

Clase 10: Algoritmos de Clustering

Bárbara Poblete

*(Basado en slides traducidos por B.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

2

Introducción

- 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

3

Métodos de clustering

- K-means
- Método jerárquico aglomerativo
- DBSCAN

4

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

5

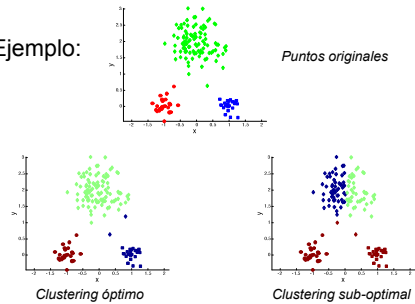
K-means

- 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

6

K-means

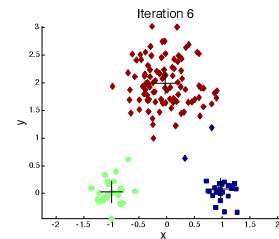
Ejemplo:



7

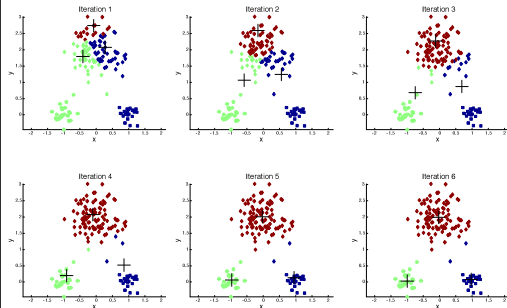
K-means

Importancia de escoger centros iniciales



8

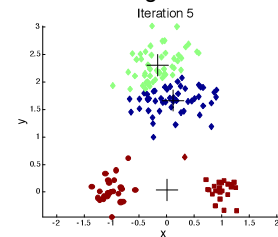
K-means



9

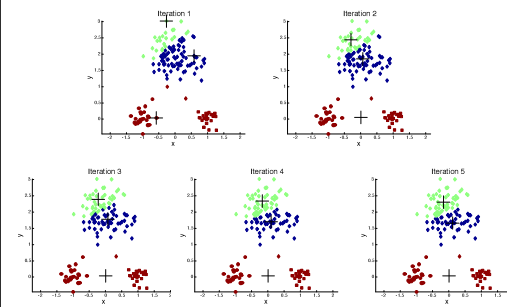
K-means

Importancia de escoger centros iniciales



10

K-means



11

K-means

Problemas para escoger centros iniciales

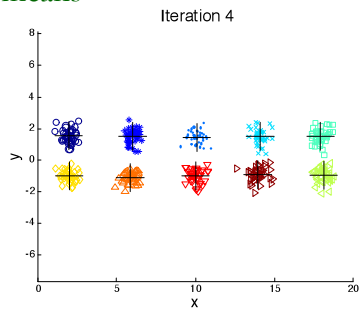
- Si hay K clusters "reales", probabilidad de escoger un centroide por cluster es baja

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

- 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)

12

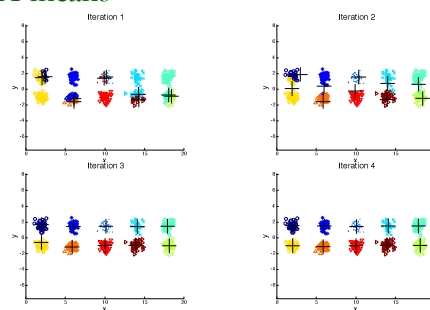
K-means



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

13

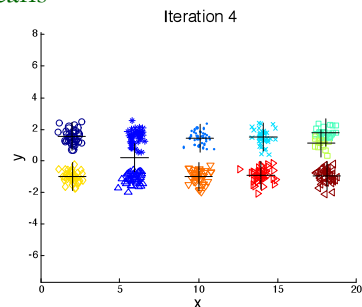
K-means



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

14

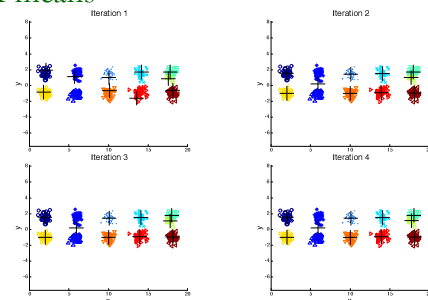
K-means



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

15

K-means



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

16

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

17

K-means

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

20

K-means

■ Bisecting K-means

- Variante que puede producir clustering jerárquico o particional

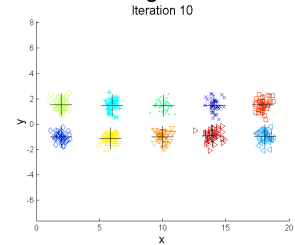
Algorithm 3 Bisecting K-means Algorithm.

```
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
```

21

K-means

■ Ejemplo de Bisecting K-means



22

K-means

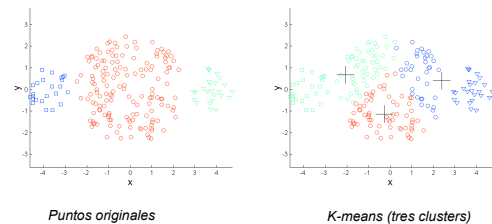
■ Limitaciones de K-means

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

23

K-means

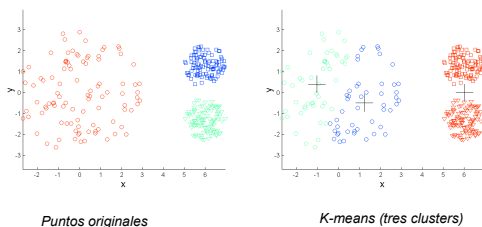
■ Ejemplo: tamaños diferentes



24

K-means

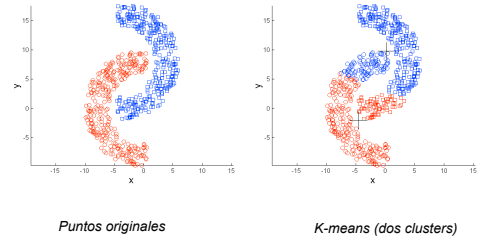
■ Ejemplo: densidades diferentes



25

K-means

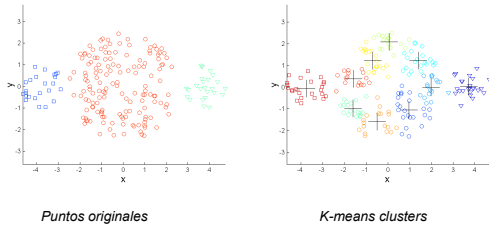
■ Ejemplo: formas no esféricas



26

K-means

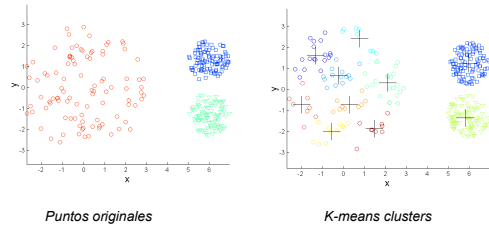
- Solución: usar K alto, luego mezclar clusters



27

K-means

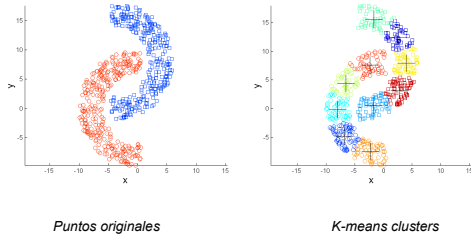
- Solución: usar K alto, luego mezclar clusters



28

K-means

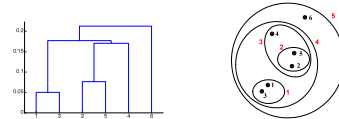
- Solución: usar K alto, luego mezclar clusters



29

Clustering Jerárquico Aglomerativo

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



30

Clustering Jerárquico Aglomerativo

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

31

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 proximidad

32

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

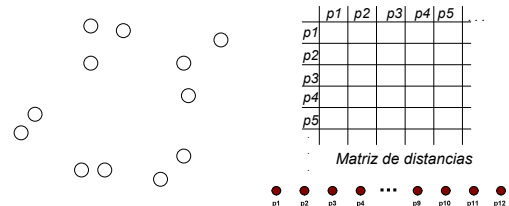
■ Operación clave: cálculo de la distancia entre clusters

- Diferentes formas de hacerlo distinguen a los diferentes algoritmos

33

Clustering Jerárquico Aglomerativo

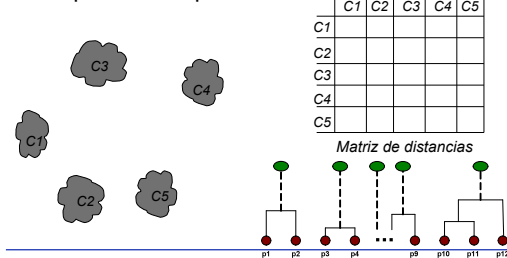
■ Situación inicial: empezar con clusters de puntos individuales y la matriz de distancias



34

Clustering Jerárquico Aglomerativo

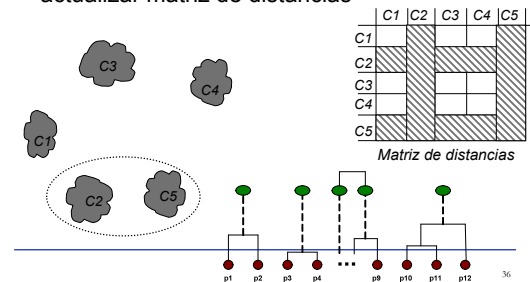
■ Después de un par de iteraciones...



35

Clustering Jerárquico Aglomerativo

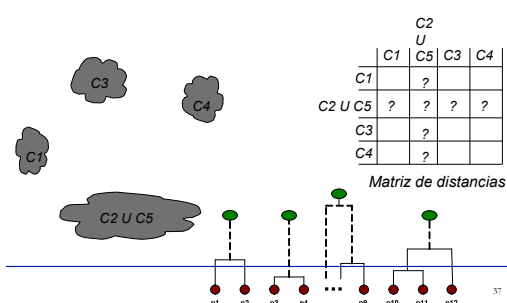
■ ... mezclar clusters más cercano (C2 y C5) y actualizar matriz de distancias



36

Clustering Jerárquico Aglomerativo

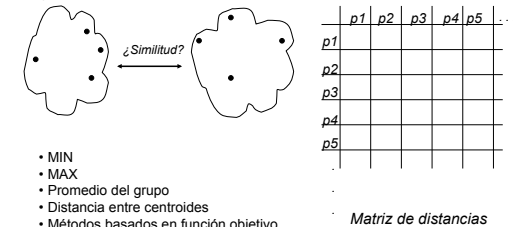
■ ¿Cómo actualizar matriz de distancias?



37

Clustering Jerárquico Aglomerativo

■ ¿Cómo definir distancias entre clusters?

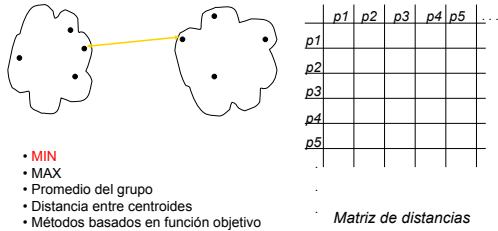


- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

38

Clustering Jerárquico Aglomerativo

¿Cómo definir distancias entre clusters?

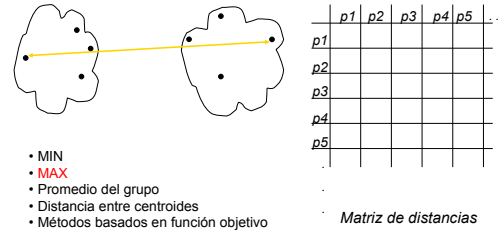


- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

39

Clustering Jerárquico Aglomerativo

¿Cómo definir distancias entre clusters?

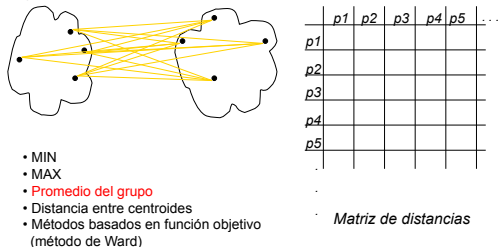


- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

40

Clustering Jerárquico Aglomerativo

¿Cómo definir distancias entre clusters?

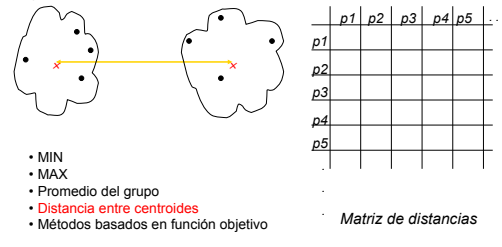


- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo (método de Ward)

41

Clustering Jerárquico Aglomerativo

¿Cómo definir distancias entre clusters?

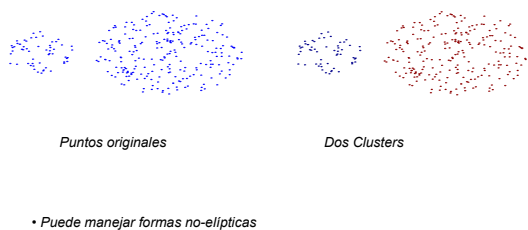


- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Métodos basados en función objetivo

42

Clustering Jerárquico Aglomerativo

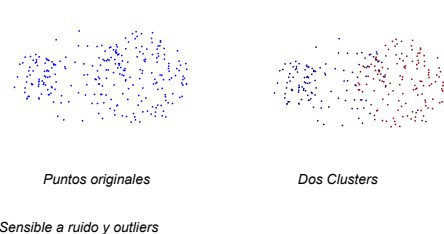
Fortaleza de distancia MIN



45

Clustering Jerárquico Aglomerativo

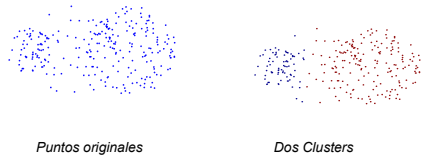
Limitaciones de distancia MIN



46

Clustering Jerárquico Aglomerativo

■ Fortaleza de MAX

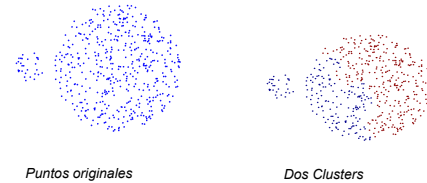


- Menos susceptible a ruido y outliers

49

Clustering Jerárquico Aglomerativo

■ Limitaciones de MAX



- Tiende a quebrar clusters grandes
- Sesgado a clusters esféricos

50

Clustering Jerárquico Aglomerativo

■ Distancia promedio de grupo

- Compromiso entre MIN y MAX
- Fortalezas
 - Menos susceptible a ruido y outliers
- Limitaciones
 - Sesgado a clusters esféricos

53

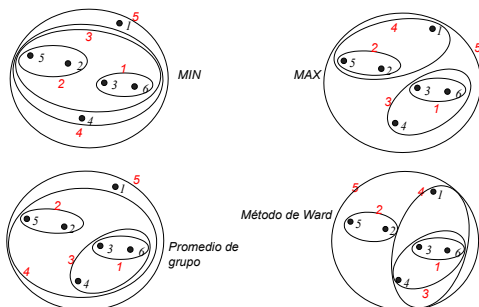
Clustering Jerárquico Aglomerativo

■ 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 susceptible a ruido y outliers
- Sesgado a clusters esféricos
- Análogo jerárquico de K-means
 - Puede usarse para inicializar K-means

54

Clustering Jerárquico Aglomerativo



55

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

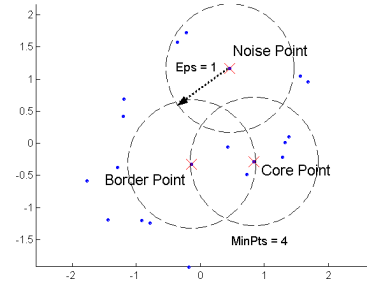
57

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

58

DBSCAN



59

DBSCAN

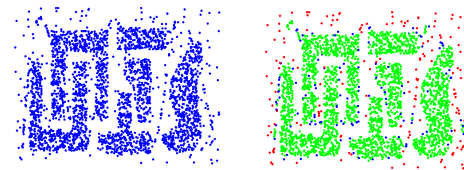
- Algoritmo DBSCAN
 - Eliminar puntos de ruido
 - Realizar clustering en los puntos que queden
- ```

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 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

```

60

## DBSCAN



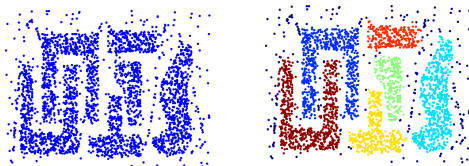
Puntos originales

Tipos de punto: core,  
border y noise

Eps = 10, MinPts = 4

61

## DBSCAN



Puntos originales

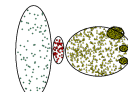
Clusters

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

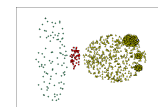
62

## DBSCAN

- No funciona bien en:

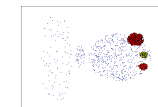


Puntos originales



(MinPts=4,  
Eps=9.92)

- Densidades variables
- Datos multidimensionales



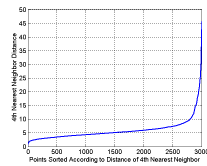
(MinPts=4,  
Eps=9.75).

63

## 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



64