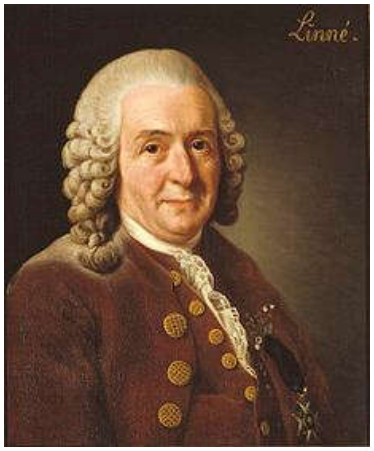
- **Enfoque de particionamiento:**
  - Construir diversas particiones y luego evaluar por algún criterio, por ejemplo, minimizando la suma de errores cuadráticos
  - Los métodos típicos: K-means, K-medoides, CLARANS

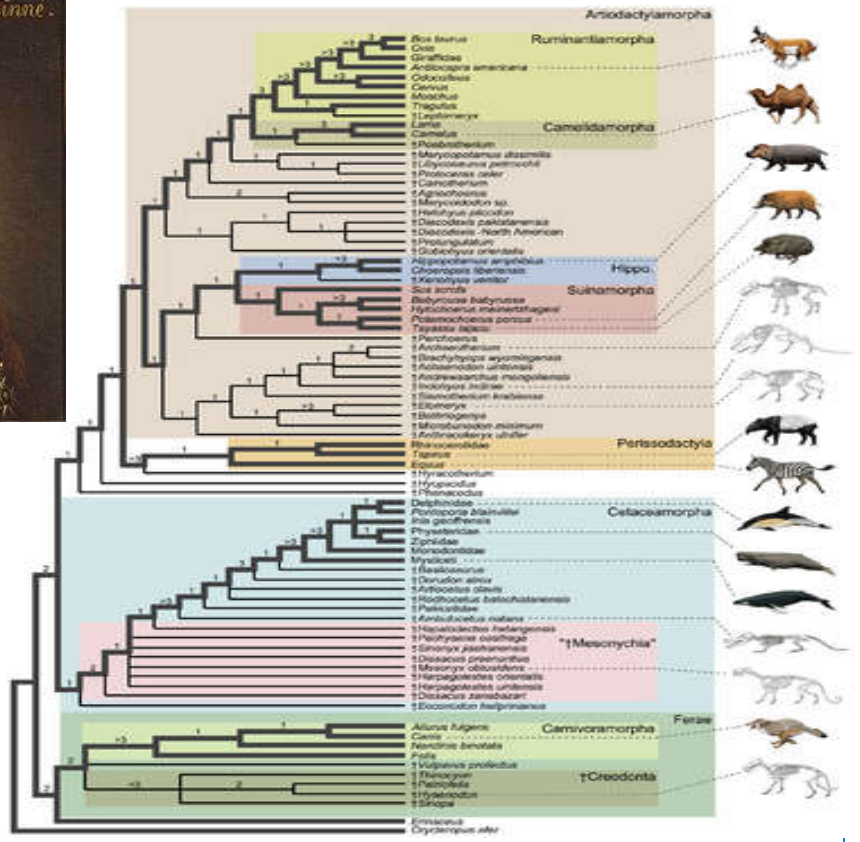


Enfoques principales de clusterización

- **Enfoque jerárquico:**

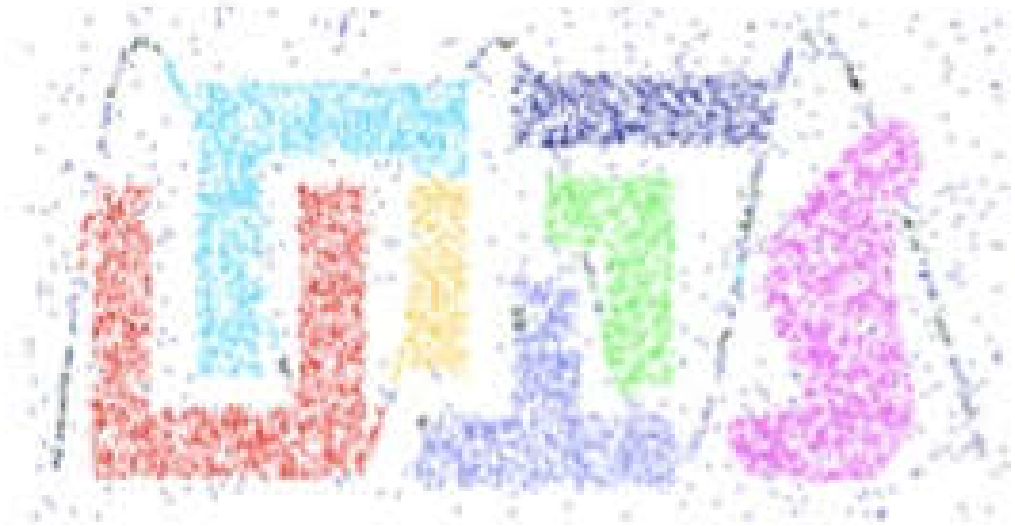
- Crear una descomposición jerárquica del conjunto de datos (u objetos) utilizando algún criterio
- Los métodos típicos: Diana, Agnes, abedul, CAMELEON



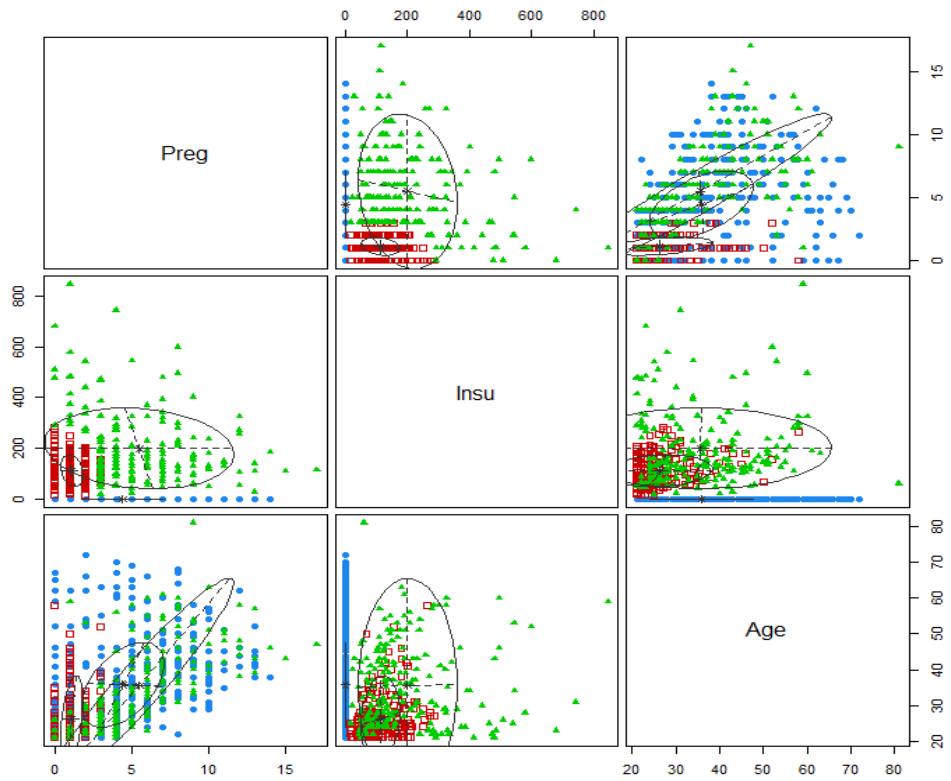


Enfoques principales de clusterización

- **Enfoque basado en la densidad:**
  - En base a funciones de conectividad y densidad
  - Los métodos típicos: DBSCAN, OPTICS, DenClue



Enfoques principales de clusterización



Basado en Modelo:

- Un modelo es la hipótesis para cada uno de los grupos y trata de encontrar el mejor ajuste de ese modelo a la otra
- Los métodos típicos: EM, SOM, COBWED

Enfoques principales de clusterización

- **Enfoque basado en Grid:**
  - basado en una estructura de varios niveles de granularidad
  - Los métodos típicos: STING, WAVECLUSTER, CLIQUE

- **Basada patrón frecuente:**
  - Basándose en el análisis de los patrones frecuentes
  - Los métodos típicos: p-Cluster

- **Basado en restricción:**
  - La agrupación, considerando las limitaciones especificadas por el usuario o específicas de la aplicación
  - Los métodos típicos: COD (obstáculos), el agrupamiento limitado

- **Basada Enlace :**
  - Los objetos son a menudo vinculados entre sí de diversas maneras
  - Enlaces masivos se pueden utilizar para los objetos de clúster: SimRank, LinkClus

Algoritmo particional : conceptos básicos

Método de creación de particiones: Partición de una base de datos D de n objetos en un conjunto de grupos k, tal que la suma de las distancias al cuadrado se reduce al mínimo (donde ci es el centro de gravedad y medois de clúster Ci)

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

Dado k, encontrar una partición de clusters k que optimiza el criterio de partición elegida

- Óptimo global: enumerar exhaustivamente todas las particiones
- Los métodos heurísticos: k-medias y algoritmos k-medoides
- k-medias (MacQueen'67, Lloyd'57 / '82): Cada grupo está representado por el centro del cúmulo
- K-medoides o PAM (Partición alrededor medoides) (Kaufman y Rousseeuw'87): Cada grupo está representado por uno de los objetos del clúster

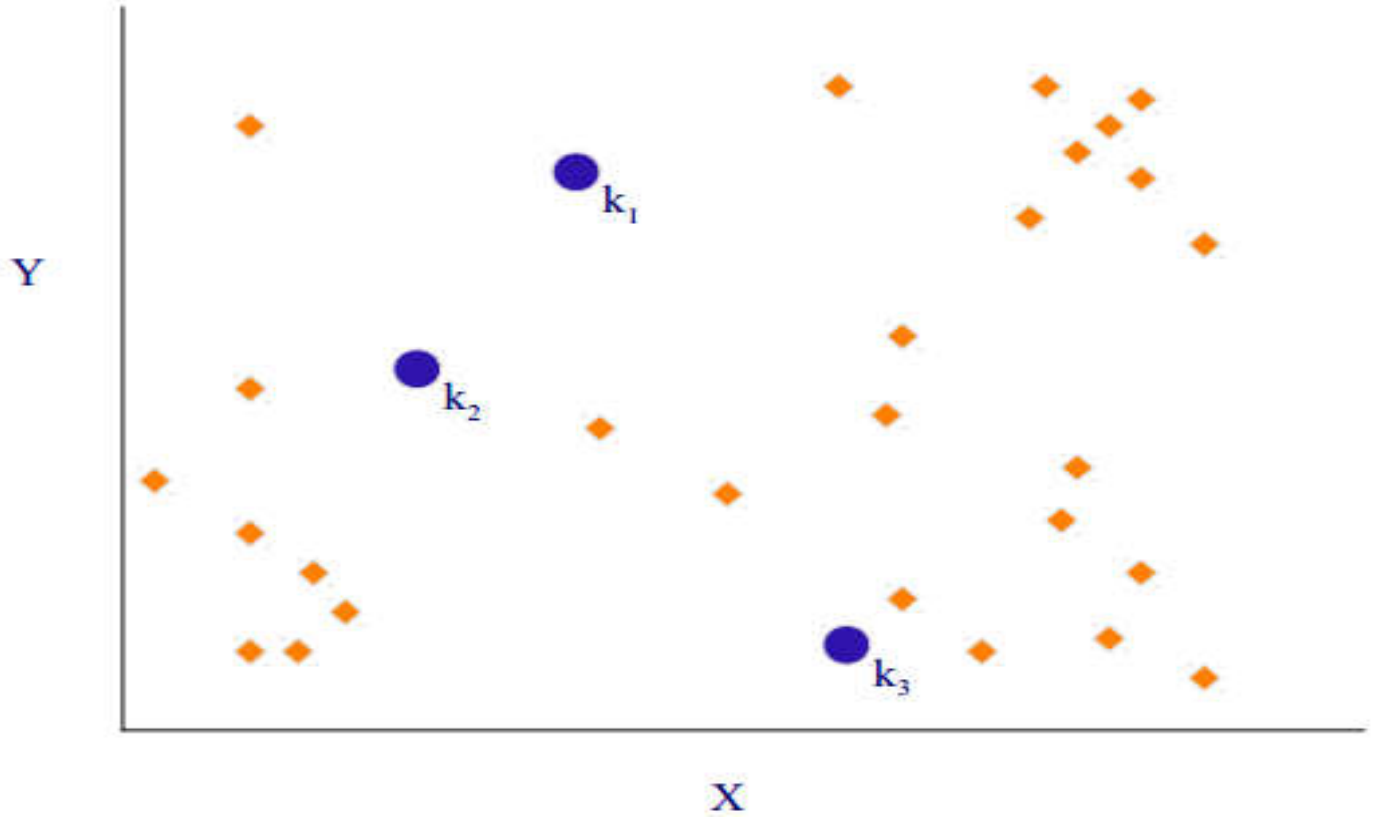
K- Means

- Enfoque de agrupación partitional
- Cada grupo está asociado con un (punto central) *centroide*
- Cada punto se asigna a la agrupación con el centroide más cercano
- Número de racimos, K, se debe especificar
- El algoritmo básico es muy simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

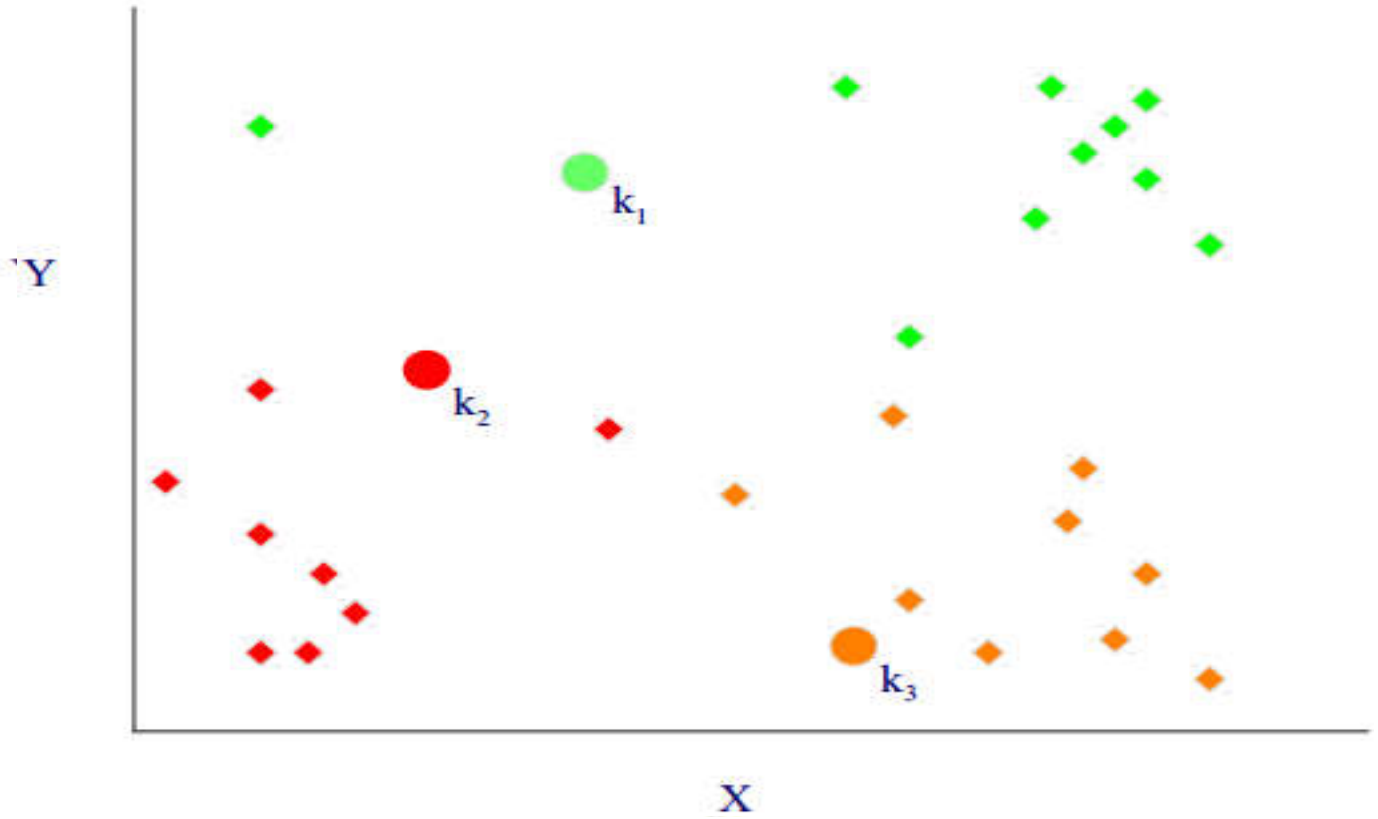
K- Means : Ejemplo, Paso 1

Iniciar los tres centros de los clúster (aleatoriamente)



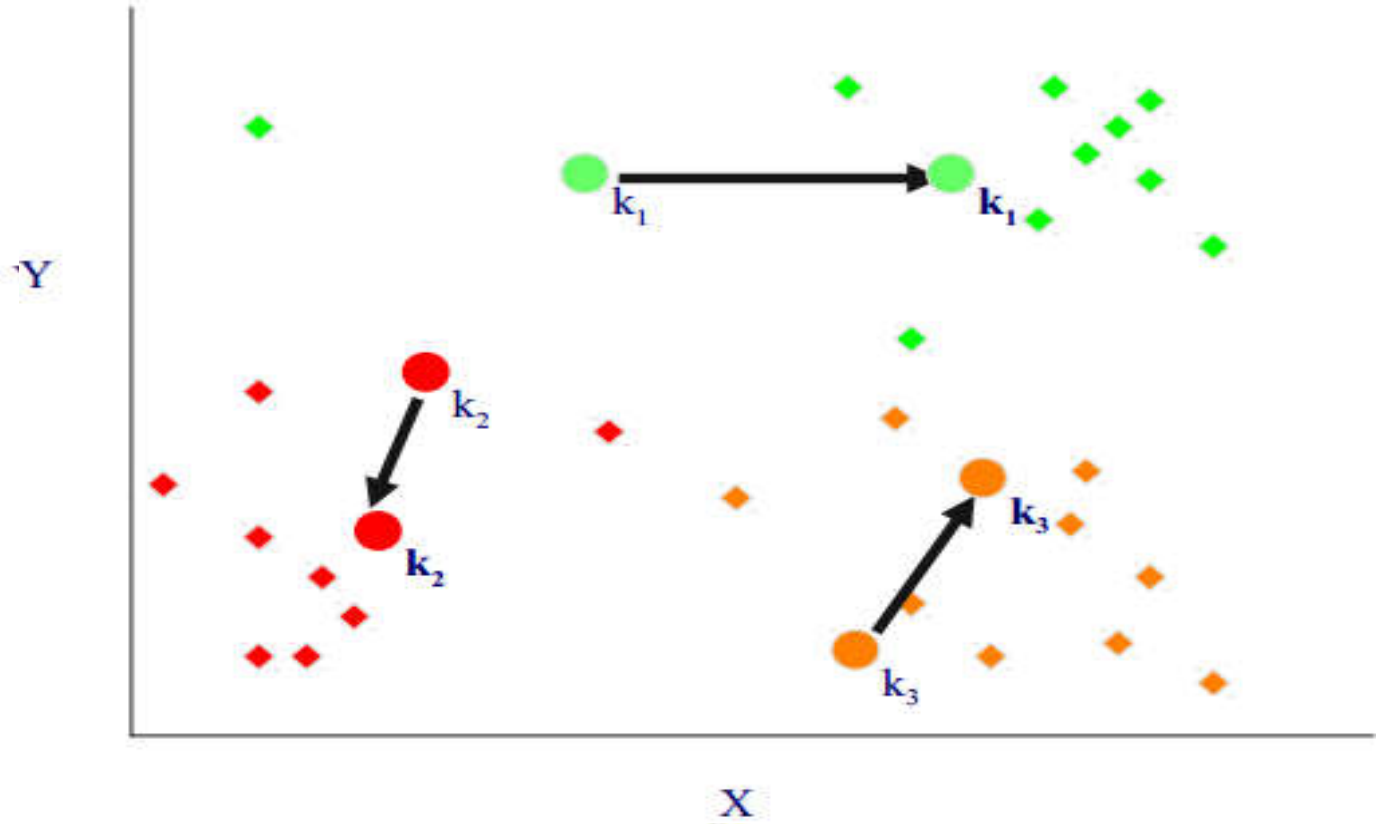
K- Means : Ejemplo, Paso 2

Asignar cada registro al centro de clúster mas cercano



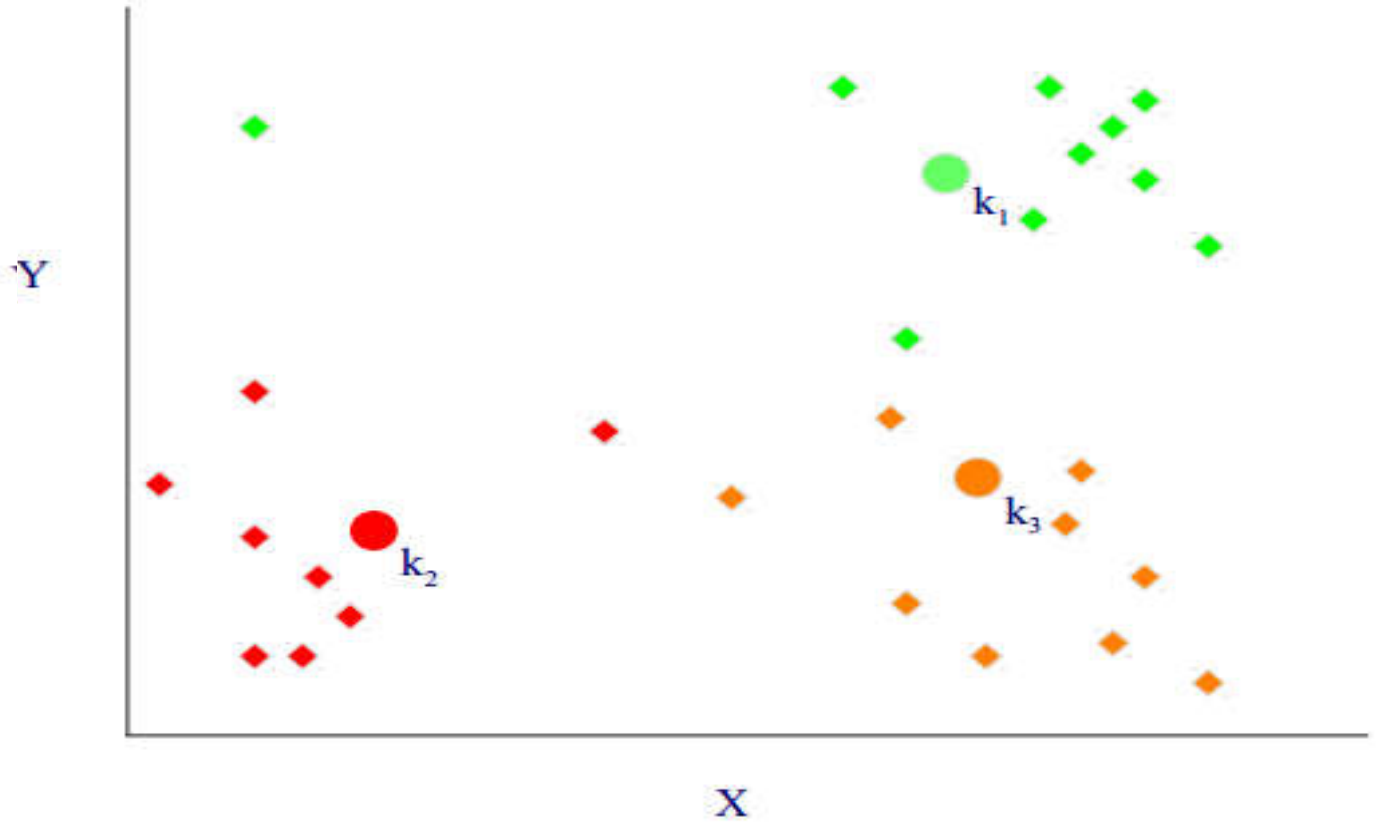
K- Means : Ejemplo, Paso 3

Desplazar cada centro clúster a la media de cada clúster

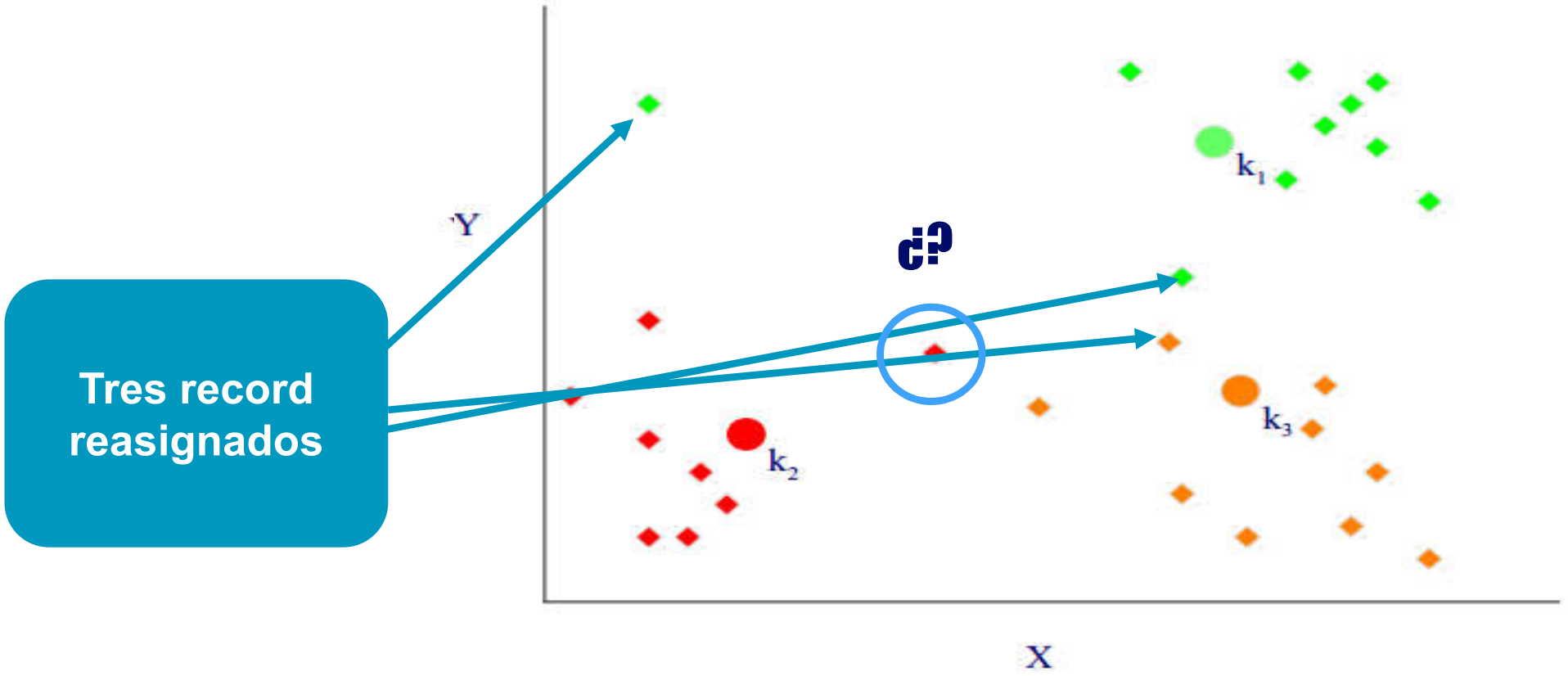


K- Means : Ejemplo, Paso 4

Reasignar los records al nuevo centro de clúster mas cercano

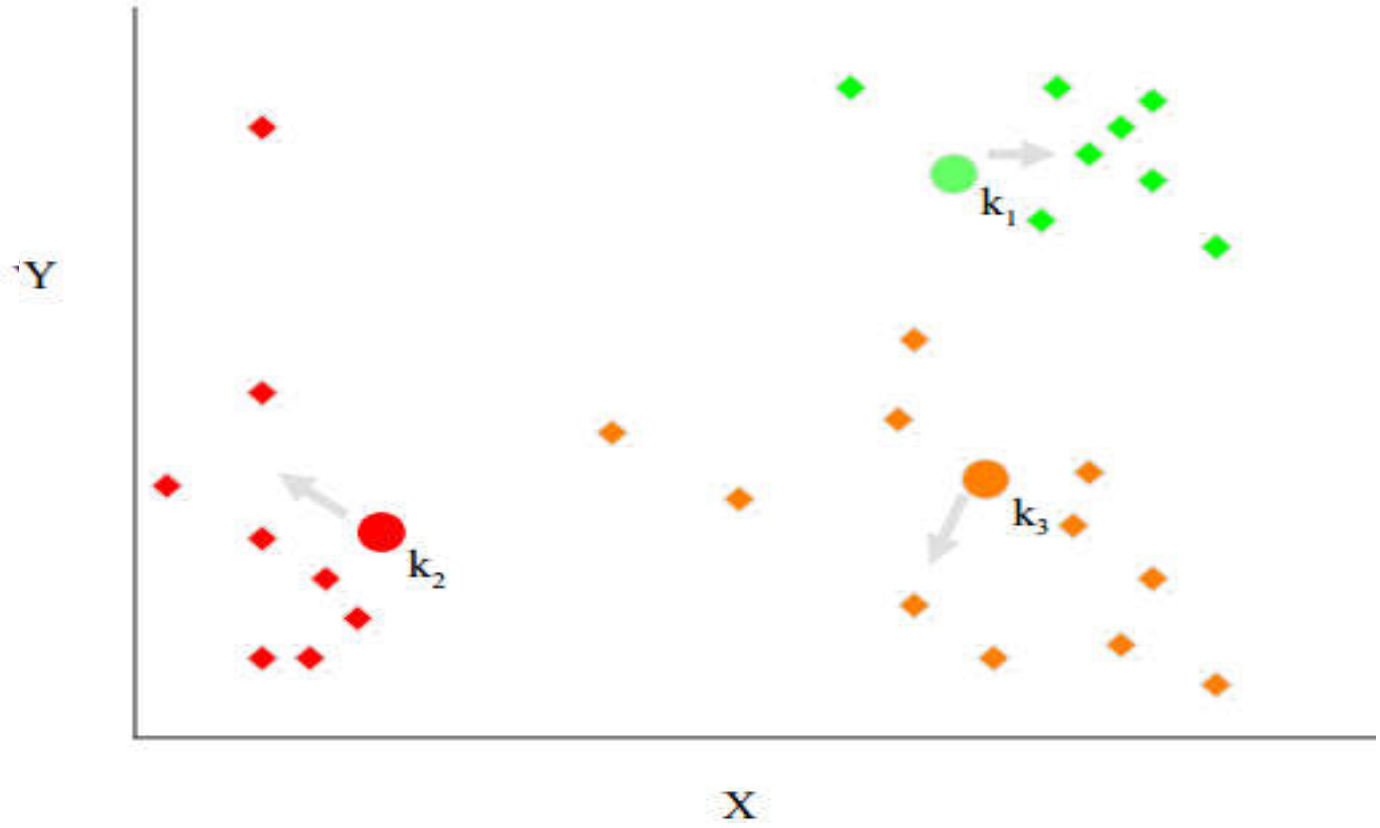


K- Means : Ejemplo, Paso 4



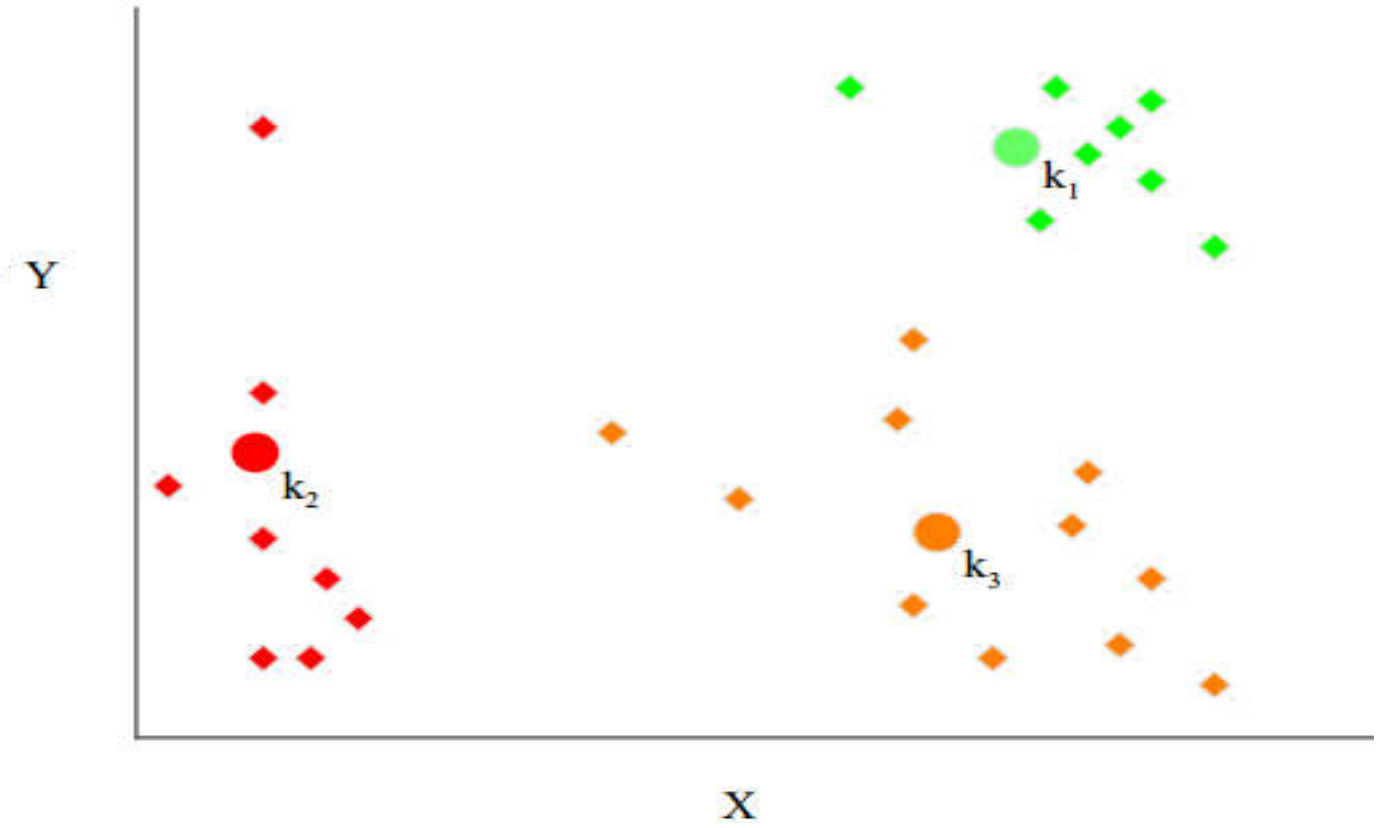
K- Means : Ejemplo, Paso 5

Re-calcular las medias de los clúester



K- Means : Ejemplo, Paso 1

Mover los centros a las medias de los clúster



K – Means : algunos detalles

- Centroides iniciales a menudo se eligen al azar.
- Clúster producidos varían de una corrida a otra.
- El centroide es (normalmente) la media de los puntos en el cluster.
- 'Cercanía' se mide por la distancia euclídea (en seguida miraremos otras), la similitud del coseno, correlación, etc.
- K-means convergerán para medidas de similitud comunes mencionados anteriormente.

K – Means : algunos detalles

- La mayor parte de la convergencia ocurre en las primeras iteraciones.
- A menudo, la condición de parada se cambia a 'Hasta que relativamente pocos puntos cambian'
- La complejidad es O (n * K * I * D)
- n = número de puntos, K = número de grupos, I = número de iteraciones, d = número de atributos

¿Que otros criterios de convergencia utilizarías?

¿En que situaciones?

K – Means : Algunos Criterios de convergencia

- Sin reasignaciones de puntos de datos a diferentes grupos (o mínimo)
- Sin (o mínimo) cambio de centroides, o
- Disminución mínima en la suma de errores cuadrados (SSE)
  - Ver siguiente diapositiva
- Detener después de X iteraciones

Evaluar los clúster K-means

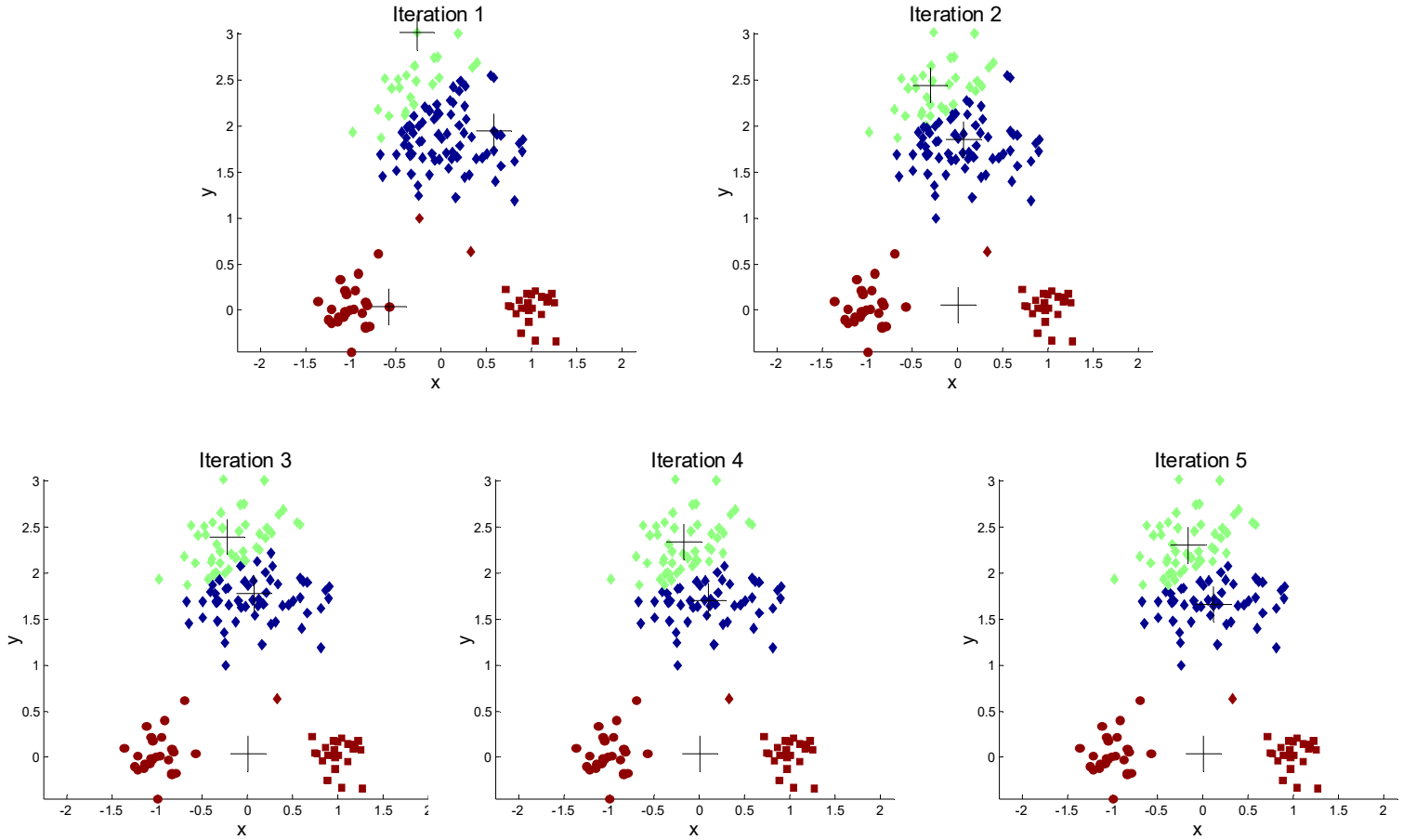
La mayor medida común es Suma de Squared Error (SSE)

- Para cada punto, el error es la distancia más cercana al clúster
- Para obtener SSE, elevamos al cuadrado estos errores y la suma de ellos.

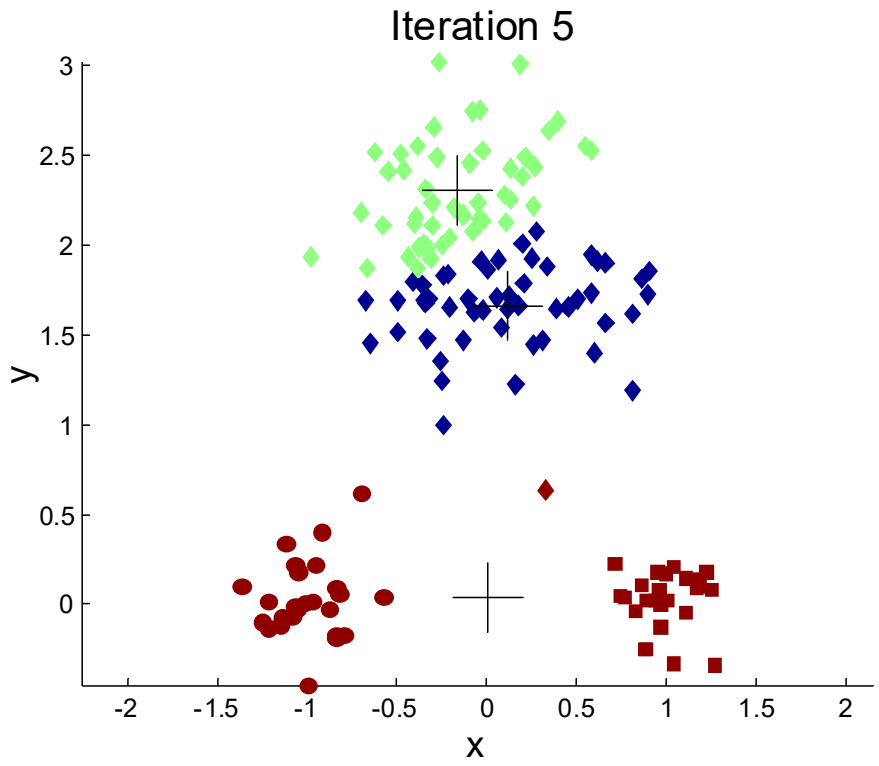
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- x es un punto de datos en el grupo Ci y Mi es el punto representativo de clúster Ci
- puede mostrar que mi corresponde al centro (media) del clúster
- Dadas dos grupos, podemos elegir el que tenga el más mínimo error
- Una manera fácil de reducir SSE es aumentar K, el número de grupos
- Una buena agrupación con menor K puede tener una SSE más baja que un pobre agrupación con mayor K

Debilidades de K-means: Semillas iniciales



Debilidades de K-means: Semillas iniciales



Debilidades de K-means: Semillas iniciales

Si hay agrupaciones K 'reales' y luego la oportunidad de seleccionar uno centroide de cada grupo es pequeño.

El azar es relativamente pequeño cuando K es grande

Si clusters son del mismo tamaño, n, entonces:

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K! n^K}{(Kn)^K} = \frac{K!}{K^K}$$

Por ejemplo, si K = 10, entonces la probabilidad = 10! / 1,010 = 0,00036

A veces los centroides iniciales ajustarse de manera "correcta", ya veces no lo hacen

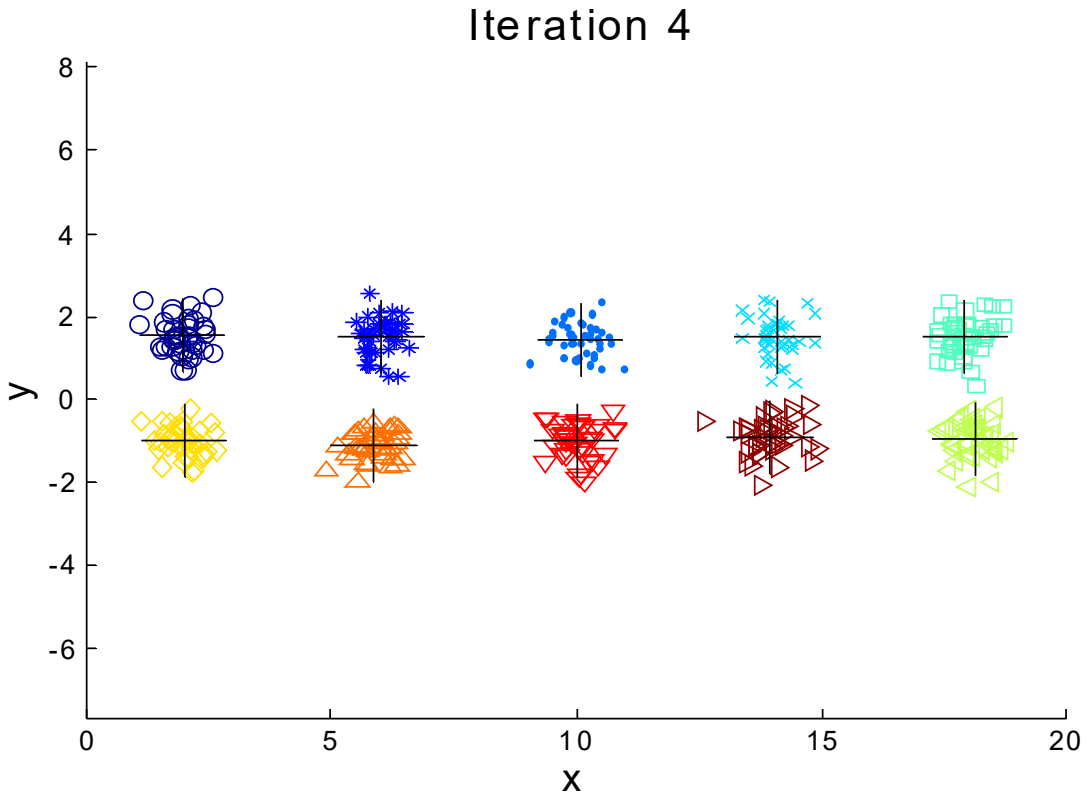
Considere un ejemplo de cinco pares de grupos

Debilidades de K-means: Semillas iniciales

Enfoques para aumentar la posibilidad de encontrar buenos grupos:

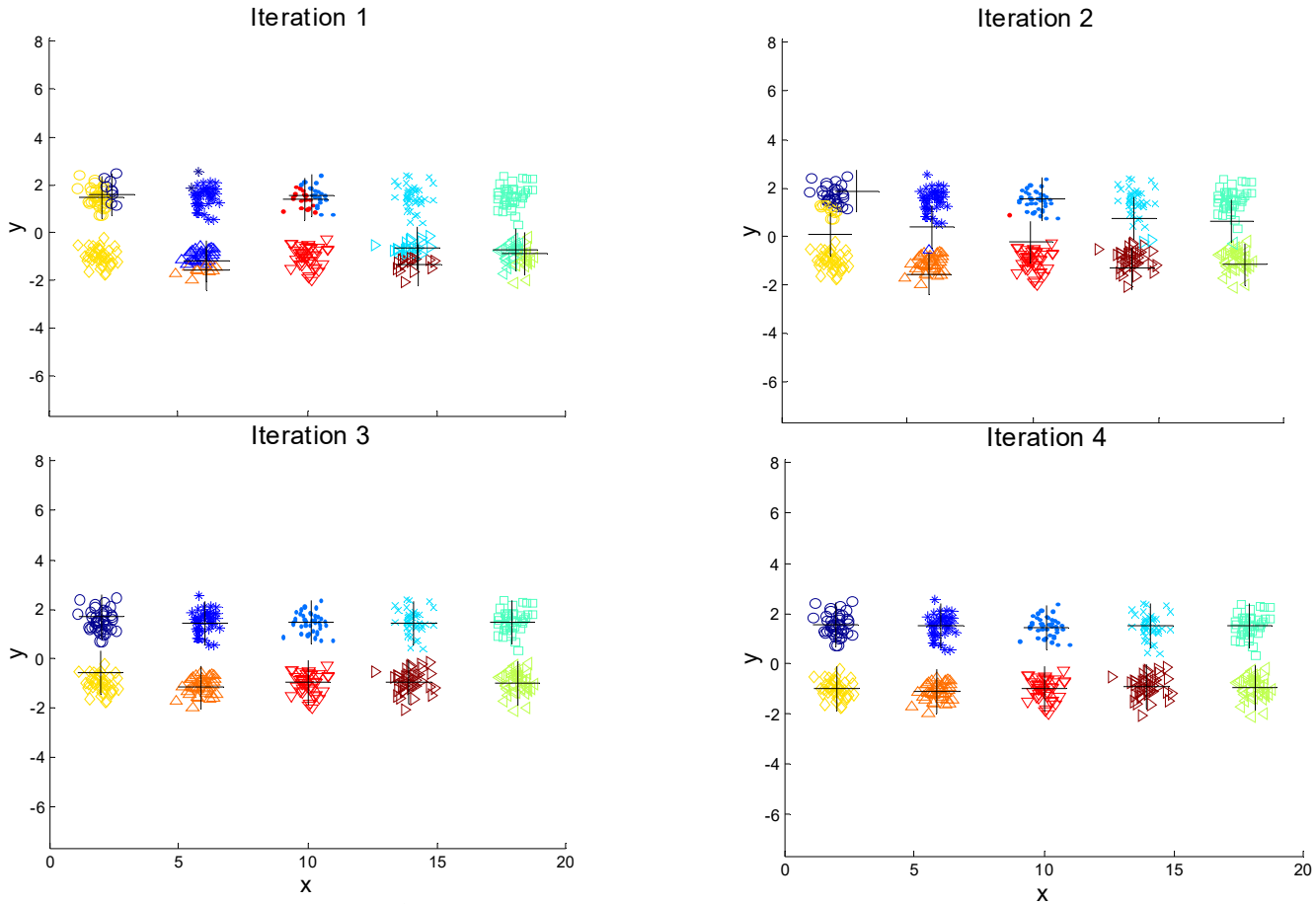
- Reiniciar varias veces con diferentes semillas al azar
  - eligieron la agrupación resultante con el suma más pequeña de error al cuadrado (SSE)
- Plazo de k-medias con diferentes números de k
  - Tenga en cuenta que la ESS de k-medias con diferentes valores de k no se puede comparar entre sí!
  - Piense: ¿qué pasa para k → n (número de ejemplos)?
- X-medios
  - comenzar con pequeñas k, entonces dividir grandes grupos y ver si mejora el resultado

El ejemplo de los 10 clúster

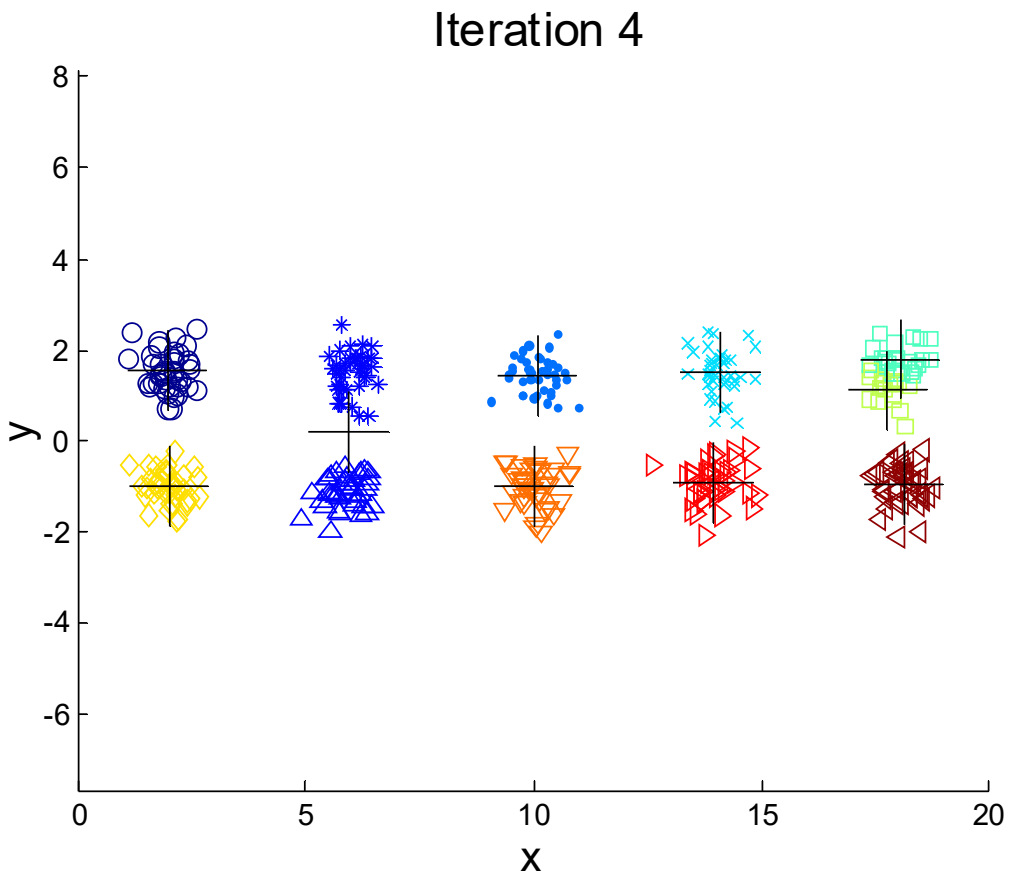


Comenzando con dos centroides iniciales en un clúster de cada par de clúster

El ejemplo de los 10 clúster



El ejemplo de los 10 clúster



Soluciones al problema de iniciación de centroides

- múltiples Corridas
  - Ayuda, pero la probabilidad no está de su lado
- Muestra y utilizar la agrupación jerárquica para determinar centroides iniciales
- Seleccione más de k centroides iniciales y luego seleccionar entre estos centroides iniciales
  - Seleccione más ampliamente separados
- Postprocesamiento
- Bisectriz K-means
  - No es tan susceptible a la inicialización

Manejo de Clúster básicos

- Básica algoritmo K-means puede producir clusters vacíos

- Varias estrategias
  - Elija el punto de que más contribuye a SSE
  - Elija un punto de la agrupación con el mayor SSE
  - Si hay varios grupos vacías, lo anterior se puede repetir varias veces.

Actualización incremental de centoides

- En el algoritmo básico K-means, centroides se actualizan después de todos los puntos se asignan a un centroide

- Una alternativa es actualizar los centroides después de cada asignación (enfoque incremental)
  - Cada asignación actualiza cero o dos centroides
  - Más costoso
  - Introduce una dependencia pedido
  - Nunca conseguir un clúster vacío
  - Puede utilizar "pesos" para cambiar el impacto

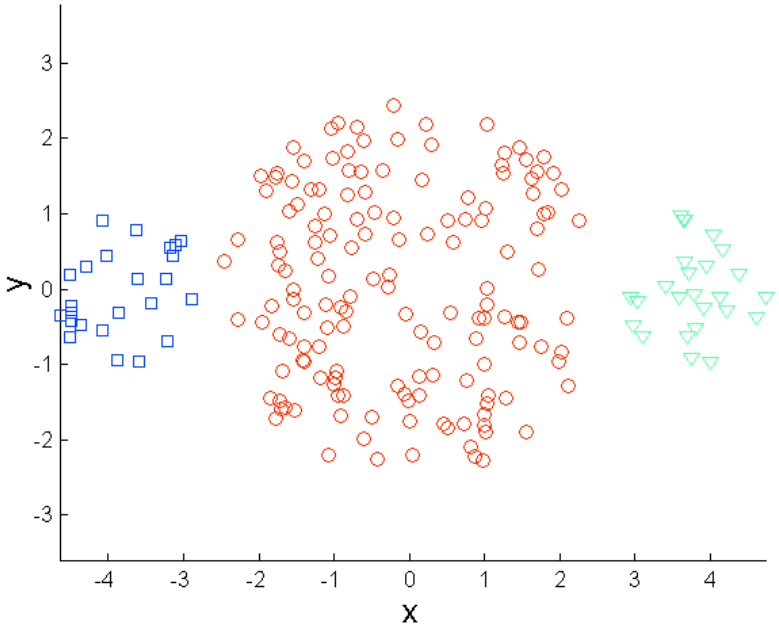
Pre-Procesamiento y Post- Procesamiento

- Pre-procesamiento
  - Normalizar los datos
  - Elimine los valores atípicos
- Postprocesamiento
  - Elimina pequeños grupos que pueden representar los valores atípicos
  - Racimos de Split 'sueltos', es decir, grupos con relativamente alta SSE
  - Combinar grupos que están "cerca" y que tienen relativamente baja SSE
  - Puede utilizar estos pasos durante el proceso de agrupación
    - ISODATA

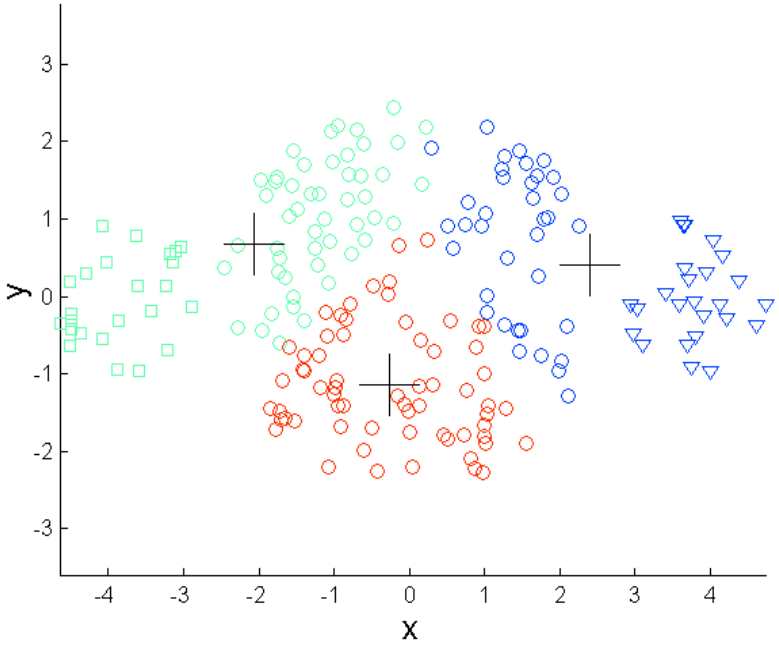
Limitaciones de K- Means

- K-medios tiene problemas cuando racimos son de diferente
  - Tamaños
  - Densidades
  - Formas para no globulares
- Kmeans tiene problemas cuando los datos contienen valores extremos.

Limitaciones de K-Menas : Diferentes tamaños

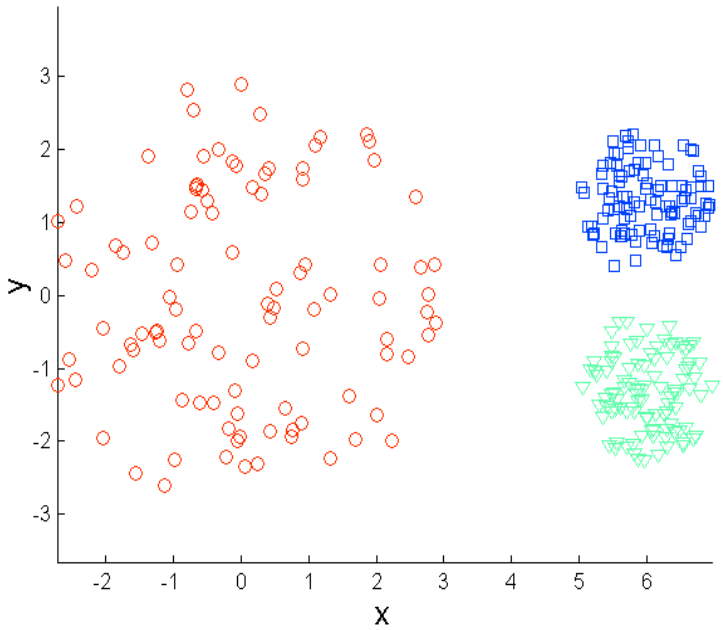


Records Originales

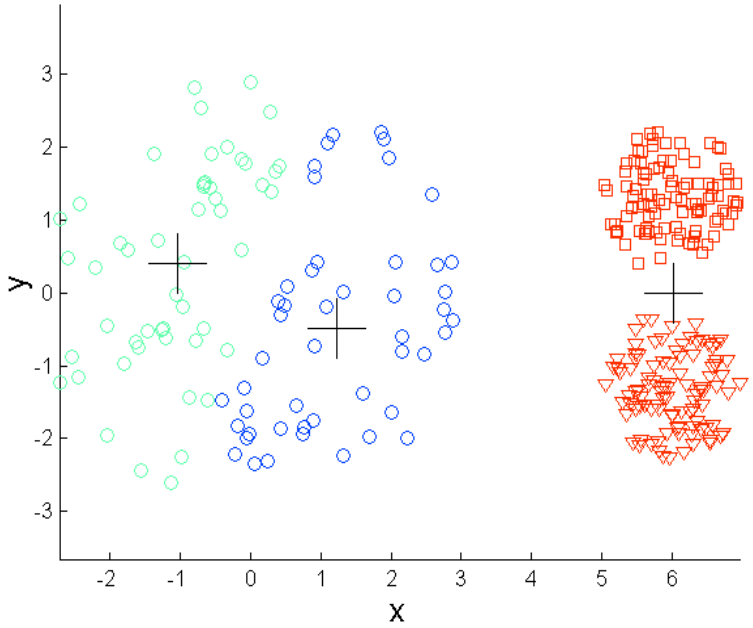


K-means (3 Clusters)

Limitaciones de K-Menas : Diferentes Densidades

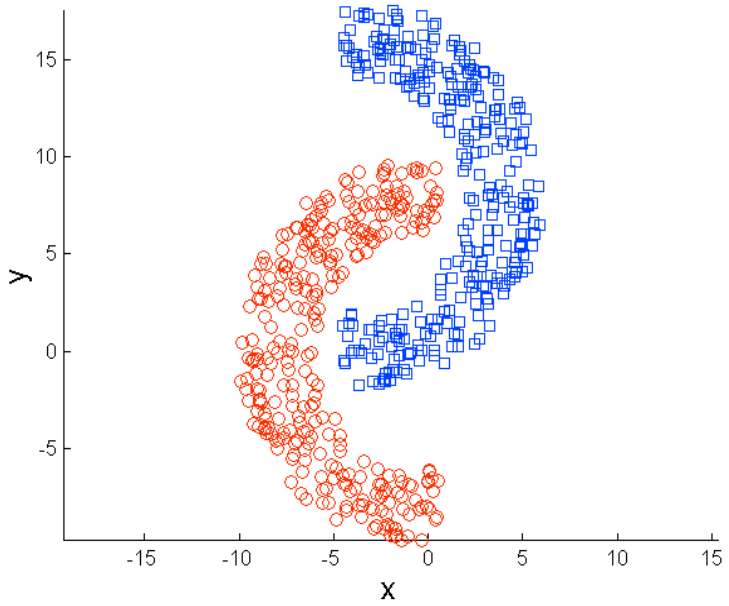


Records Originales

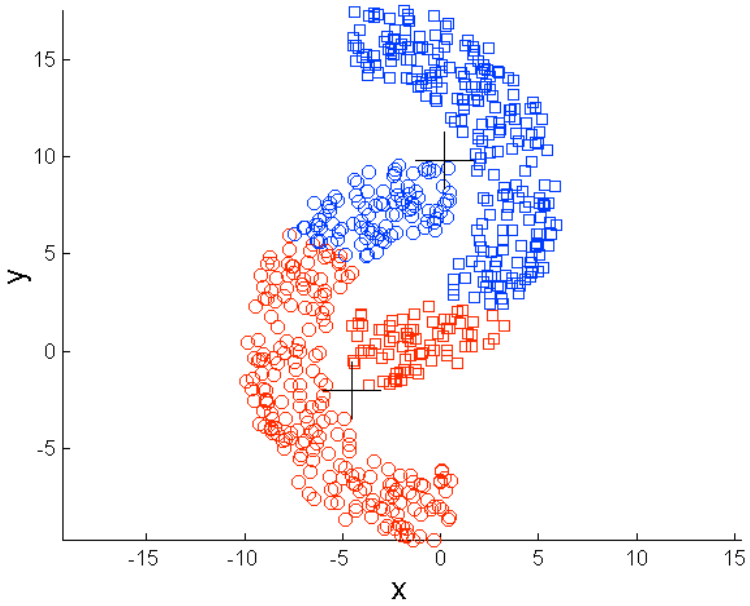


K-means (3 Clusters)

Limitaciones de K-Menas : Formas no globulares

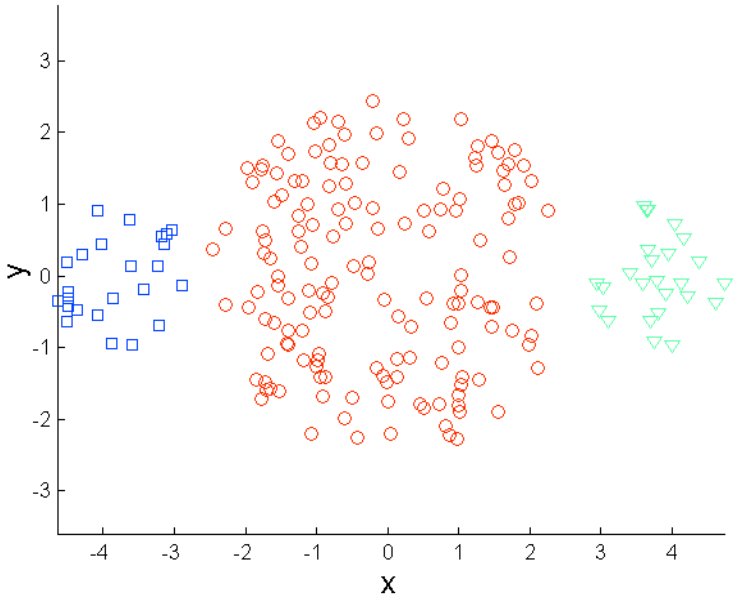


Records Originales

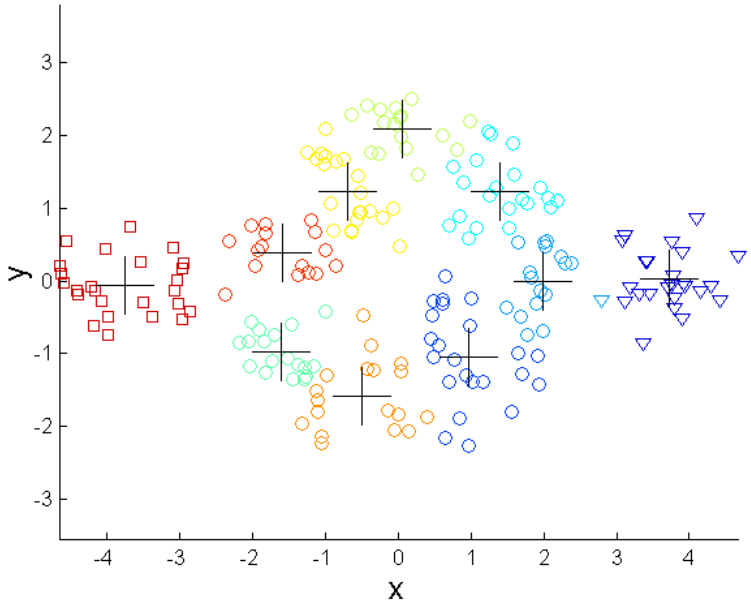


K-means (3 Clusters)

Una posible solución

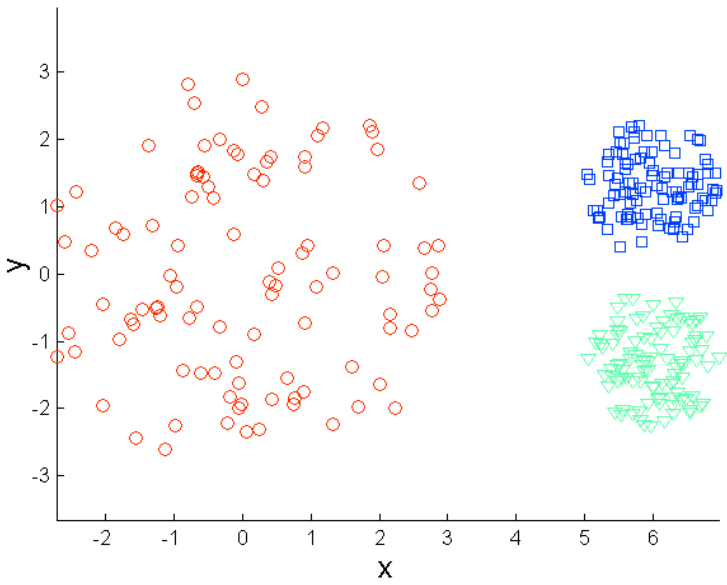


Records Originales

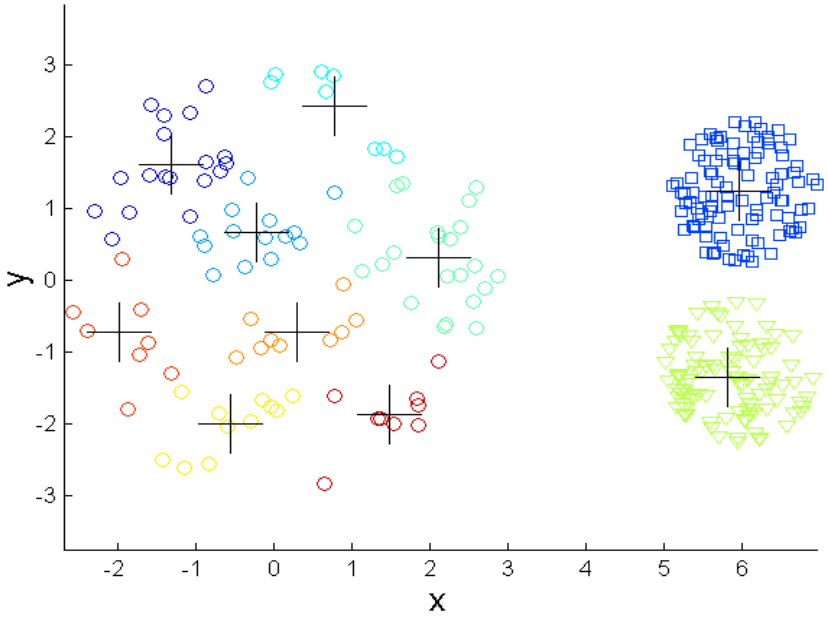


K-means Clusters

Una posible solución

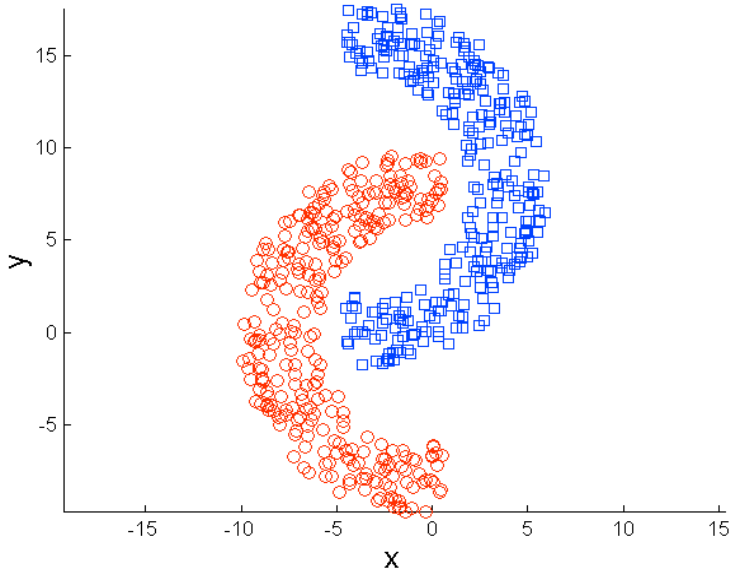


Records Originales

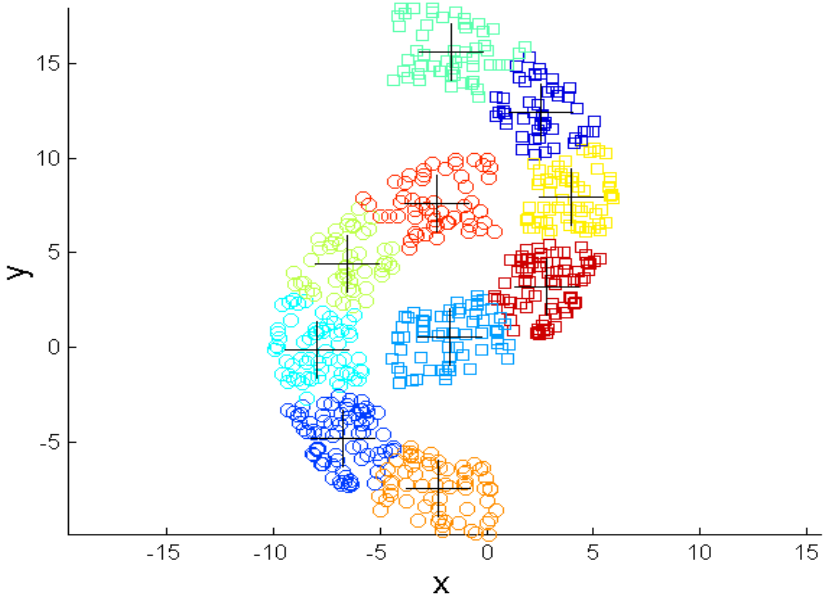


K-means Clusters

Una posible solución



Records Originales

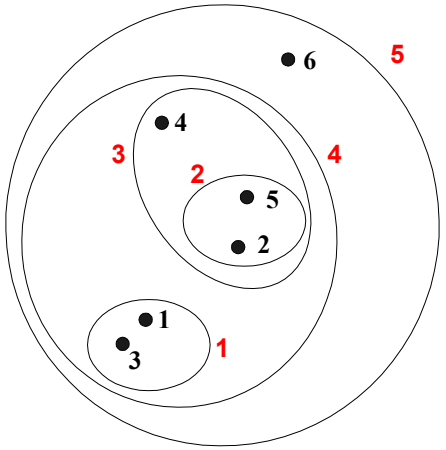


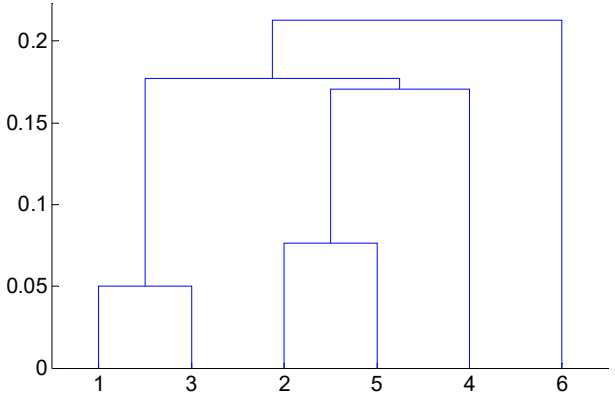
K-means Clusters

Clusters Jerárquicos

Clúster jerarquicos

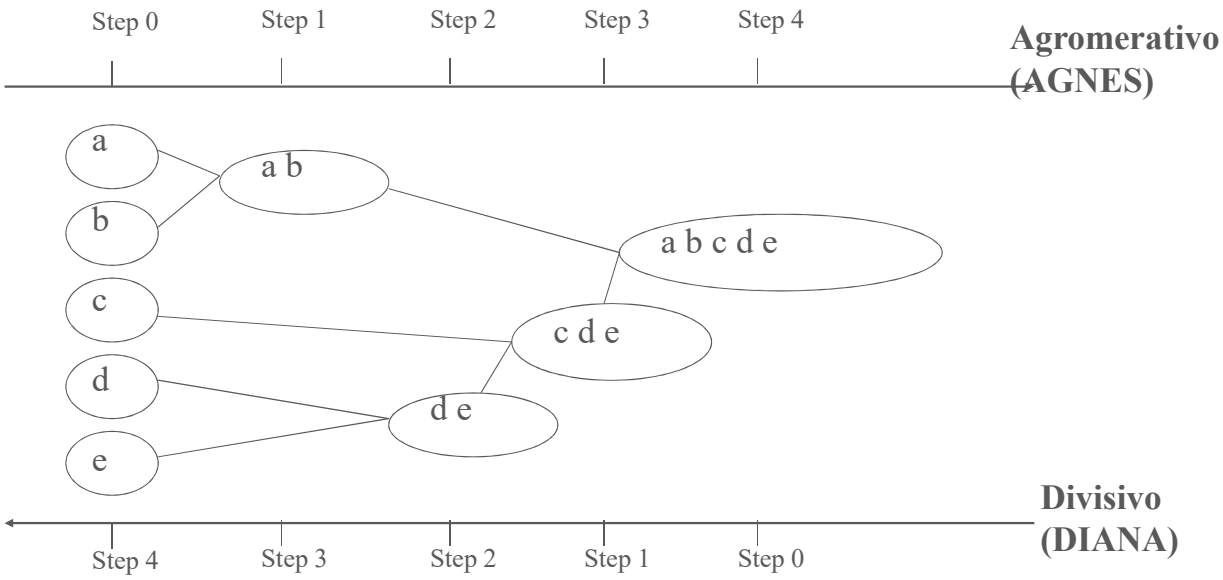
- Produce un conjunto de grupos anidados organizados como un árbol jerárquico
- Puede ser visualizado como un dendrograma
  - Un árbol como diagrama que registra las secuencias de fusiones o escisiones





Clúster jerarquicos

- Utilice matriz de distancia como criterio de agrupamiento. Este método no requiere el número de grupos k como una entrada, pero necesita una condición de terminación



Clases Clúster jerárquicos: Divisivo - DIANA (Divisive Analysis)

- Top Down (enfoque divisivo / aglomerativo)
- • Comience con un racimo universales
- • Encuentra dos grupos
- • Proceder de forma recursiva en cada subgrupo
- • Puede ser muy rápido

Clases Clúster jerárquicos: Aglomerativo AGNES (Agglomerative Nesting)

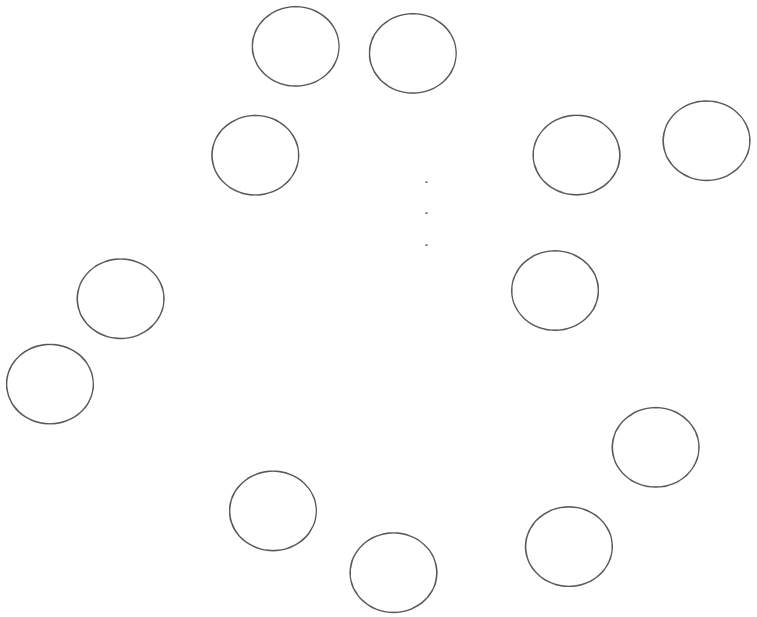
- Botton UP (aglomeración)
- Comience con racimos de instancia única
- En cada paso, unirse a los dos grupos más cercanos
- La decisión de diseño: la distancia entre clusters
  - e.g. Dos casos más cercanos en clústeres
    - frente a la distancia entre los medios

Clases Clúster jerárquicos: Aglomerativo AGNES (Agglomerative Nesting)

Algoritmo básico:

1. Calcular la matriz de proximidad
2. A cada record asignarle un clúster
3. **Repeat**
4. fusionar dos clúster que estén mas cercanos
5. Actualizar la matriz de proximidad
6. **Until** que quede un solo clúster

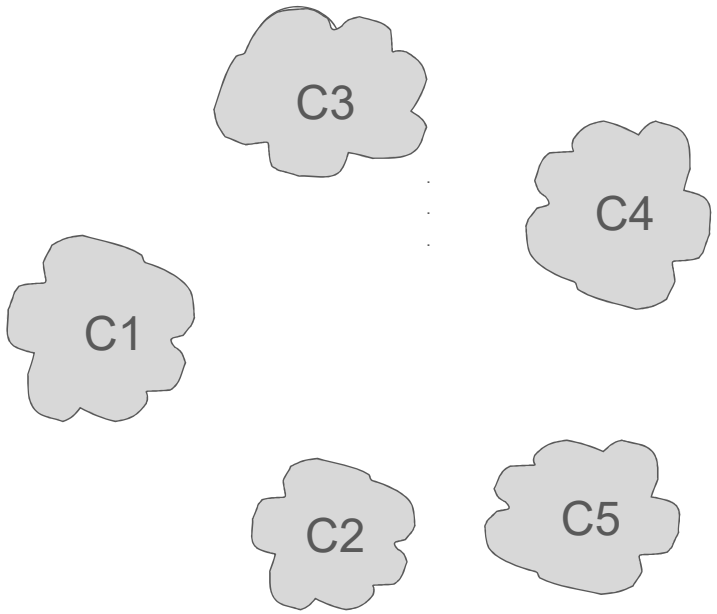
Situación inicial



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

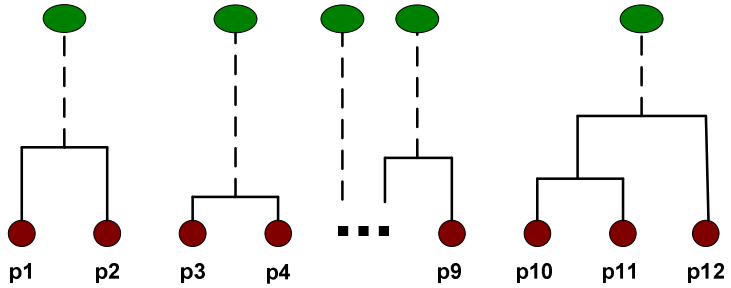
Matriz de proximidad

Situación intermedia

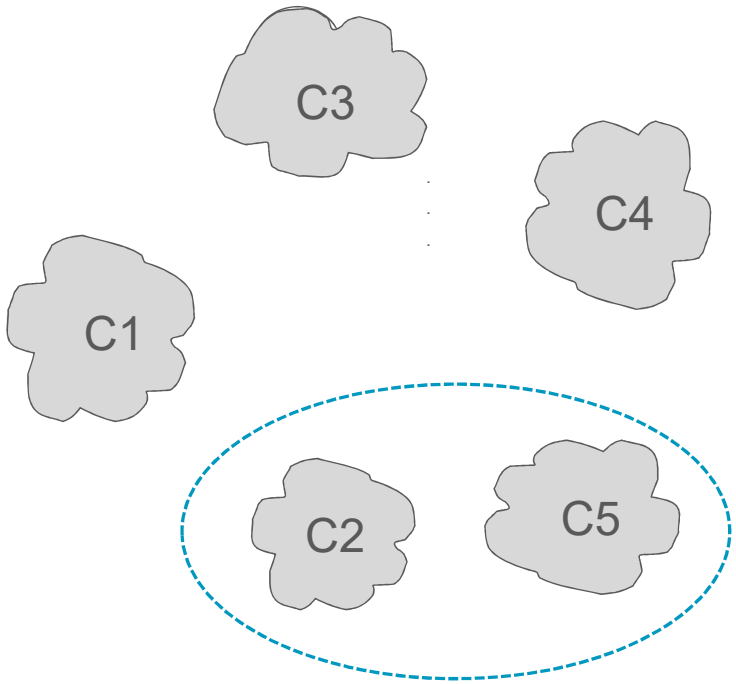


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidad

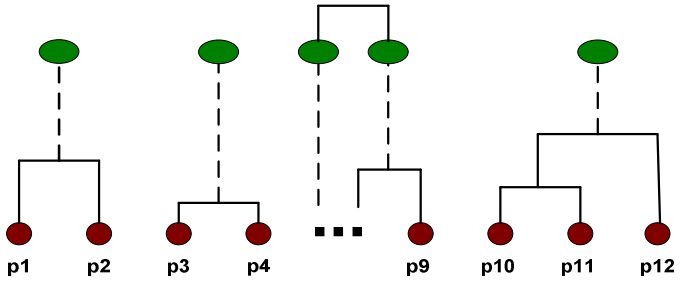


Situación intermedia

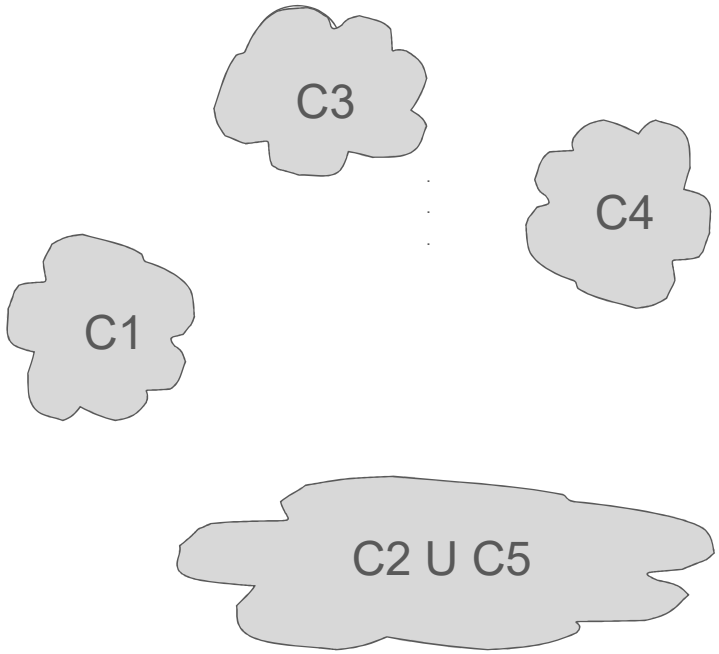


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidad

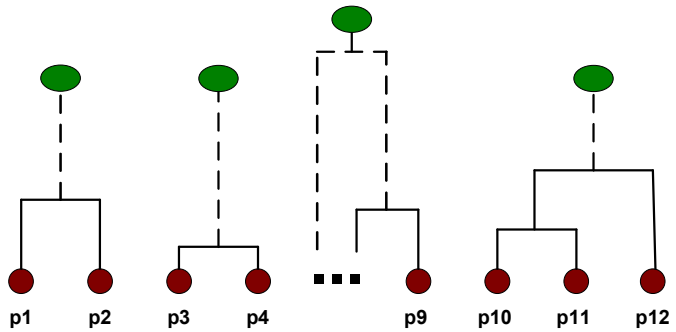


Situación intermedia

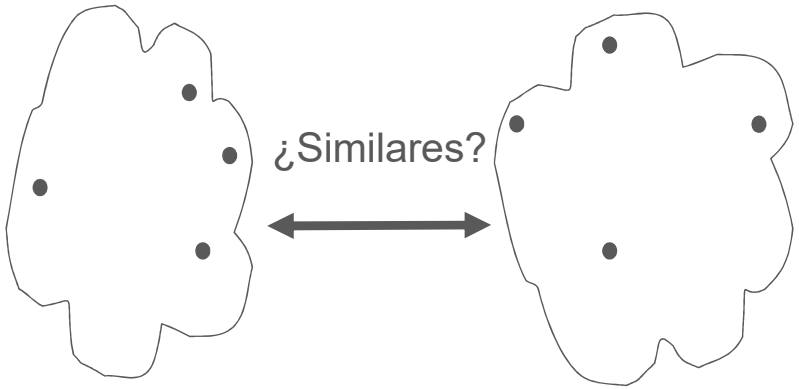


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de proximidad



Como definir la similaridad inter-Clúster

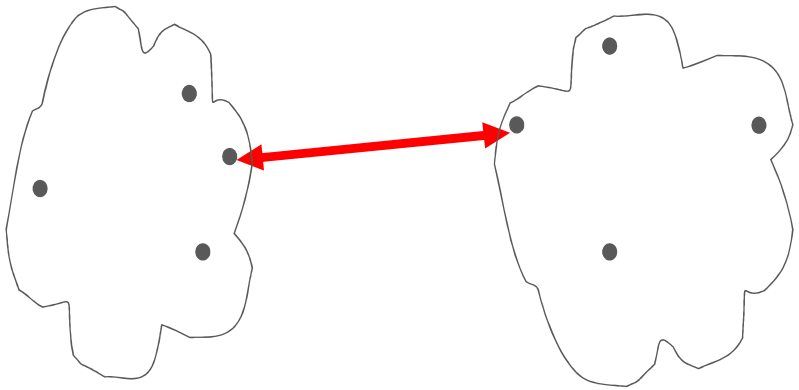


- MIN
- MAX
- Promedio de grupo
- Distancia entre centroides
- Otros métodos basados en una función objetico
  - Ward's Method : Uso de error cuadrático

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

Matriz de proximidad

Como definir la similaridad inter-Clúster

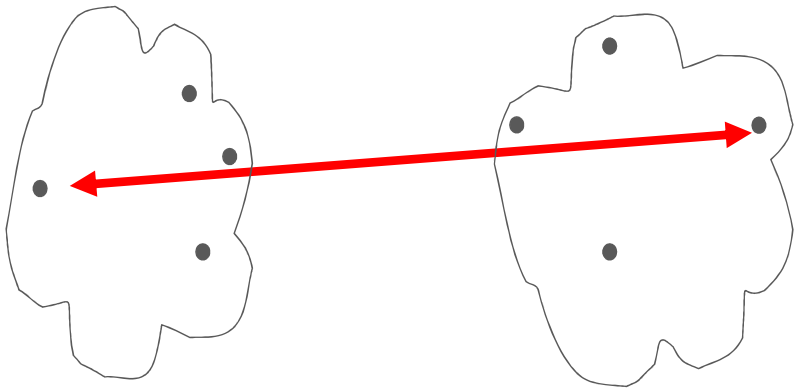


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
⋮						

Matriz de proximidad

- **MIN**
- MAX
- Promedio de grupo
- Distancia entre centroides
- Otros métodos basados en una función objetico
  - Ward's Method : Uso de error cuadrático

Como definir la similaridad inter-Clúster

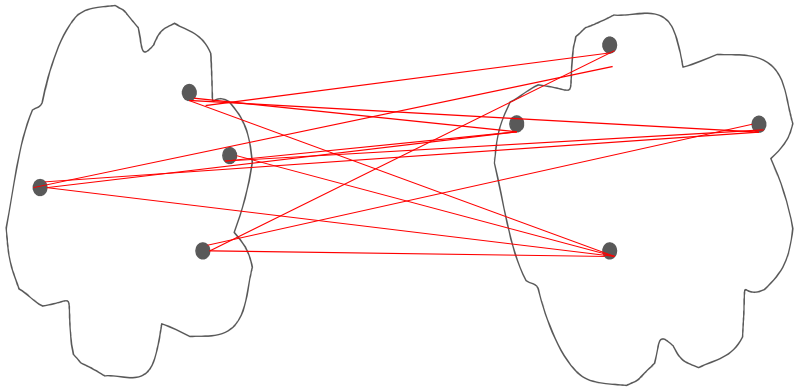


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Matriz de proximidad

- MIN
- **MAX**
- Promedio de grupo
- Distancia entre centroides
- Otros métodos basados en una función objetico
  - Ward's Method : Uso de error cuadrático

Como definir la similaridad inter-Clúster

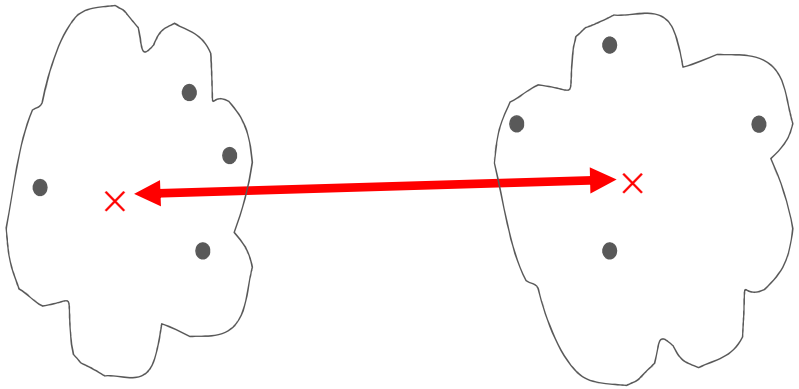


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Matriz de proximidad

- MIN
- MAX
- **Promedio de grupo**
- Distancia entre centroides
- Otros métodos basados en una función objetico
  - Ward's Method : Uso de error cuadrático

Como definir la similaridad inter-Clúster



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

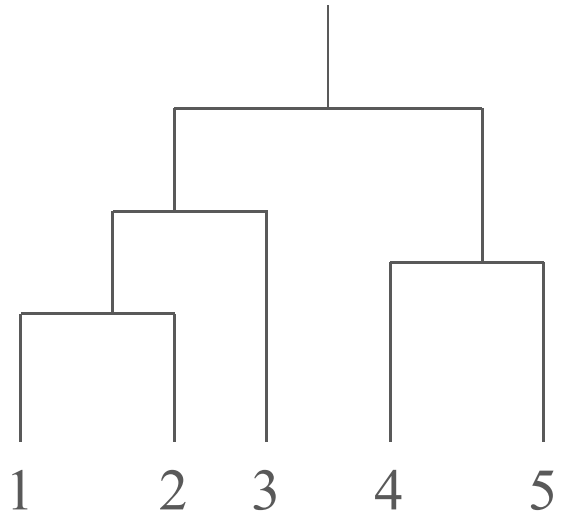
Matriz de proximidad

- MIN
- MAX
- Promedio de grupo
- **Distancia entre centroides**
- Otros métodos basados en una función objetico
  - Ward's Method : Uso de error cuadrático

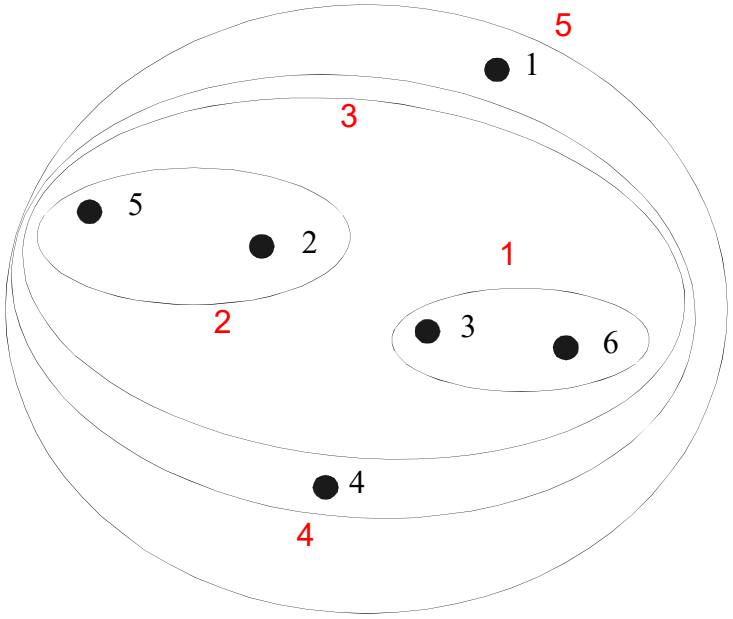
Similariedad de clúster : MIN o Link sencillo

- Similitud de los dos grupos se basa en los dos puntos más similares (más cercanos) en los diferentes grupos
  - Determinado por un par de puntos, es decir, por un enlace en el gráfico de proximidad.

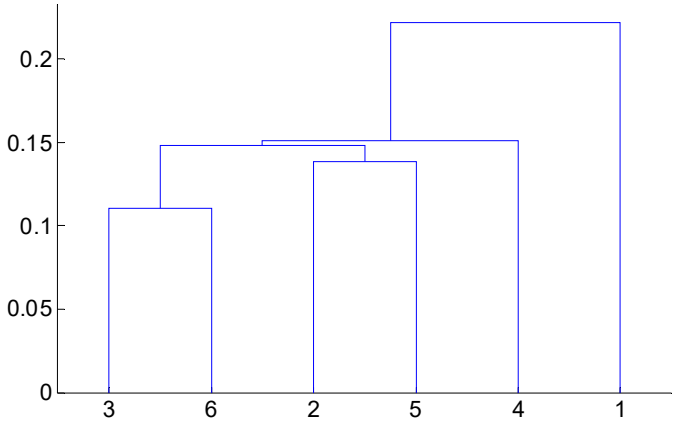
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Similariedad de clúster : MIN o Link sensillo

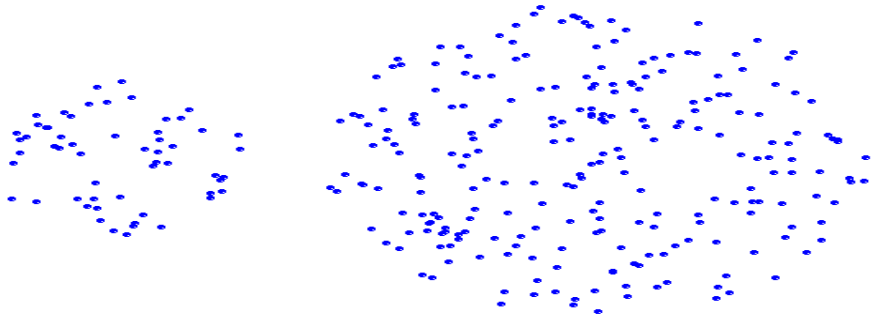


Clúster cercanos

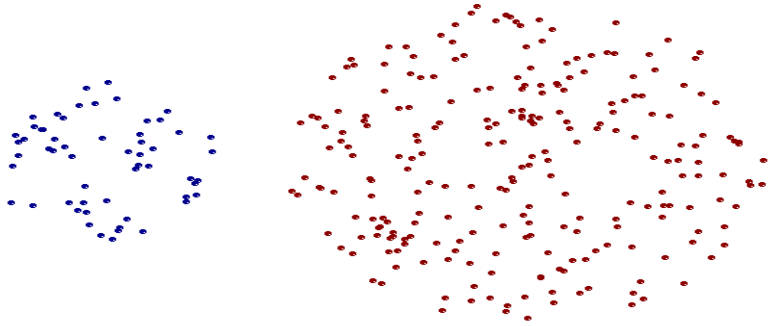


Dendrograma

Fortalezas de MIN



Record originales

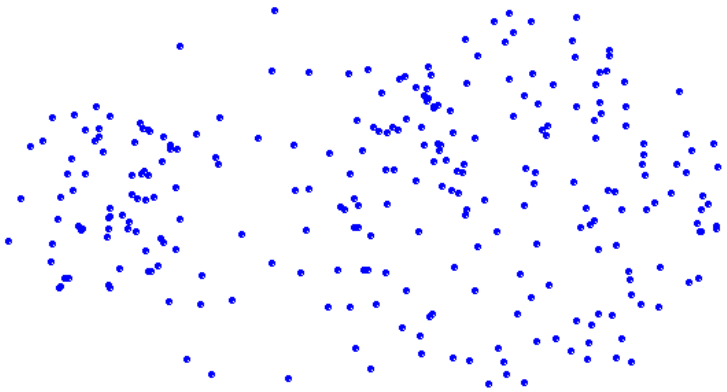


Dos clúster

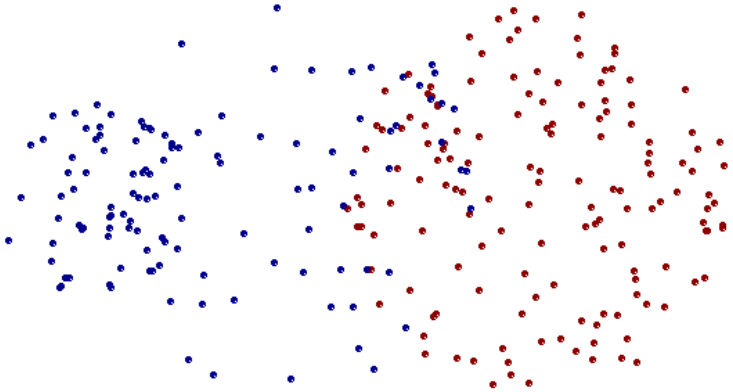
- Puede afrontar clúster de formas no elípticas

Debilidades de MIN

Record originales



Dos clúster

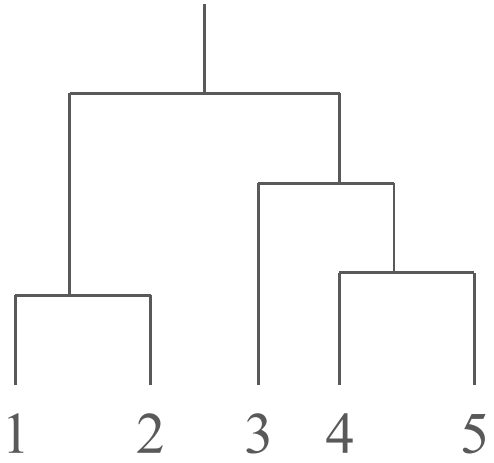


- Sensible ante valores atípicos

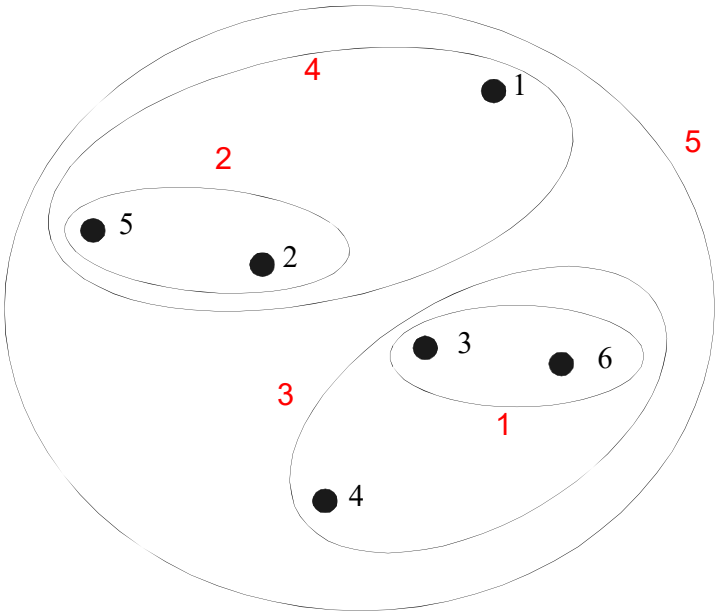
Similariedad de clúster : MAX o Link Completo

- Similitud de los dos grupos se basa en los dos puntos menos similares (más lejanos) en los diferentes grupos
  - Determinado por todos los pares de puntos en los dos grupos

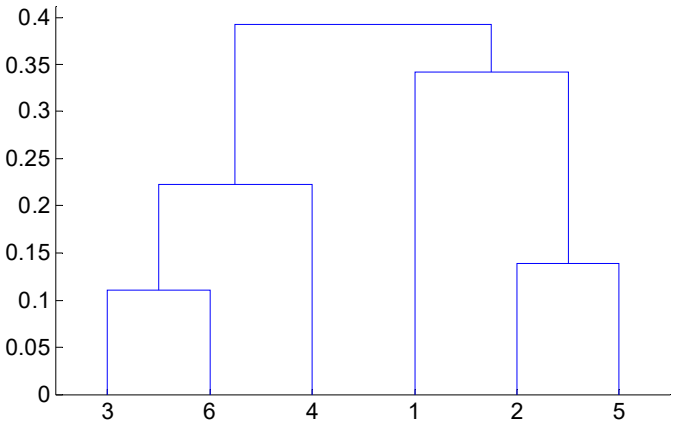
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Similariedad de clúster : MAX o Link Completo



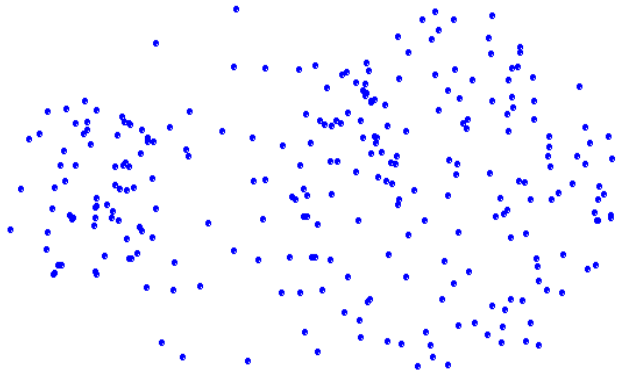
Clúster vesinos



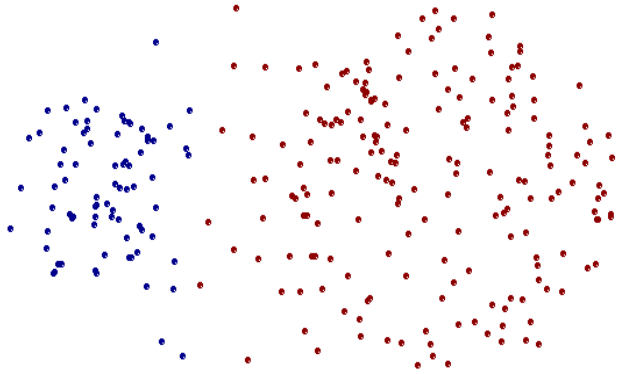
Dendrograma

Fortalezas de MAX

Record originales



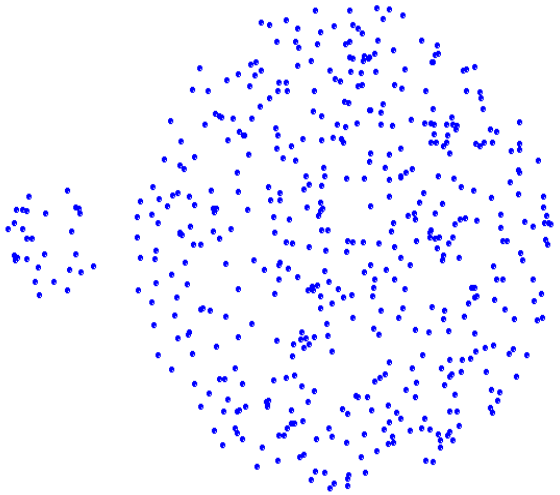
Dos clúster



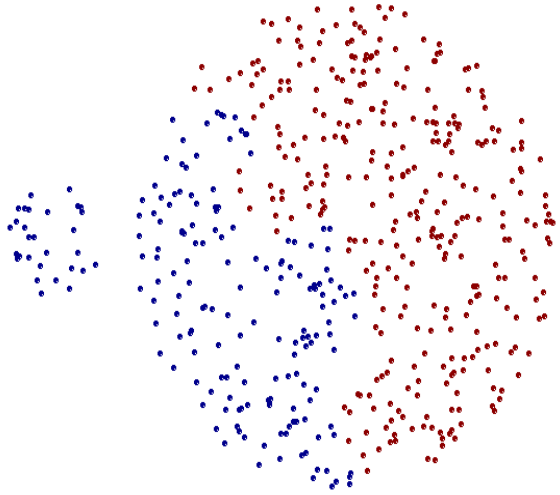
❑ Menos susceptible a valores atípico

Limitaciones de MAX

Record originales



Dos clúster



- Tiende a generar clúster amplios
- Sesgado hacia clúster globulares

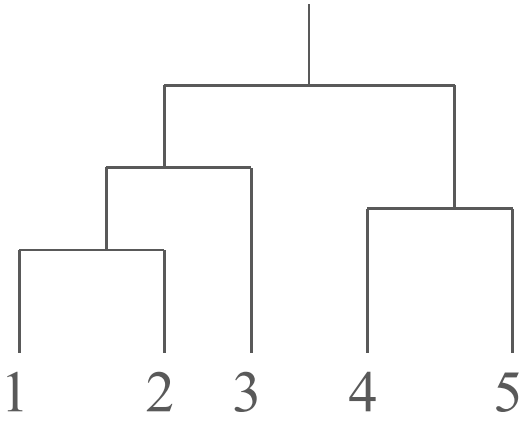
Similariedad de clúster : Promedio de grupo

- La proximidad de dos grupos es el promedio de la proximidad de pares entre los puntos en los dos grupos.

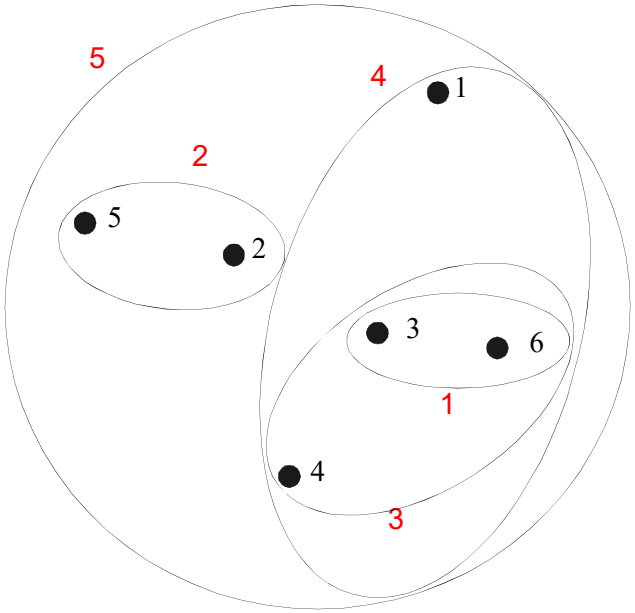
$$\text{proximidad}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Necesidad de utilizar la conectividad media de escalabilidad desde la proximidad totales favorece a las grandes agrupaciones

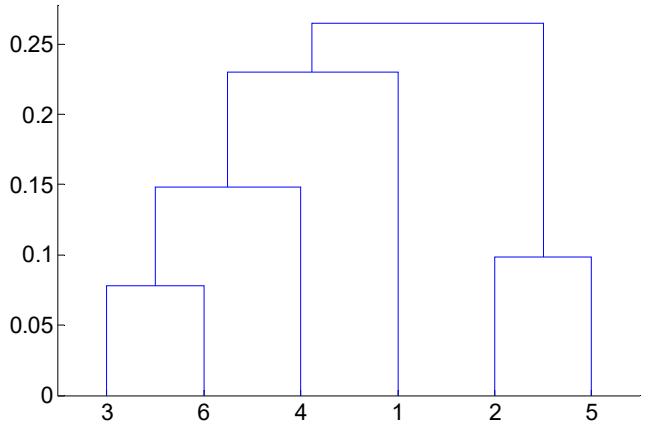
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Similariedad de clúster : Promedio de grupo



Clúster vecinos



Dendrograma

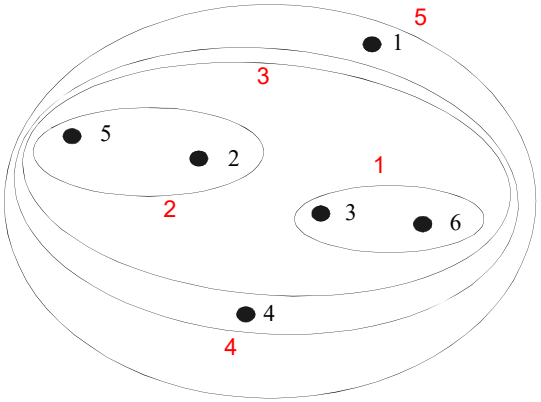
Similariedad de clúster : Promedio de grupo

- Compromiso entre individual y completa Enlace
- fortalezas
  Menos susceptible al ruido y los valores extremos
- Limitaciones
  Sesgada hacia los cúmulos globulares

Similariedad de clúster : Ward's Method

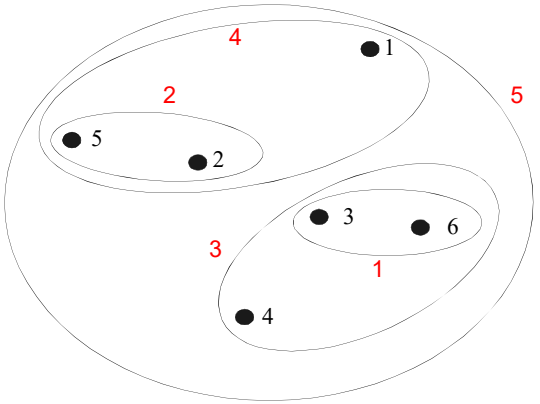
- Similitud de dos grupos se basa en el aumento de error al cuadrado, cuando se fusionan dos clusters
  - Similar al promedio del grupo si la distancia entre los puntos es la distancia al cuadrado
- Menos susceptible al ruido y los valores extremos
- Sesgada hacia los cúmulos globulares
- Análogo jerárquica de K-means
  - Puede ser utilizado para inicializar K-means

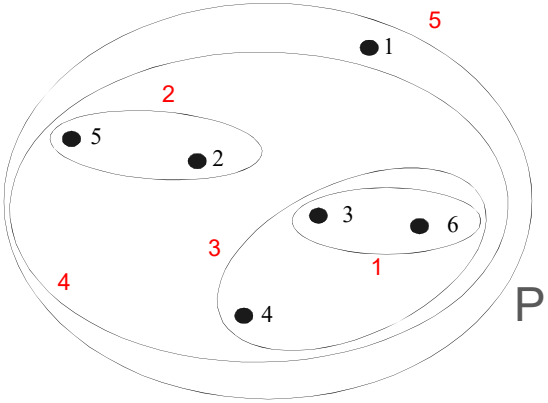
Similariedad de clúster : Comparación



MIN

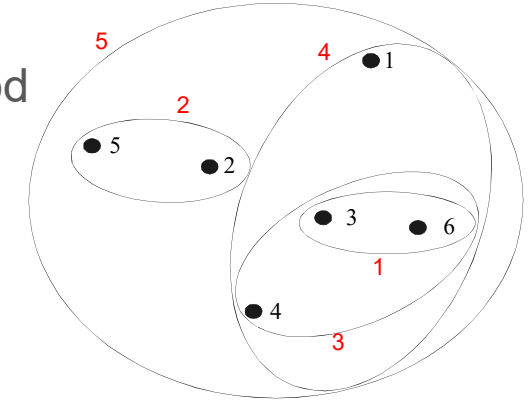
MAX





Promedio de grupo

Ward's Method



Clúster jerárquicos : Requerimientos de tiempo y espacio

- O (N2) Espacio, ya que utiliza la matriz de proximidad.
  - N es el número de puntos.

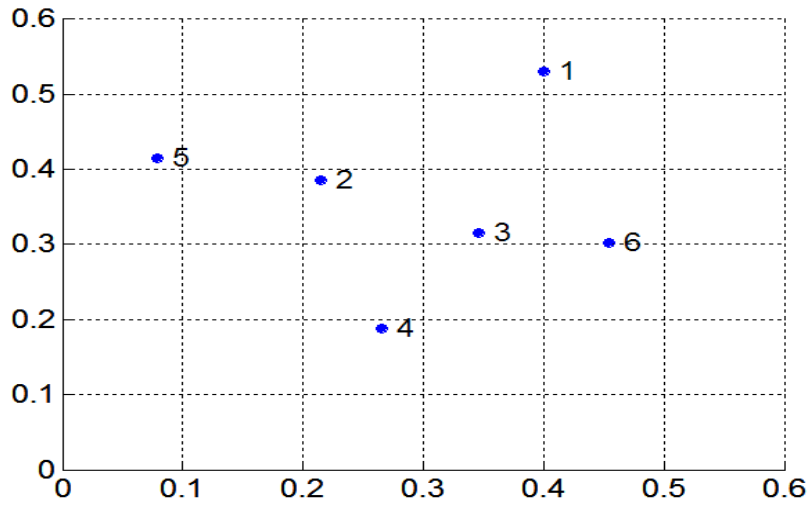
- O (N3) Tiempo. en muchos casos
  - Hay pasos N y en cada paso del tamaño, N2, matriz de proximidad deben actualizarse y buscaron
  - La complejidad puede ser reducida a O (log N2 (N)) tiempo para algunos enfoques

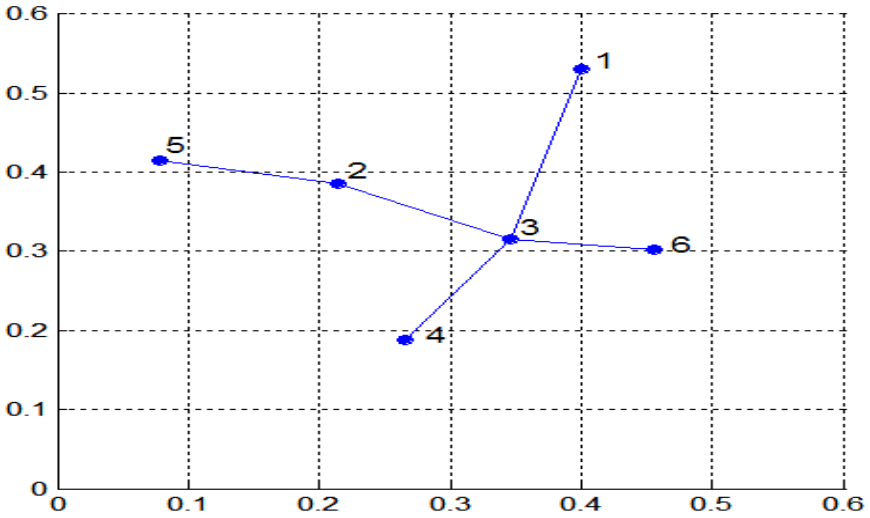
Clúster jerárquicos : Problemas y limitaciones

- Una vez que se tomó la decisión de combinar los dos grupos, no se puede deshacer
- Sin función objetivo se minimiza directamente
- Diferentes esquemas tienen problemas con uno o más de los siguientes:
  - La sensibilidad al ruido y los valores extremos
  - El manejo de diferentes grupos de tamaño y formas convexas Dificultad
  - Romper grandes grupos

Clúster jerárquicos Divisivo

- Construir MST (Minimum Spanning Tree)
  - Comience con un árbol que consiste en cualquier punto
  - En pasos sucesivos, busque el par más cercano de los puntos (p, q) de tal manera que un punto (P) está en el árbol actual, pero el otro (q) no se
  - Añadir q al árbol y poner un borde entre p y q





Clúster jerárquicos Divisivo

Algoritmo básico:

Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

DBSCAN

DBSCAN

DBSCAN es un algoritmo basado en la densidad.
Descubre clúster de forma arbitraria en bases de datos espaciales con el ruido

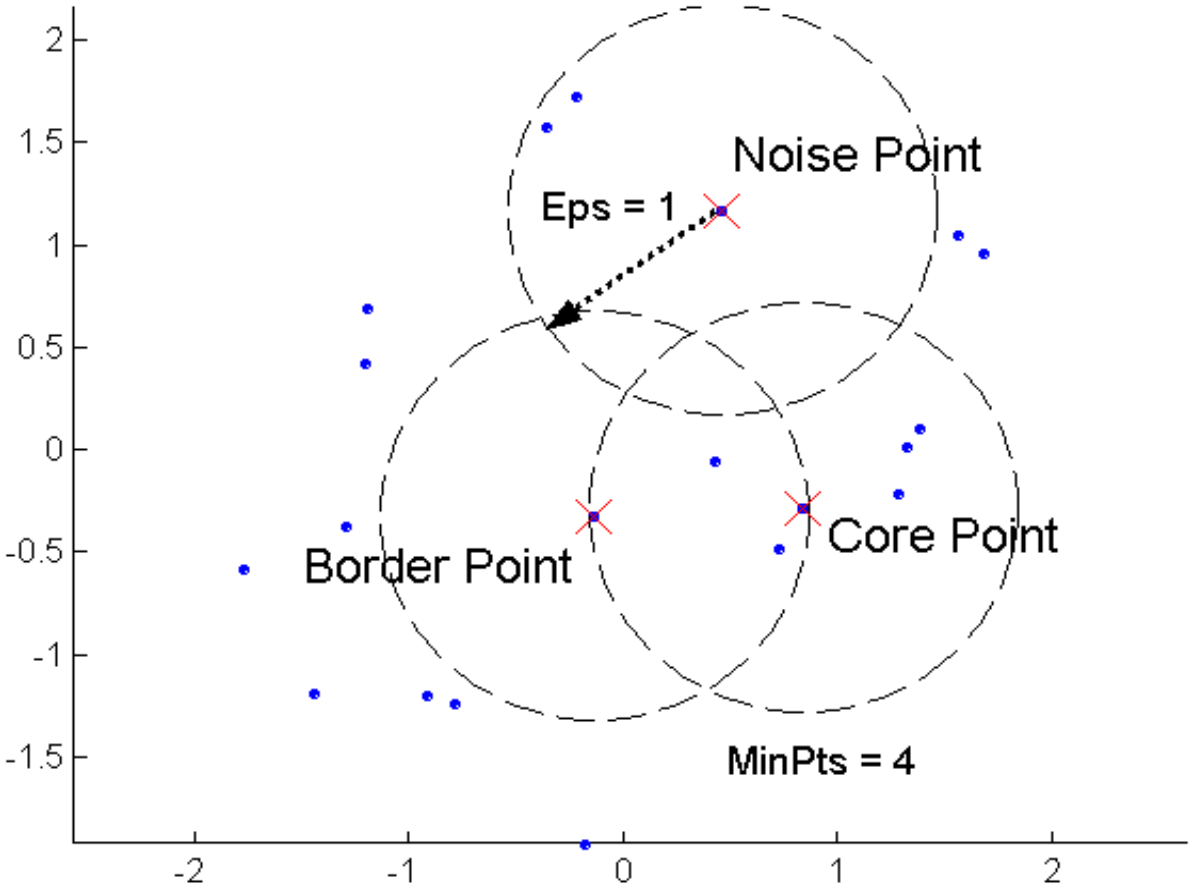
Densidad = número de puntos dentro de un radio específico (EPS)

Un punto es un punto central si tiene más de un número determinado de puntos (MinPts) dentro de Eps
Estos son los puntos que están en el interior de un clúster

Un punto fronterizo tiene menos de MinPts dentro de Eps, pero está en el entorno de un punto básico

Un punto de ruido es cualquier punto que no es un punto central o un punto fronterizo.

DBSCAN


Noise Point
Eps = 1
Border Point
Core Point
MinPts = 4
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
0
1
2

DBSCAN

Elimina los puntos atípicos (outiers)

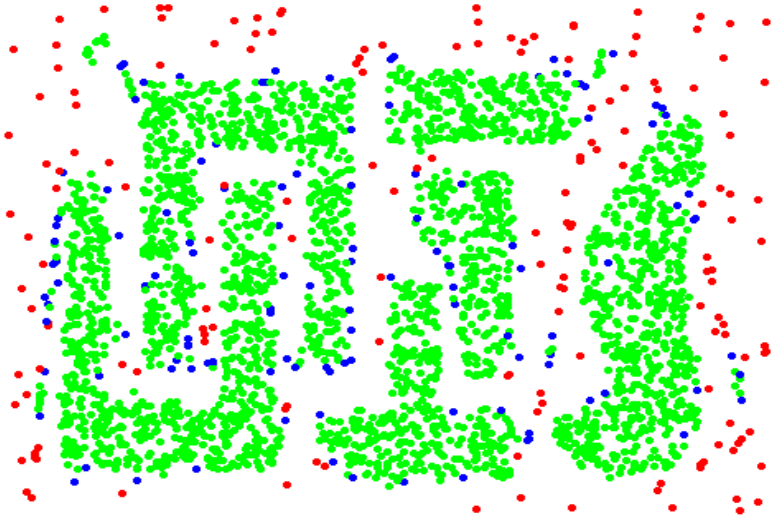
Realizar la agrupación de los puntos restantes

current_cluster_label $\leftarrow 1$
for all core points **do**
  if the core point has no cluster label **then**
    current_cluster_label $\leftarrow$ *current_cluster_label* $+ 1$
    Label the current core point with cluster label *current_cluster_label*
  end if
  for all points in the *Eps*-neighborhood, except i^{th} the point itself **do**
    if the point does not have a cluster label **then**
      Label the point with cluster label *current_cluster_label*
    end if
  end for
end for

DBSCAN

Record Originates

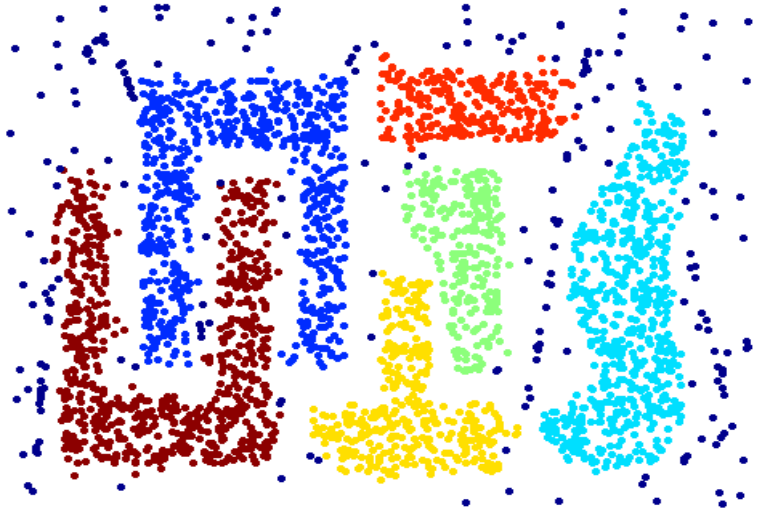
Eps = 10,
MinPts = 4



Tipos de puntos: core, border y noise

DBSCAN

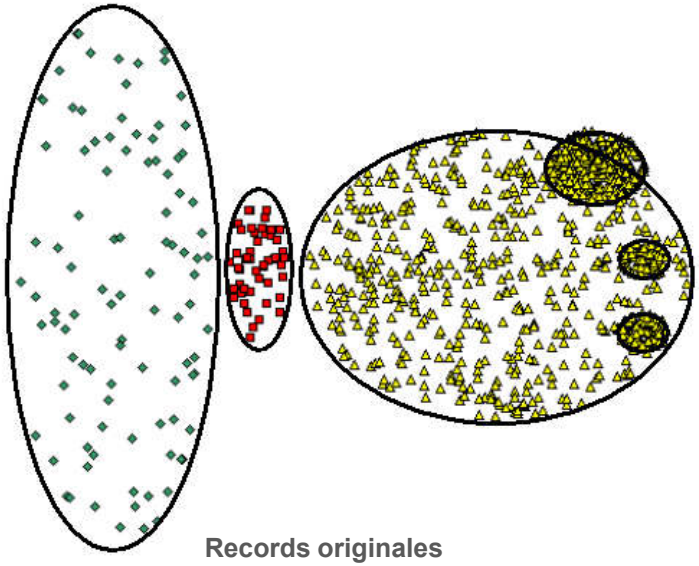
Records originaes



Clusters

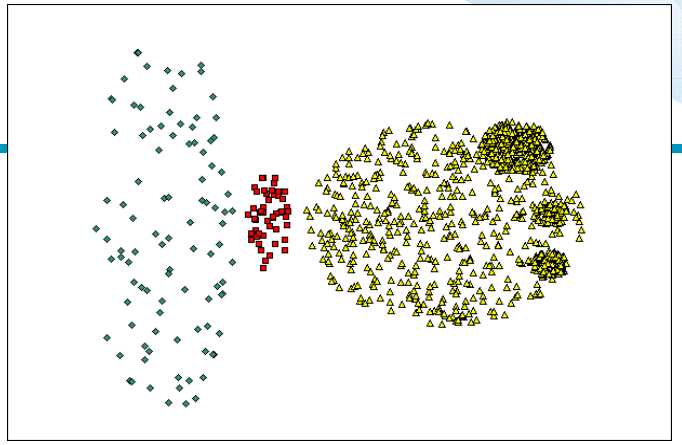
- Buenos contra el ruido
- Pueden manejar diferentes formas y tamaños de clusters

DBSCAN

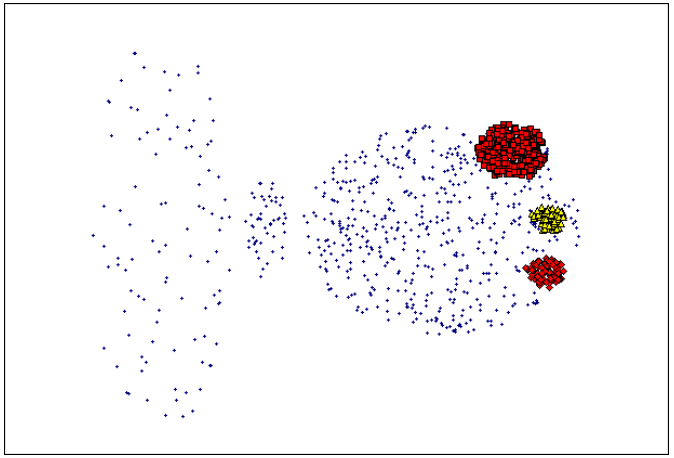


Records originales

- **Variedad de densidades**
- **Datos de alta dimensionaidad**



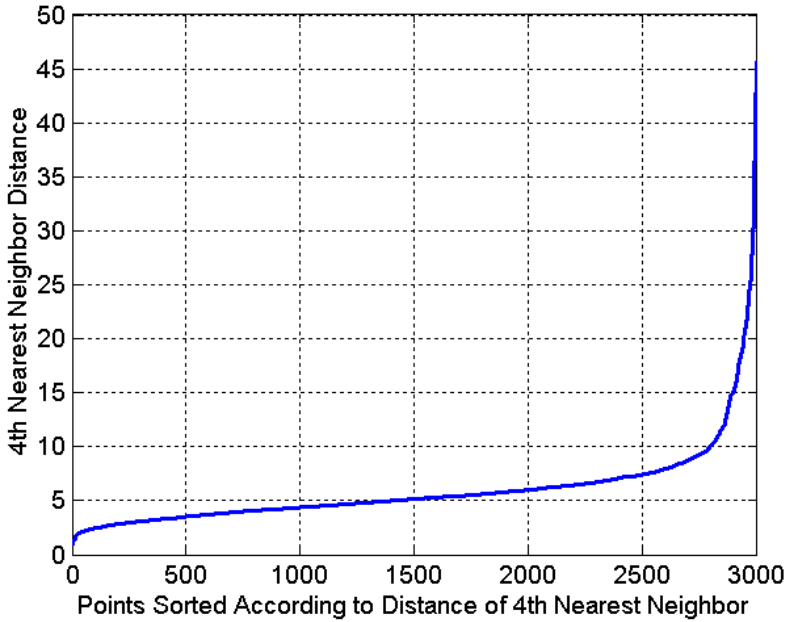
(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

DBSCAN

- La Idea es que los puntos en un clúster, Kth sus vecinos más cercanos están a aproximadamente la misma distancia
- Los puntos de ruido tienen la orden k vecino más cercano en la distancia más lejos
- Así, la trama ordenada según la distancia de todos los puntos de su orden k vecino más cercano



Evaluación de clusters

Validación de clusters

- Para la clasificación supervisada tenemos una variedad de medidas para evaluar lo bien que nuestro modelo es
- Exactitud, precisión, recordar

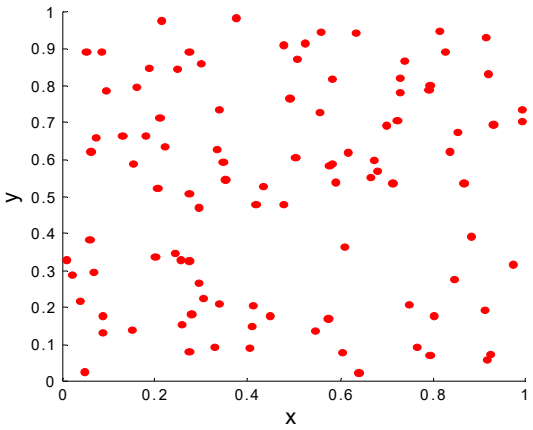
- Para el análisis de conglomerados, la cuestión análoga es la forma de evaluar la "bondad" de las agrupaciones resultantes?

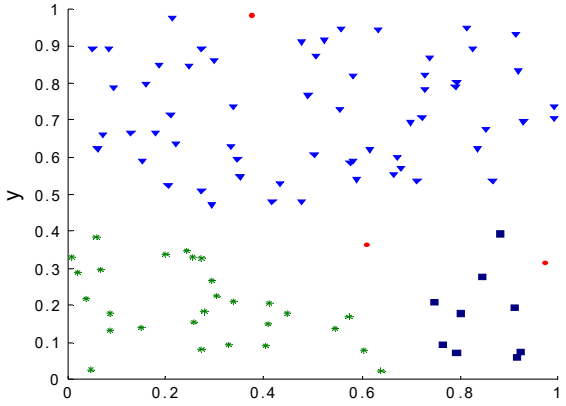
- Sin embargo, "racimos están en el ojo del espectador"!

- Entonces ¿por qué queremos para evaluarlos?
- Para evitar la búsqueda de patrones de ruido
- Para comparar los algoritmos de agrupamiento
- Para comparar dos conjuntos de agrupaciones
- Para comparar los dos grupos

Validación de clusters

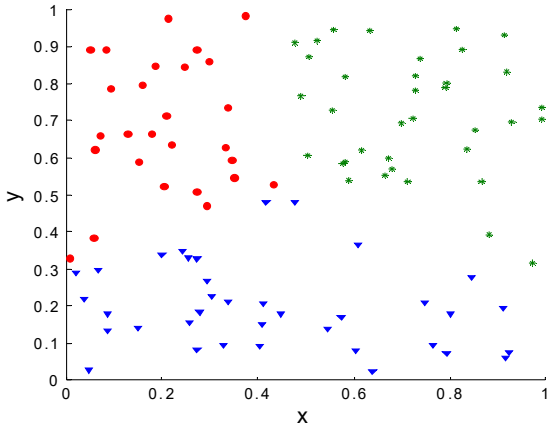
Pentos aleatorios

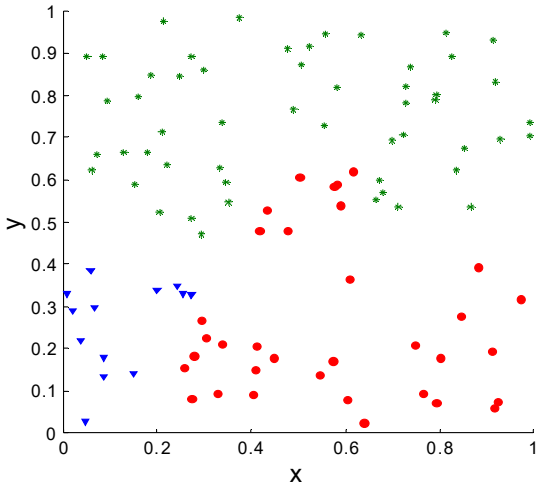




DBSCAN

K-means





Complete Link

Validación de clusters

- La determinación de la tendencia agrupación de un conjunto de datos, es decir, distinguir si la estructura no aleatoria realmente existe en los datos.
- La comparación de los resultados de un análisis de conglomerados para conocida externamente resultados, etiquetas de clase por ejemplo, para dado externamente.
- La evaluación de qué tan bien los resultados de un análisis de conglomerados se ajustan a los datos sin referencia a información externa. - Utilice sólo los datos
- Comparando los resultados de dos conjuntos diferentes de análisis de conglomerados para determinar cuál es mejor.
- La determinación del número "correcto" de las agrupaciones. Para 2, 3, y 4, podemos distinguir, además, si queremos evaluar toda la agrupación o sólo grupos individuales.

Mediciones de Validación de clusters

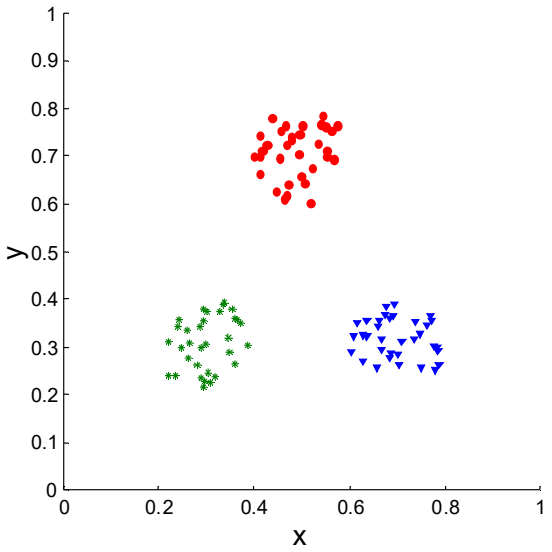
- medidas numéricas que se aplican para juzgar diversos aspectos de la validez de clúster, se clasifican en los siguientes tres tipos.
- Índice externa: Se utiliza para medir el grado en que las etiquetas de racimo que coincida con una alimentación externa etiquetas de clase.
- entropía
- Índice interna: Se utiliza para medir la bondad de una estructura de agrupación sin respeto a la información externa.
- Suma de errores al cuadrado (SSE)
- Índice relativa: Se utiliza para comparar dos agrupamiento diferente de las agrupaciones.
- A menudo un índice externo o interno se utiliza para esta función, por ejemplo, SSE o entropía
- A veces, estos se denominan como criterios en lugar de índices
- Sin embargo, a veces criterio es la estrategia general y el índice es la medida numérica que implementa el criterio.

Mediciones de Validación de clusters via correlación

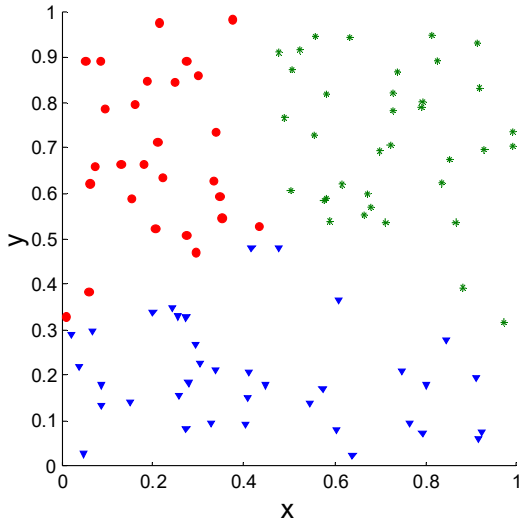
dos matrices
Matriz de proximidad
"Matriz de incidencia
Una fila y una columna para cada punto de datos
Una entrada es 1 si el par asociado de puntos pertenecen al mismo grupo
Una entrada es 0 si el par asociado de puntos pertenece a diferentes grupos de
Calcular la correlación entre las dos matrices
Puesto que las matrices son simétricas, sólo la correlación entre? N (n-1) / 2 entradas debe ser calculada.
Alta correlación indica que los puntos que pertenecen al mismo grupo están cerca uno del otro.
No es una buena medida para algunos densidad de clusters basados contigüidad.

Medisiones de Validación de clusters via correlación

Correlación de la incidencia y de proximidad matrices para la K-significa la agrupación de los dos conjuntos de datos siguientes.



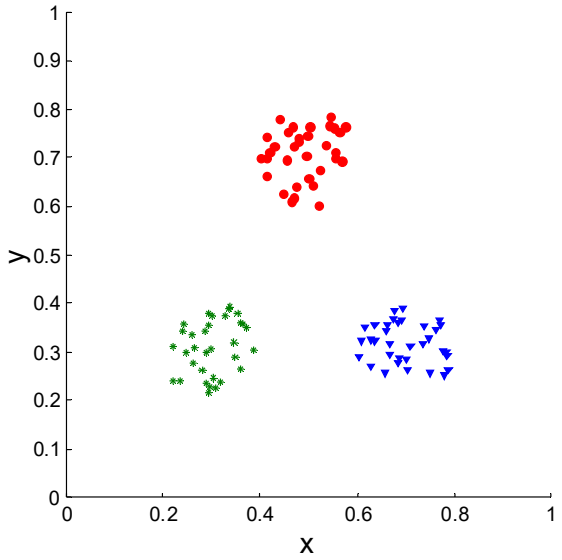
Corr = -0.9235

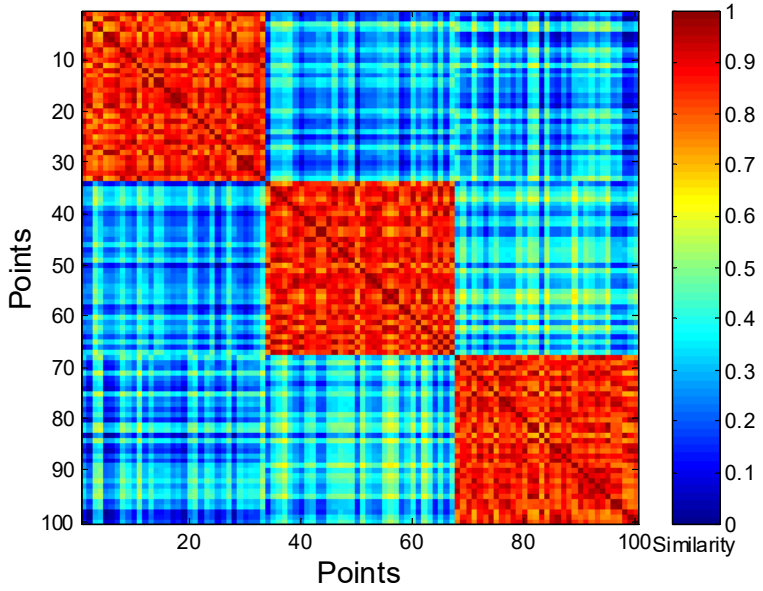


Corr = -0.5810

Medisiones de Validación de clusters via correlación

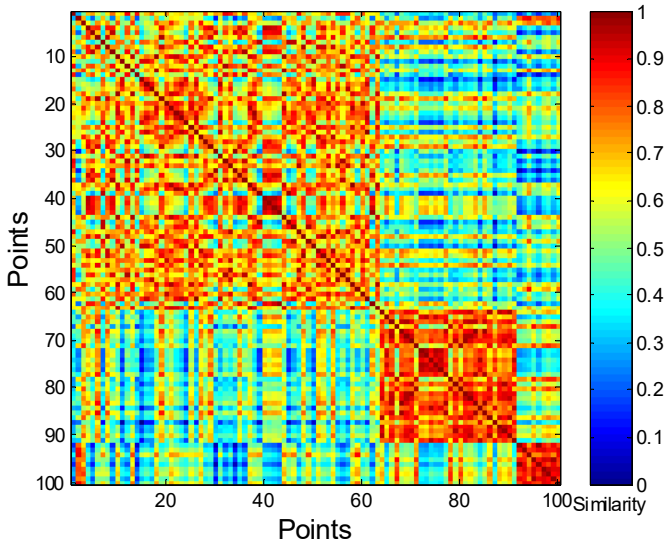
Ordenar la matriz de similitud con respecto a agruparse etiquetas e inspeccionar visualmente.

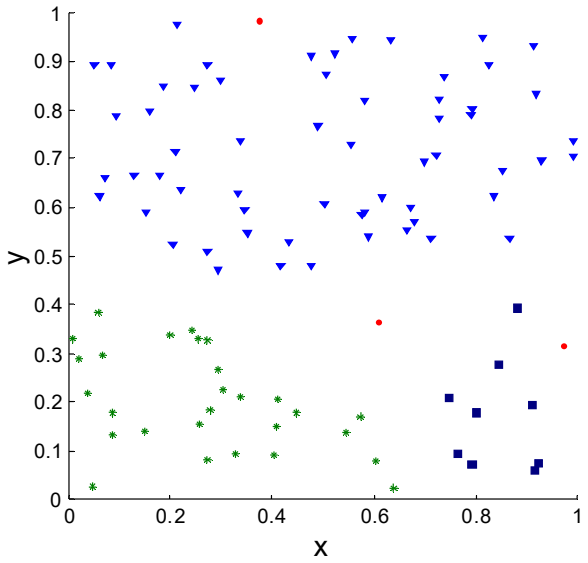




Medisiones de Validación de clusters via correlación

Clusters en los datos aleatorios no son tan nítidas

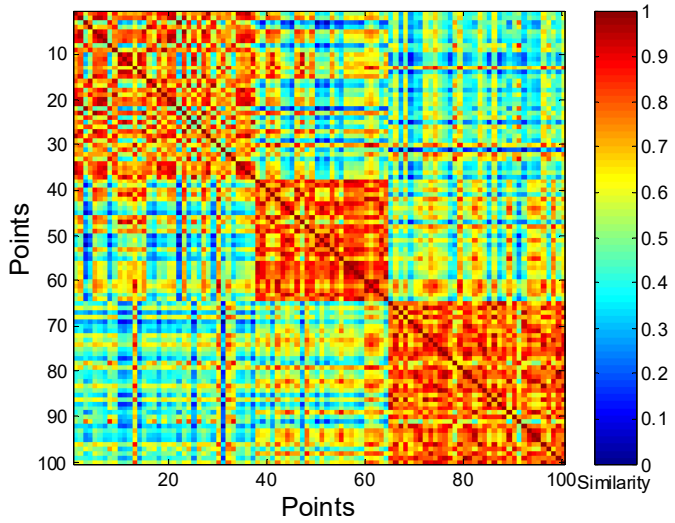


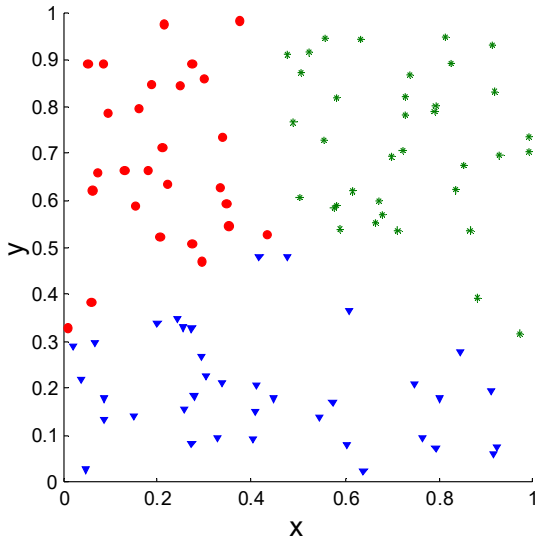


DBSCAN

Medisiones de Validación de clusters via correlación

Clusters en los datos aleatorios no son tan nítidos

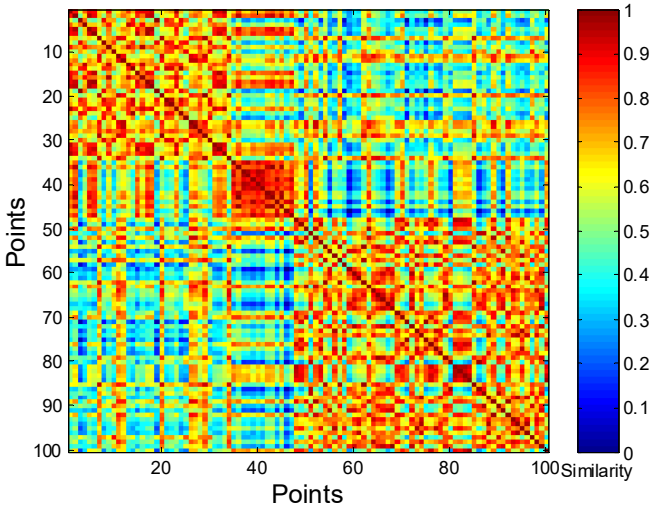


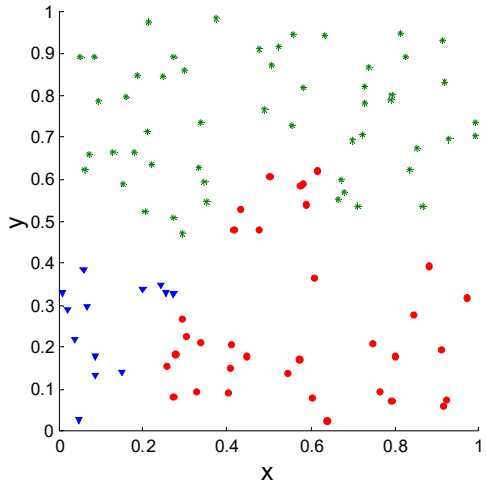


K-means

Medisiones de Validación de clusters via correlación

Clusters en los datos aleatorios no son tan nítidos

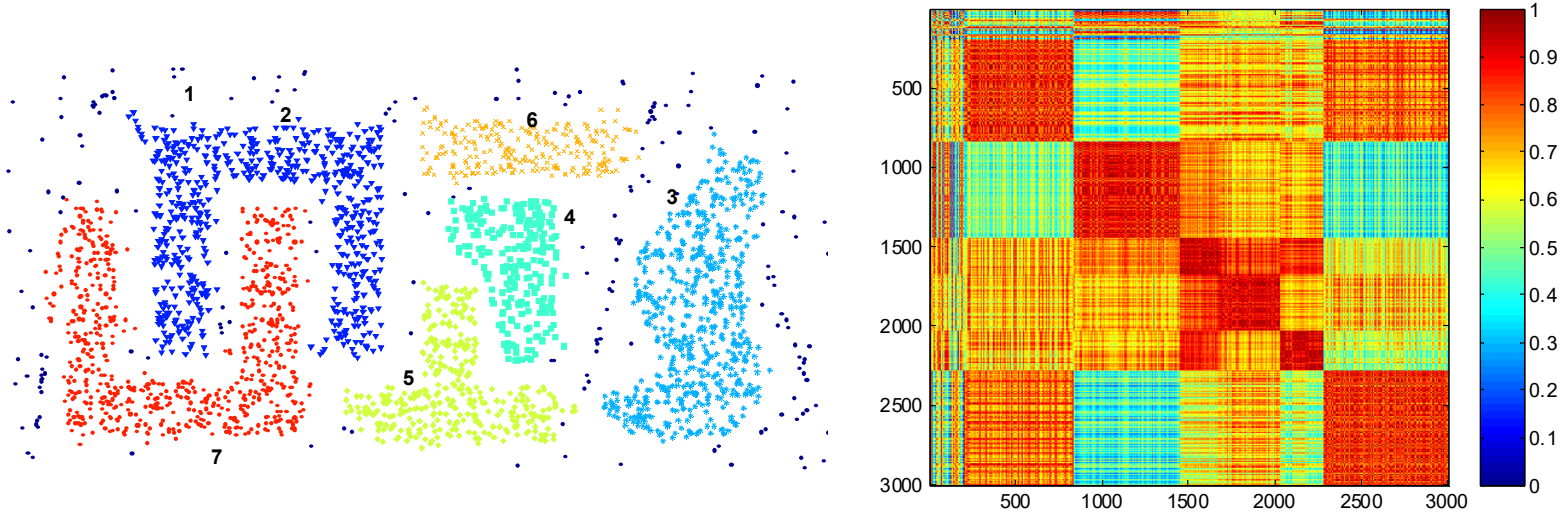




Complete Link

Medisiones de Validación de clusters via correlación

Clusters en los datos aleatorios no son tan nítidos



DBSCAN

Medidas internas: SSE

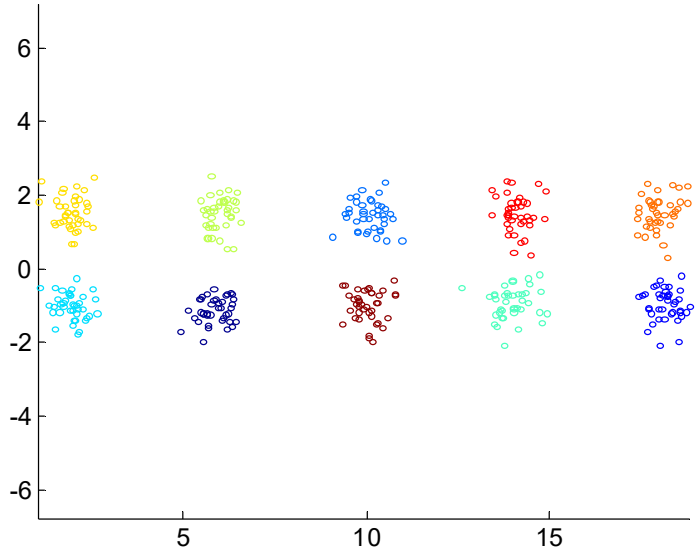
Las agrupaciones de las figuras más complicadas no están bien separados

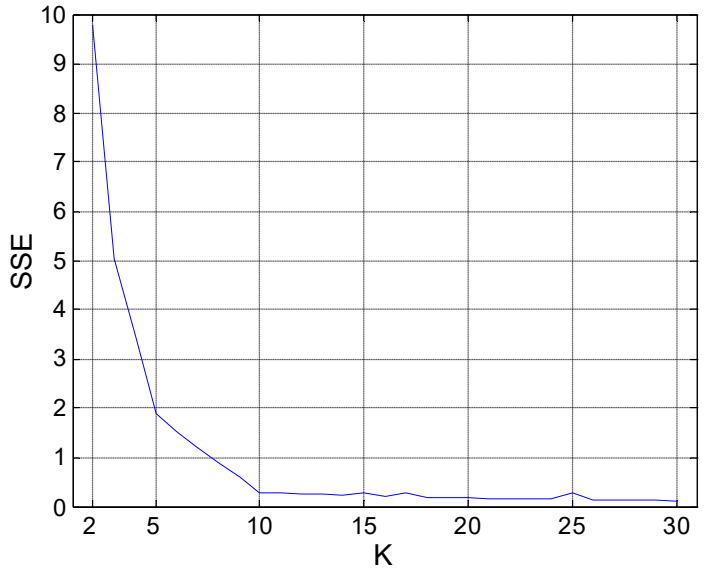
Índice interna: Se utiliza para medir la bondad de una estructura de agrupación sin respeto a la información externa

SSE

SSE es bueno para la comparación de dos agrupamientos o dos grupos (media SSE).

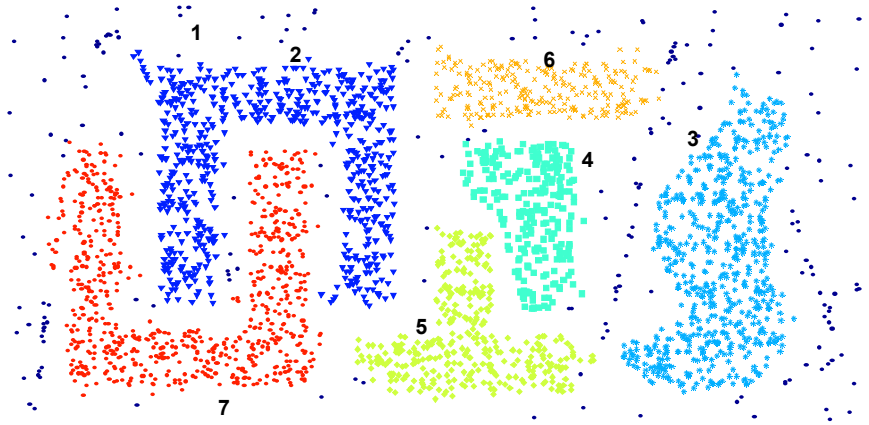
También puede utilizarse para estimar el número de agrupaciones

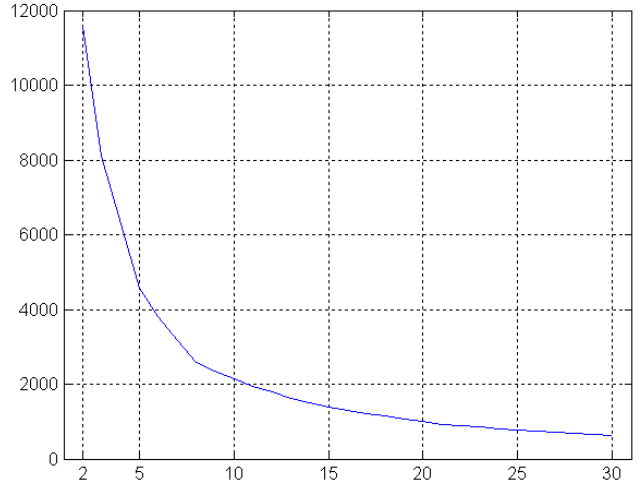




Medidas internas: SSE

Curva SSE par conjunto de datos mas complicados





SSE de cluster usando k-means

Vaidación interna

Cohesión clúster: Medidas cuán estrechamente relacionados son objetos de un clúster

Ejemplo: SSE

La separación de clúster: Medir la forma distinta o bien separa un cluster es de otras agrupaciones

Ejemplo: error al cuadrado

Cohesión se mide por la suma dentro de clúster de cuadrados (SSE)

$$\boldsymbol{WSS} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x} \in \boldsymbol{C}_i} (\boldsymbol{x} - \boldsymbol{m}_{\boldsymbol{i}})^2$$

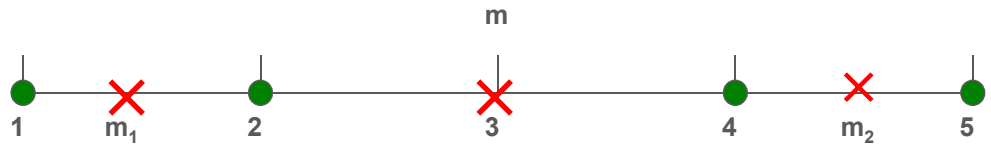
La separación se mide por la suma entre grupo de cuadrados

$$BSS = \sum_i |C_i|(m - m_i)^2$$

Vaidación interna

Example: SSE

- BSS + WSS = constante



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

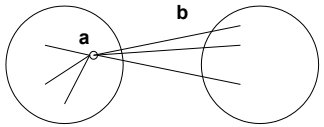
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Validación interna: Coeficiente de silueta

combinar ideas de la cohesión y la separación, pero para puntos individuales, así como las agrupaciones y la agrupación

Para un punto individual, i
Calcular una distancia media de i a los elementos en la agrupación
Calcula b = min (distancia media de i de puntos en otro grupo)
El coeficiente de la silueta de un punto viene dada entonces por s = 1 - a / b si a <b, (o s = b / a - 1 si un b, no es el caso habitual)
Típicamente entre 0 y 1.
Cuanto más cerca de 1, mejor.



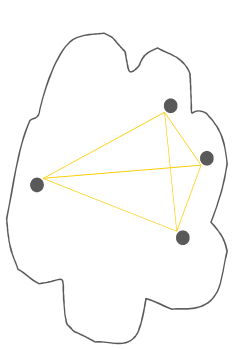
Se puede calcular el ancho promedio de la silueta por un grupo o una agrupación

Validación interna

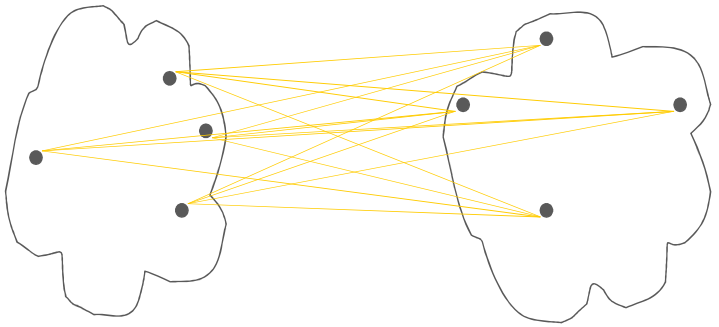
Un enfoque basado en el gráfico de proximidad se puede utilizar también para la cohesión y separación.

la cohesión del clúster es la suma del peso de todos los enlaces dentro de un grupo.

la separación del clúster es la suma de los pesos entre los nodos del clúster y los nodos fuera del clúster.



cohesion



separation

Vaidación interna

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Preguntas ...

¿Cuál criterio es el mejor?

¿Cómo elejir entre criterios de validación interna, externa o de comparación?

… Por otro lado …


CUANDO DICE QUE PREFIERE
QUE TENGA CUERPO ANTES QUE CARA
YO AMO EL GYM

¡¡ Ojo !!

Caundo se toman decisiones se debe considerar el problema especifico, el perfil del o de so decisores, las partes interesadas y en la medida de lo posible tener múltiples métricas con diferentes perspectivas.

Como veremos mas adelante el mismo espiritad se conserva en aprendizaje supervisado

THANK YOU!

ANY QUESTIONS?