DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

CC5206/CC71Q Minería de Datos
Análisis de Clusters

Benjamin Bustos

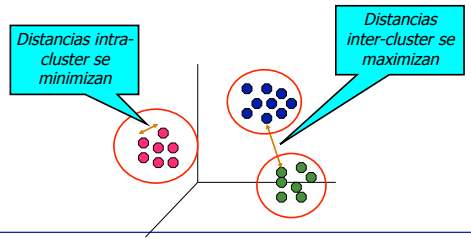
Departamento de Ciencias de la Computación
Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Introducción

- ¿Qué es análisis de clusters?
  - Encontrar grupos de objetos tal que los objetos en un grupo sean similares (o relacionados) entre sí y que sean diferentes (o no relacionados) a los objetos en otros grupos

Introducción

- ¿Qué es análisis de clusters?

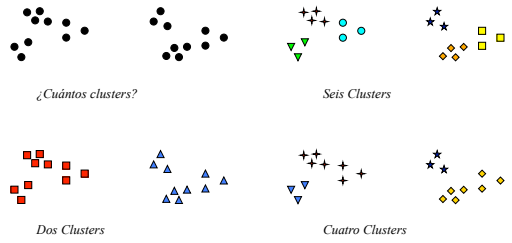


Introducción

- Análisis de clusters es una tarea esencial para muchas aplicaciones. Permite:
  - Encontrar clusters naturales y describir sus propiedades (data understanding)
  - Encontrar agrupamientos útiles (data class identification)
  - Encontrar representantes para grupos homogeneos (data reduction)
  - Encontrar objetos inusuales (outliers detection)
  - Encontrar perturbaciones aleatorios de los datos (noise detection),
  - Etc.

Introducción

- Noción de cluster puede ser ambigua

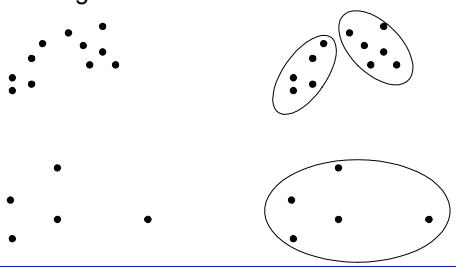


Introducción

- Tipos de clustering
  - Un clustering es un conjunto de clusters
  - Distinción importante entre conjuntos de clusters jerárquicos y particionales
  - Clustering Particional
    - Divide los datos en subconjuntos sin traslape (clusters), tal que cada dato está en un solo subconjunto
  - Clustering Jerárquico
    - Un conjunto de clusters anidados, organizados como un árbol

Introducción

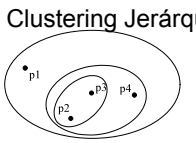
- Clustering Particional



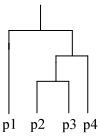
Puntos originales *Clustering Particional*

Introducción

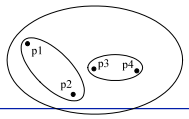
- Clustering Jerárquico



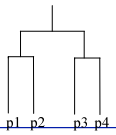
Clustering Jerárquico tradicional



Dendograma tradicional



Clustering Jerárquico no tradicional



Dendograma no tradicional

Introducción

- Tipos de clusters
  - Bien separados
  - Basados en un centro
  - Contiguos
  - Basados en densidad
  - Propiedad o Conceptual

Introducción

- Clusters bien separados
  - Un cluster es un conjunto de puntos tal que cualquier punto en un cluster está más cerca (más similar) a cualquier otro punto en el cluster que a cualquier punto no en el cluster

Tres clusters bien separados

Introducción

- Clusters basados en un centro
  - Un cluster es un conjunto de objetos tal que un objeto en él esta más cerca (más similar) al centro del cluster que al centro de cualquier otro
  - El centro de un cluster puede ser el centroide, el promedio de todos los puntos en el cluster, o el medioide, el punto más "representativo" del cluster

Cuatro clusters basados en centros

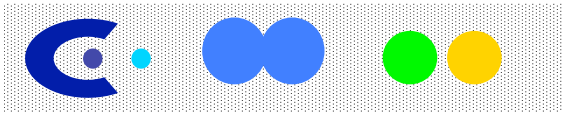
Introducción

- Clusters contiguos (vecino más cercano o transitivo)
  - Un cluster es un conjunto de puntos tal que un punto en un cluster está más cerca (más similar) a uno o más puntos en el cluster que a cualquier punto no en el cluster

Ocho clusters contiguos

Introducción

- Clusters basados en densidad
  - Un cluster es una región densa de puntos, separada por regiones de baja densidad de otras regiones de alta densidad
  - Usado cuando los clusters son irregulares o están entrelazados, y cuando hay ruido y outliers



Seis clusters basados en densidad

13

Introducción

- Clusters conceptuales o que comparten propiedad
  - Encuentra clusters que comparten alguna propiedad común o representan un concepto particular

Dos círculos que traslapan

14

Métodos de clustering

- K-means
- Método jerárquico aglomerativo
- DBSCAN

15

K-means

- Método de clustering particional
  - Cada cluster está asociado a un centroide
  - Cada punto se asigna al cluster cuyo centroide sea el más cercano
  - Número de clusters, K, parámetro del método
- Algoritmo

Algorithm 1 Basic K-means Algorithm.

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

16

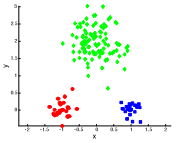
K-means

- Detalles del algoritmo
  - Centroides iniciales: aleatorios
    - Clusters varían dependiendo de la elección
  - Centroide es (típicamente) la media de los puntos en el cluster
  - "Cercanía" se mide con alguna distancia
  - K-means converge para distancias "usuales"
  - En general la convergencia sucede con pocas iteraciones
    - Iterar hasta que cambien "pocos" puntos de cluster
  - Complejidad es O(n * K * I * d)
    - n puntos, K centros, I iteraciones, d dimensiones

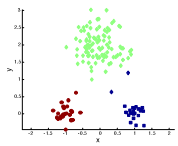
17

K-means

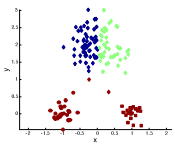
- Ejemplo:



Puntos originales



Clustering óptimo

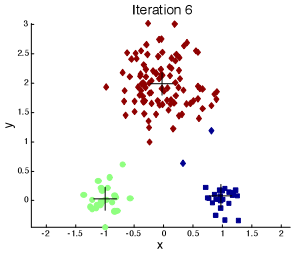


Clustering sub-optimal

18

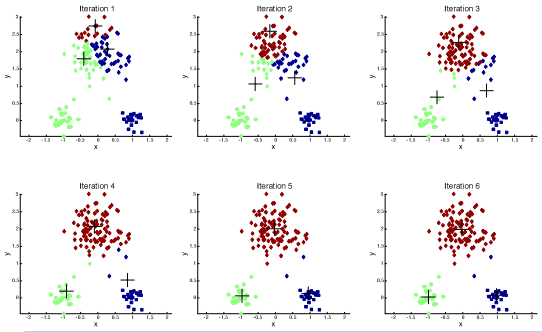
K-means

- Importancia de escoger centros iniciales



19

K-means



20

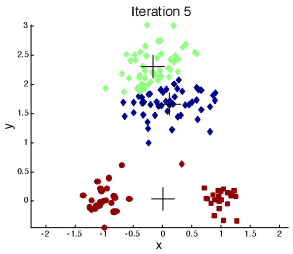
K-means

- Evaluando clusters obtenidos con K-means
  - Medida más común: Sum of Squared Error (SSE)
  - Por cada punto, error es distancia al cluster más cercano
  - $SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$
  - x: punto en cluster C_i, m_i: centroide de C_i
  - Dados dos clusters, escoger el que tenga menor error
  - Forma de reducir SSE: aumentar K
    - Buen clustering reduce SSE, incluso para menor K

21

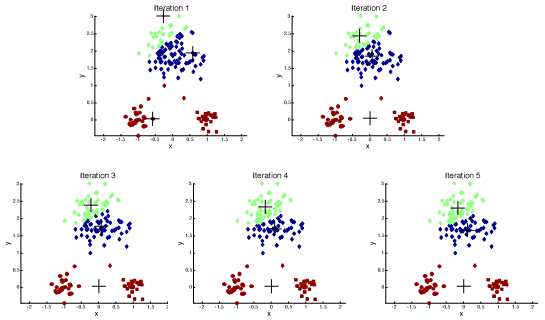
K-means

- Importancia de escoger centros iniciales



22

K-means



23

K-means

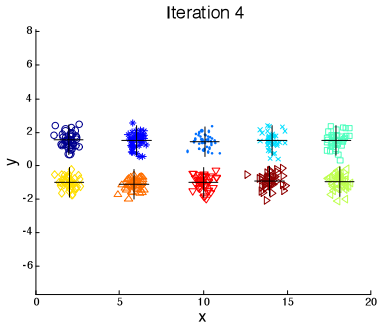
- Problemas para escoger centros iniciales
  - Si hay K clusters "reales", probabilidad de escoger un centroide por cluster es baja

$$P = \frac{\#\text{ formas de escoger un centroide de cada cluster}}{\#\text{ formas de escoger K centroides}} = \frac{K! n^K}{(Kn)^K} = \frac{K!}{K^K}$$

  - Ej.: si K = 10, entonces $P = 10!/10^{10} = 0.00036$
  - Centroides iniciales pueden o no "reajustarse" en la forma correcta (converger a óptimo local)
  - Ejemplo: cinco pares de clusters (10 en total)

24

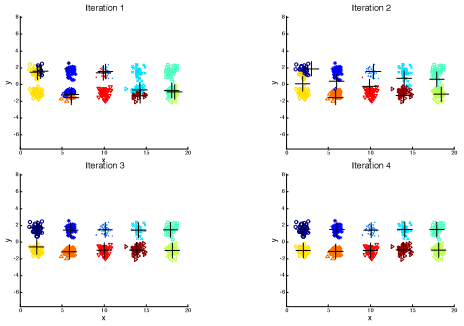
K-means



Empezando con dos centroides iniciales en un cluster por cada par de clusters

25

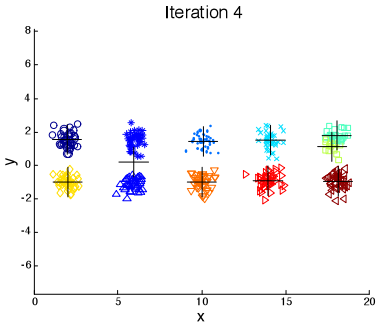
K-means



Empezando con dos centroides iniciales en un cluster por cada par de clusters

26

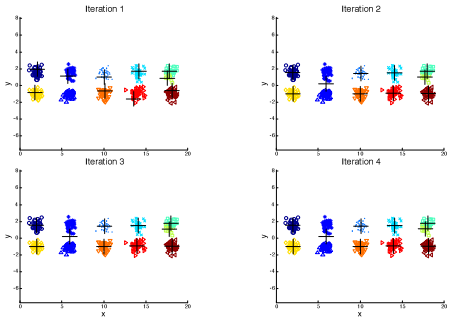
K-means



Empezando con algunos pares de clusters con tres centroides iniciales, mientras otros tienen sólo uno.

27

K-means



Empezando con algunos pares de clusters con tres centroides iniciales, mientras otros tienen sólo uno.

28

K-means

- Solución al problema de centroides iniciales
  - Ejecutar varias veces K-means
    - Ayuda, pero la probabilidad sigue siendo baja
  - Muestrear y usar clustering jerárquico para determinar centroides iniciales
  - Elegir más de K centroides iniciales
    - Luego escoger los K más separados entre sí
  - Post procesamiento
  - Bisecting K-means
    - No tan susceptible a problemas de inicialización

29

K-means

- Manejando clusters vacíos
  - Algoritmo K-means puede retornar clusters vacíos
  - Estrategias para encontrar centroide de reemplazo:
    - Escoger el punto más lejano a su centroide (punto que contribuye más al SSE)
    - Escoger punto del cluster con máximo SSE
      - Típicamente resulta en dividir dicho cluster
    - Si hay varios clusters vacíos, repetir estas estrategias varias veces

30

K-means

- Actualizar centros incrementalmente
  - Algoritmo original: centroides se actualizan después de asignar todos los puntos
  - Alternativa: actualizar centroides después de cada asignación
    - Cada asignación actualiza cero o dos centroides
    - Más costoso
    - Introduce dependencia en el orden de los puntos
    - Nunca se obtiene un cluster vacío

K-means

- Preprocesamiento
  - Normalizar los datos
  - Eliminar outliers
- Postprocesamiento
  - Eliminar clusters pequeños que puedan representar outliers
  - Partir clusters "sueltos" (con alto SSE)
  - Mezclar clusters cercanos y con bajo SSE

K-means

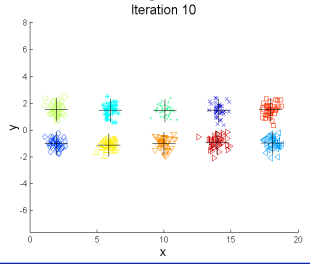
- Bisecting K-means
  - Variante que puede producir clustering jerárquico o particional

Algorithm 3 Bisecting K-means Algorithm.

```
1: Initialize the list of clusters to contain the cluster containing all points.
2: repeat
3:    Select a cluster from the list of clusters
4:    for i = 1 to number_of_iterations do
5:       Bisect the selected cluster using basic K-means
6:    end for
7:    Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: until Until the list of clusters contains K clusters
```


K-means

- Ejemplo de Bisecting K-means

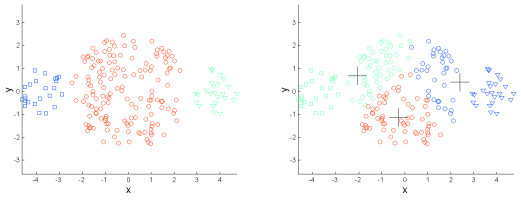


K-means

- Limitaciones de K-means
  - Clusters de diferente tamaño
  - Clusters de diferentes densidades
  - Clusters con formas no esféricas
- K-means no es robusto a outliers

K-means

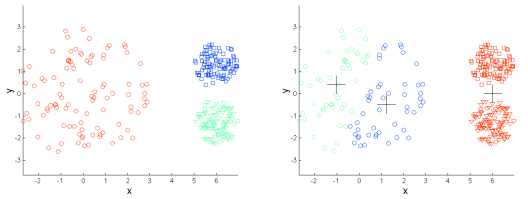
- Ejemplo: tamaños diferentes



Puntos originales *K-means (tres clusters)*

K-means

- Ejemplo: densidades diferentes

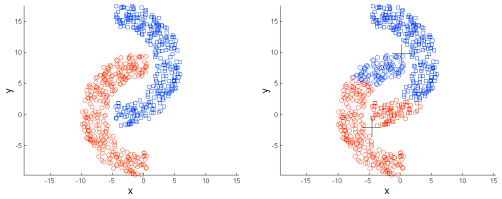


Puntos originales *K-means (tres clusters)*

37

K-means

- Ejemplo: formas no esféricas

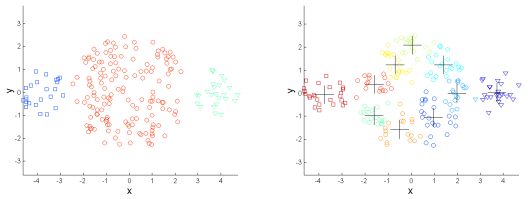


Puntos originales *K-means (dos clusters)*

38

K-means

- Solución: usar K alto, luego mezclar clusters

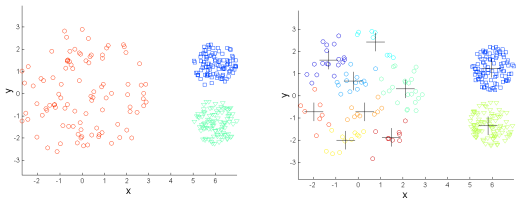


Puntos originales *K-means clusters*

39

K-means

- Solución: usar K alto, luego mezclar clusters

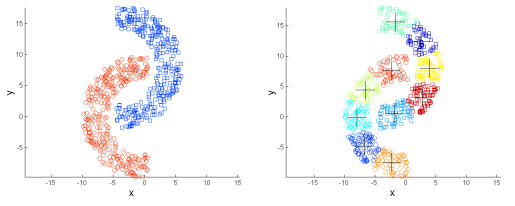


Puntos originales *K-means clusters*

40

K-means

- Solución: usar K alto, luego mezclar clusters

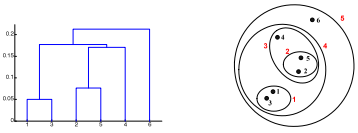


Puntos originales *K-means clusters*

41

Clustering Jerárquico Aglomerativo

- Produce un conjunto de clusters anidados organizados en un árbol jerárquico
- Puede visualizarse como un dendograma
  - Diagrama parecido a un árbol que registra la secuencias de mezclas o divisiones



42

Clustering Jerárquico Aglomerativo

- Fortalezas
  - No tiene que suponer un número a priori de clusters
    - Se puede obtener cualquier número de clusters deseado "cortando" el dendograma en el nivel apropiado
  - Clusters pueden corresponer a taxonomía
    - Ejemplos en biología

43

Clustering Jerárquico Aglomerativo

- Tipos principales de clustering jerárquico
  - Aglomerativo
    - Empezar con cada punto como cluster individual
    - En cada paso, mezclar el par de clusters más cercano hasta que quede sólo un cluster (o *k* clusters)
  - Divisivo
    - Empezar con un cluster que contenga todos los puntos
    - En cada paso, dividir un cluster en dos hasta que todo cluster contenga un solo punto (o haya *k* clusters)
- Requieren calcular matriz de distancias

44

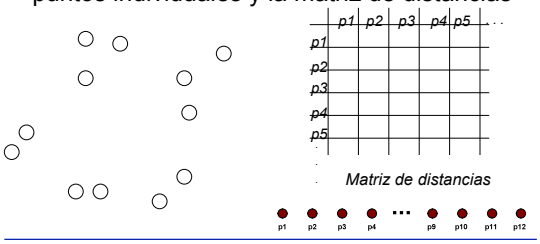
Clustering Jerárquico Aglomerativo

- Algoritmo básico (aglomerativo)
  1. Calcular matriz de distancias
  2. Sea cada punto un cluster
  3. **Repetir**
  4. Mezclar par de clusters más cercano
  5. Actualizar matriz de distancias
  6. **Hasta** que quede sólo un cluster
- Operación clave: cálculo de la distancia entre clusters
  - Diferentes formas de hacerlo distinguen a los diferentes algoritmos

45

Clustering Jerárquico Aglomerativo

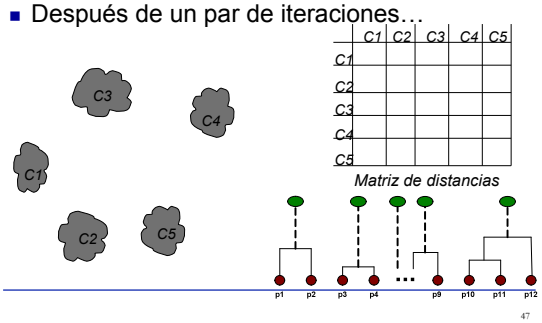
- Situación inicial: empezar con clusters de puntos individuales y la matriz de distancias



46

Clustering Jerárquico Aglomerativo

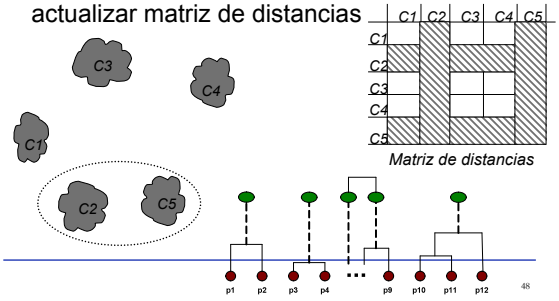
- Después de un par de iteraciones…



47

Clustering Jerárquico Aglomerativo

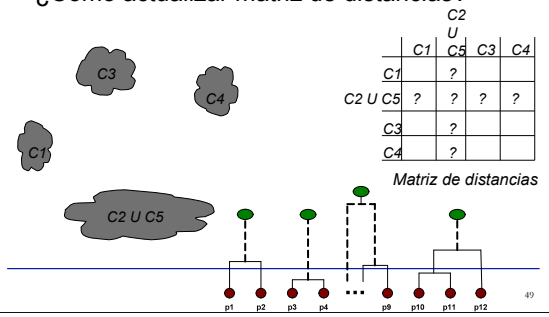
- … mezclar clusters más cercano (C2 y C5) y actualizar matriz de distancias



48

Clustering Jerárquico Aglomerativo

- ¿Cómo actualizar matriz de distancias?

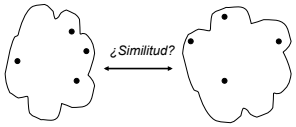


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de distancias

Clustering Jerárquico Aglomerativo

- ¿Cómo definir distancias entre clusters?



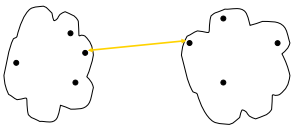
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

Matriz de distancias

Clustering Jerárquico Aglomerativo

- ¿Cómo definir distancias entre clusters?



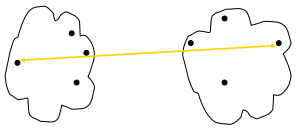
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

Matriz de distancias

Clustering Jerárquico Aglomerativo

- ¿Cómo definir distancias entre clusters?



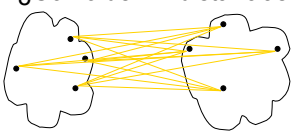
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

Matriz de distancias

Clustering Jerárquico Aglomerativo

- ¿Cómo definir distancias entre clusters?



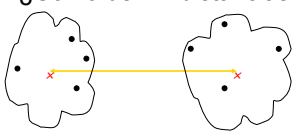
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo (método de Ward)

Matriz de distancias

Clustering Jerárquico Aglomerativo

- ¿Cómo definir distancias entre clusters?



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

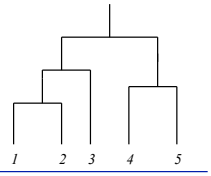
- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

Matriz de distancias

Clustering Jerárquico Aglomerativo

- Distancia MIN (o single link)
  - Similitud se basa en los dos puntos más cercanos del par de clusters
    - Determinado por un par de puntos, i.e., por un enlace en el grafo de distancias

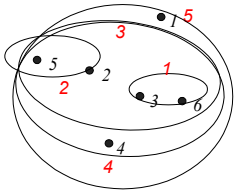
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

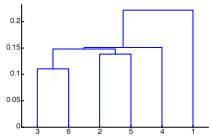


55

Clustering Jerárquico Aglomerativo

- Distancia MIN (o single link)



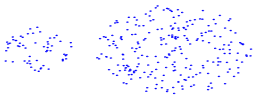


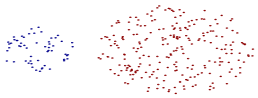
Clusters anidados *Dendrograma*

56

Clustering Jerárquico Aglomerativo

- Fortaleza de distancia MIN





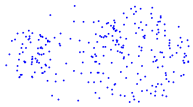
Puntos originales *Dos Clusters*

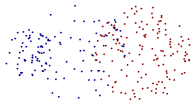
- *Puede manejar formas no-elípticas*

57

Clustering Jerárquico Aglomerativo

- Limitaciones de distancia MIN





Puntos originales *Dos Clusters*

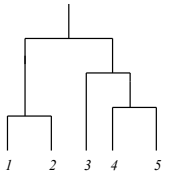
- *Sensible a ruido y outliers*

58

Clustering Jerárquico Aglomerativo

- Distancia MAX (o complete linkage)
  - Similitud se basa en los dos puntos más lejanos del par de clusters
    - Determinado por todos los pares de puntos en los clusters

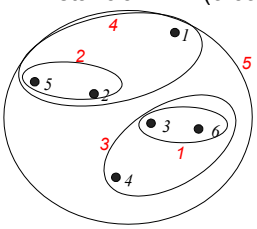
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

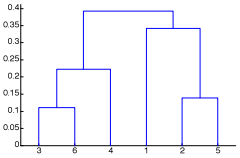


59

Clustering Jerárquico Aglomerativo

- Distancia MAX (o complete linkage)



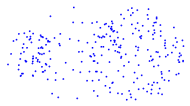


Clusters anidados *Dendrograma*

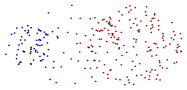
60

Clustering Jerárquico Aglomerativo

- Fortaleza de MAX



Puntos originales

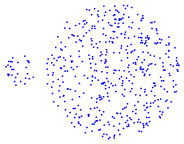


Dos Clusters

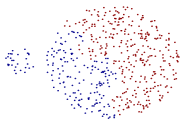
- *Menos suceptible a ruido y outliers*

Clustering Jerárquico Aglomerativo

- Limitaciones de MAX



Puntos originales



Dos Clusters

- *Tiende a quebrar clusters grandes*
- *Sesgado a clusters esféricos*

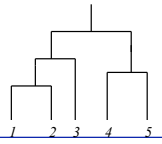
Clustering Jerárquico Aglomerativo

- Distancia promedio de grupo
  - Similitud es el promedio de distancias entre pares de puntos de los clusters

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

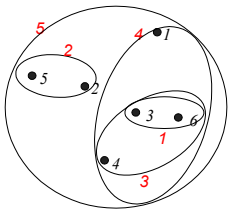
  - Método intermedio entre single link y complete link

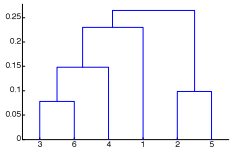
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Clustering Jerárquico Aglomerativo

- Distancia promedio de grupo





Clusters anidados *Dendrograma*

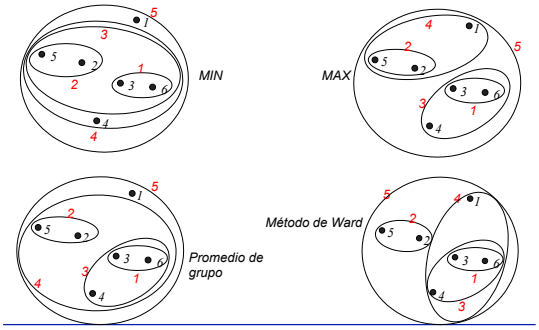
Clustering Jerárquico Aglomerativo

- Distancia promedio de grupo
  - Compromiso entre MIN y MAX
  - Fortalezas
    - Menos suceptible a ruido y outliers
  - Limitaciones
    - Sesgado a clusters esféricos

Clustering Jerárquico Aglomerativo

- Método de Ward
  - Similitud entre clusters se basa en el incremento del SSE cuando se mezclan dos clusters
    - Similar a distancia promedio de grupo si la distancia entre puntos es distancia cuadrada
  - Menos suceptible a ruido y outliers
  - Sesgado a clusters esféricos
  - Analogo jerárquico de K-means
    - Pueda usarse para inicializar K-means

Clustering Jerárquico Aglomerativo



Clustering Jerárquico Aglomerativo

- Requerimientos de tiempo y espacio
  - Espacio: O(N^2) para guardar matriz de distancias
    - N: número de puntos
  - Tiempo: O(N^3) en muchos casos
    - Para N pasos, se debe actualizar matriz de similitud en cada paso
    - Complejidad puede reducirse a O(N^2 log N) usando listas ordenadas o heaps

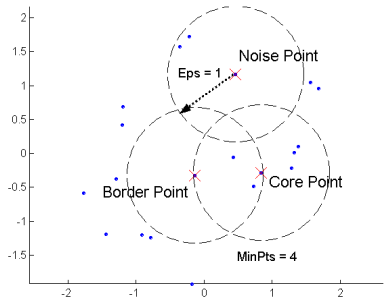
Clustering Jerárquico Aglomerativo

- Problemas y limitaciones
  - Una vez decidido unir dos clusters, no se puede deshacer
  - No hay una función objetivo que sea directamente minimizada
  - Problemas de los diferentes esquemas:
    - Sensibles a ruido y outliers
    - Dificultad para manejar clusters de distinto tamaño
    - Pueden romper clusters grandes

DBSCAN

- Algoritmo de clustering basado en densidad
  - Densidad = número de puntos dentro de un radio especificado (Eps)
  - Punto "core": tiene más puntos que un valor especificado de puntos (MinPts) a distancia Eps
    - Éstos son los puntos dentro del cluster
  - Punto "border": tiene menos que MinPts puntos en el radio Eps, pero esta en la vecindad de un punto core
  - Punto "noise": cualquier punto que no sea core ni border

DBSCAN

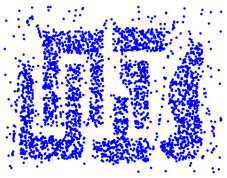


DBSCAN

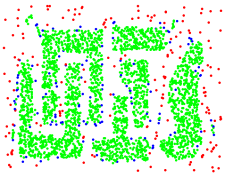
- Algoritmo DBSCAN
  - Eliminar puntos de ruido
  - Realizar clustering en los puntos que queden

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← 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



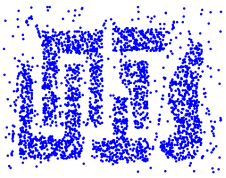
Puntos originales



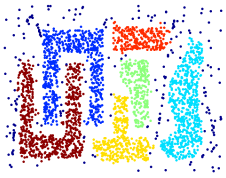
Tipos de punto: core, border y noise

Eps = 10, MinPts = 4

DBSCAN



Puntos originales

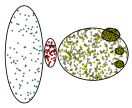


Clusters

- Resistente a ruido
- Puede encontrar clusters de diferentes formas y tamaños

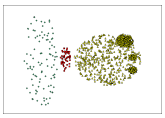
DBSCAN

- No funciona bien en:

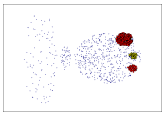


Puntos originales

- Densidades variables
- Datos multidimensionales



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

DBSCAN

- Determinando Eps y MinPts
  - Para los puntos en un cluster, su k-NN se encuentra aprox. a la misma distancia
  - Puntos de ruido tienen su k-NN a mayor distancia
  - Graficar distancia ordenada de cada punto a su k-NN