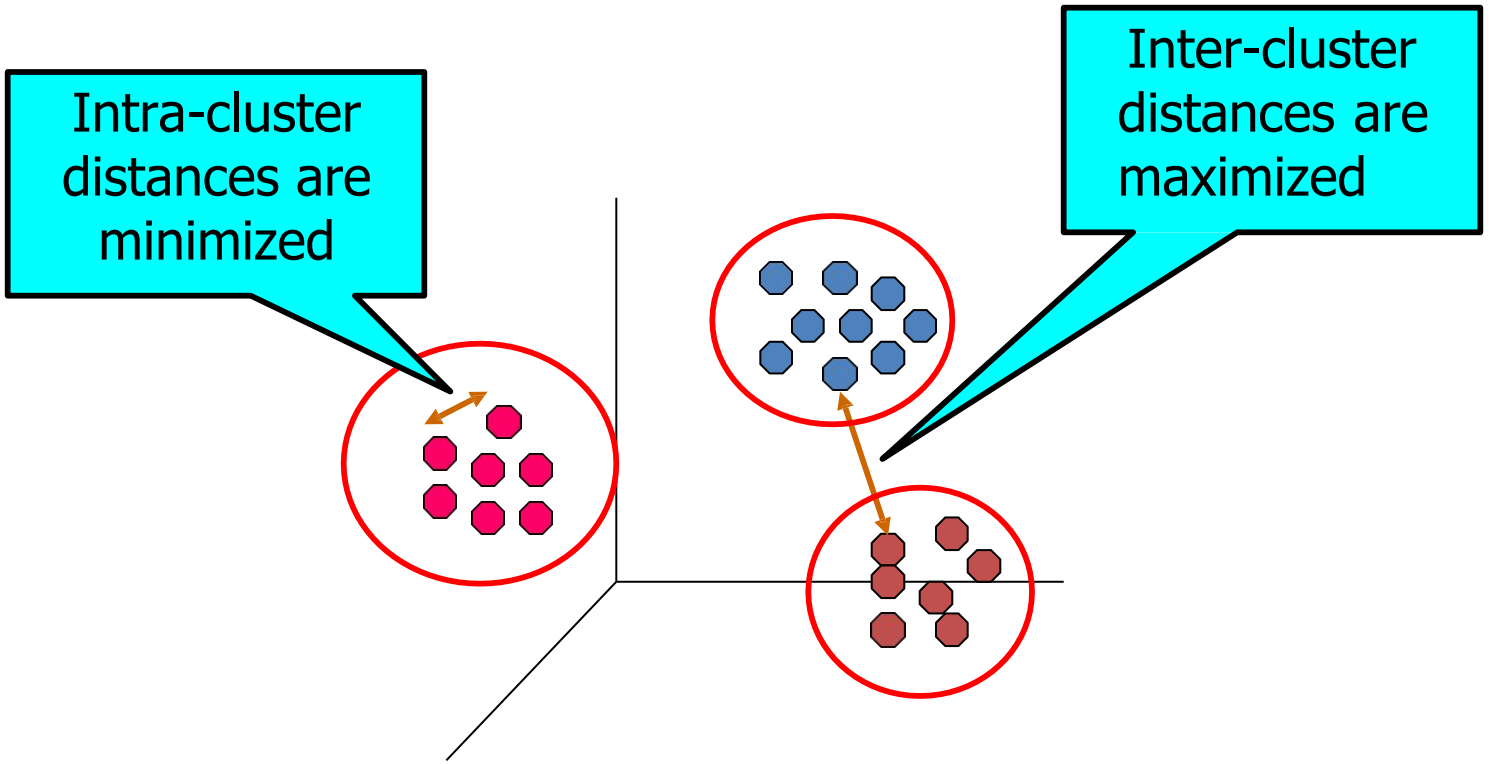
Clustering – Métodos de Aglomeramiento

Análisis de Agrupamientos/Clusters

- Consiste en encontrar grupos tales que los objetos dentro de un mismo grupo sean similares (o estén relacionados) entre sí y diferentes (o no relacionados) a los objetos de los otros grupos



Similitud y Distancia

DISTANCIA inversa a SIMILITUD.

- Muchísimas formas de calcular la distancia. Algunos ejemplos…
  - ***Distancia Euclídea:*** $\sqrt{\sum_{i=1}^{n} (x_i - y)^2}$
  - ***Distancia de Manhattan:*** $\sum_{i=1}^{n} |x_i - y_i|$

  } Valores Continuos (conveniente estandarizar antes)

  - ***Distancias por Diferencia:***
    ejemplo: if *x*=*y* then *D*=0 else *D*=1

  } Valores Discretos

- Y muchísimas formas más…

¿Cuál es la cantidad adecuada de grupos?

¿Cuántos grupos?

6

2

4

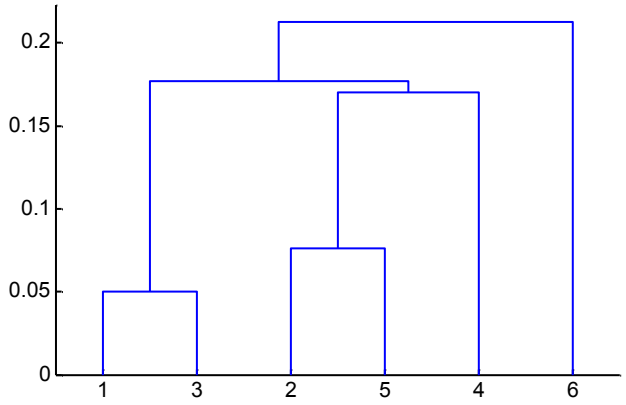
Tipos de agrupamientos (clustering)

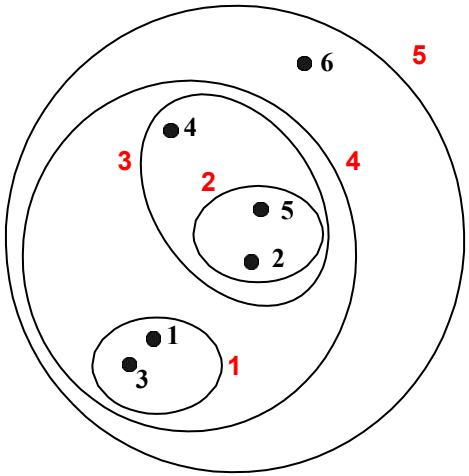
- Jerárquico
  - Consiste de un conjunto de grupos anidados organizados como una estructura de árbol

- No Jerárquico – (Ejemplo: K-means)
  - Se divide a los objetos de datos en subconjuntos (clusters) no superpuestos de forma tal que un objeto pertenece solo a un subconjunto

Clustering Jerárquico

- Genera un conjunto de grupos anidados organizados como una estructura de árbol
- Se puede visualizar como un dendograma
  - Un diagrama tipo árbol que registra las secuencias de uniones y separaciones





Clustering Jerárquico - Fortalezas

- No asume ninguna cantidad de grupos en particular
  - Se puede obtener el número deseado de grupos 'cortando' el dendograma en el nivel apropiado.
- Puede corresponderse con taxonomías significativas
  - Por ejemplo en ciencias biológicas (reino animal, reconstrucción genética, …)

Clustering Jerárquico

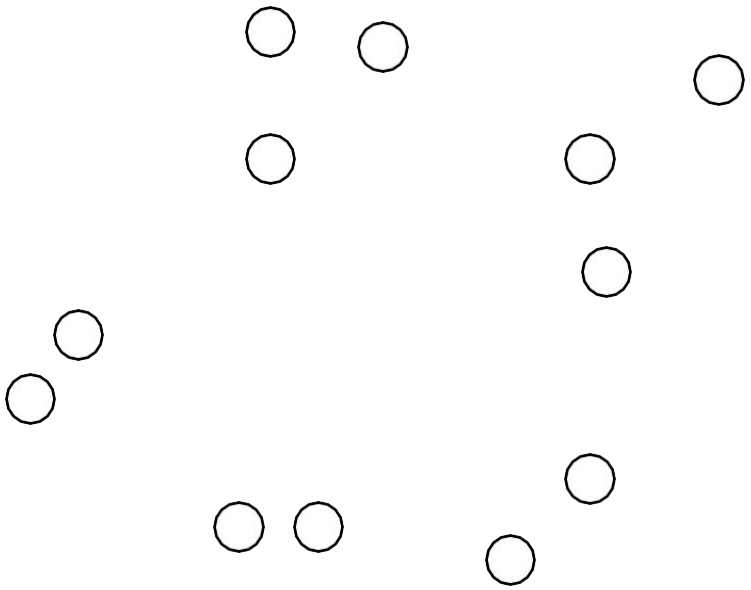
- Dos tipos principales de clustering jerárquico
  - Aglomerativo:
    - Comienza con los elementos como clusters individuales
    - En cada paso agrupa los pares de clusters más cercanos hasta que queda uno solo
  - Divisivo:
    - Comienza con un único cluster que incluye a todos los elementos
    - En cada paso divide un cluster hasta que cada cluster contiene un único elemento
- Los algoritmos jerárquicos tradicionales utilizan una matriz de similaridad o distancia
  - Unen o separan un cluster por vez

Algoritmo de Clustering Aglomerativo

- Es la técnica más popular de clustering jerárquico
- El algoritmo básico es simple
  1. Calcular la matriz de proximidad
  2. Tomar a cada elemento como un cluster
  3. **Repetir**
  4. Unir los dos clusters más cercanos
  5. Actualizar la matriz de proximidad
  6. **Hasta** que solo queda un cluster
- La operación clave es el cómputo de la proximidad entre dos clusters
  - Existen diferentes algoritmos de acuerdo a la forma de definir la distancia entre clusters

Algoritmo de Clustering Aglomerativo
Situación inicial

- Comienza con los elementos como clusters individuales

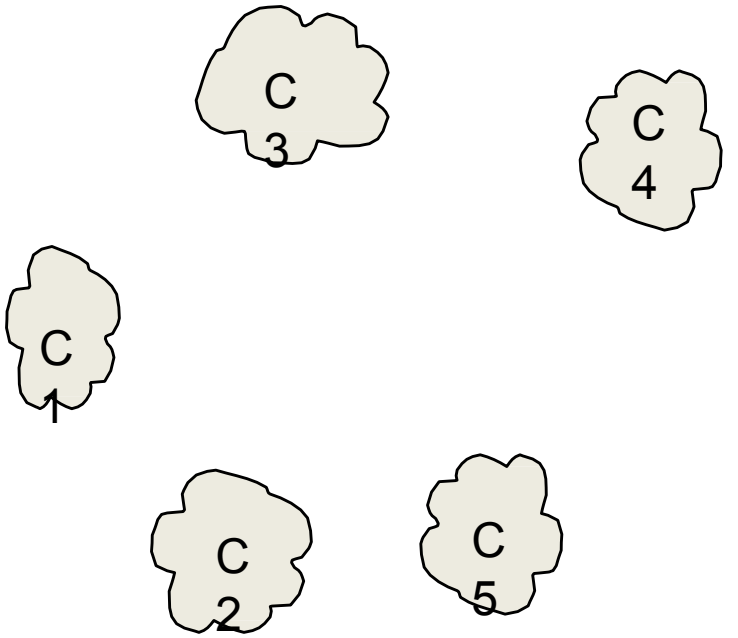


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
:						

Matriz de Proximidad

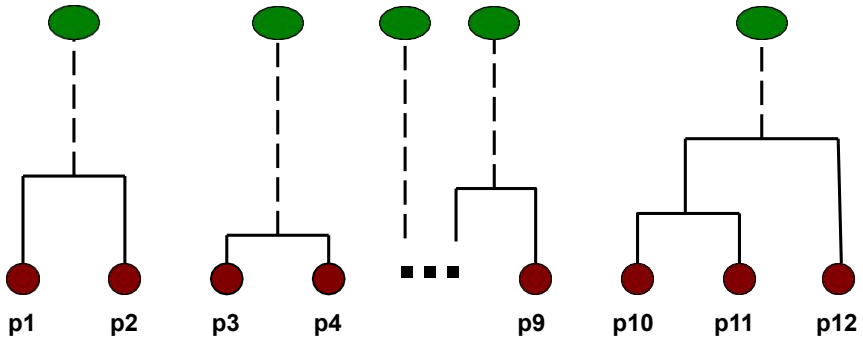
Algoritmo de Clustering Aglomerativo
Situación intermedia

- Luego de algunos pasos se tienen grupos



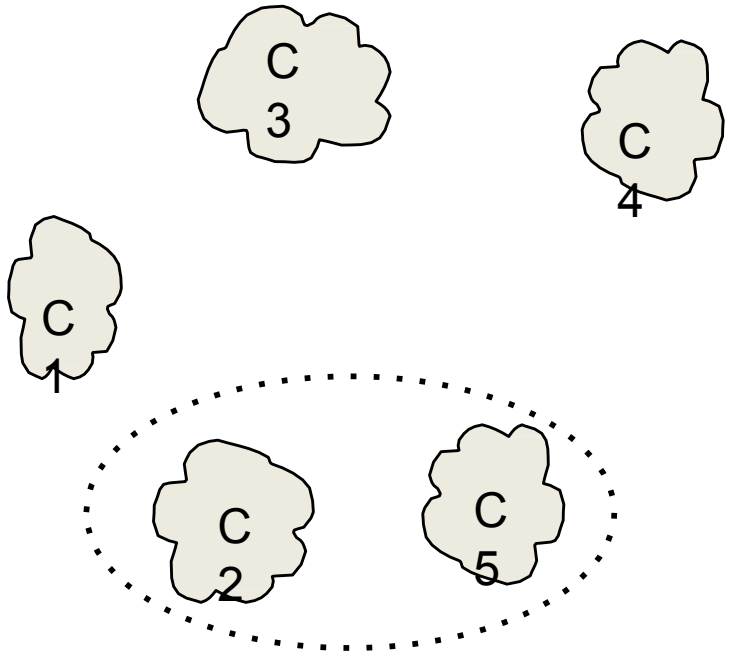
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de Proximidad



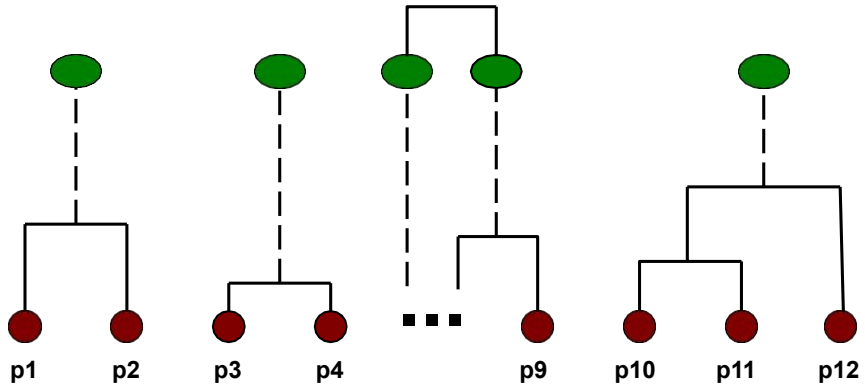
Algoritmo de Clustering Aglomerativo Situación Intermedia

- Se unen los dos grupos más cercanos (C2 y C5) y se actualiza la matriz.
- ¿Qué criterio se utiliza para la actualización?

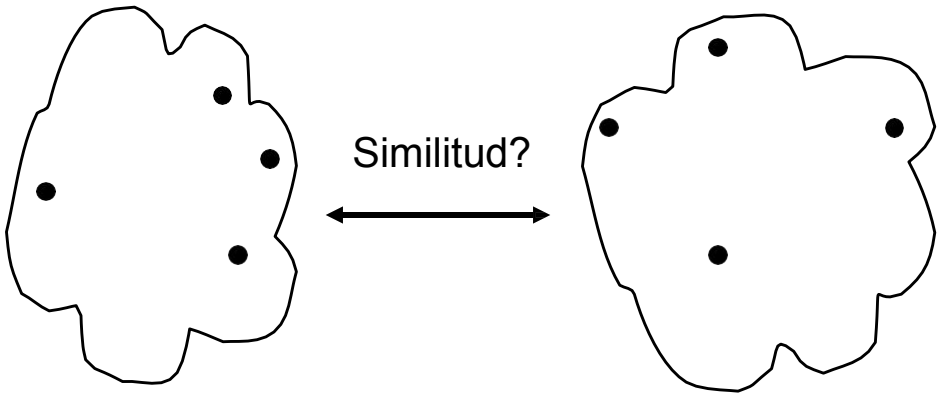


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de Proximidad



Algoritmo de Clustering Aglomerativo
Definición de similaridad entre clusters (linkage)

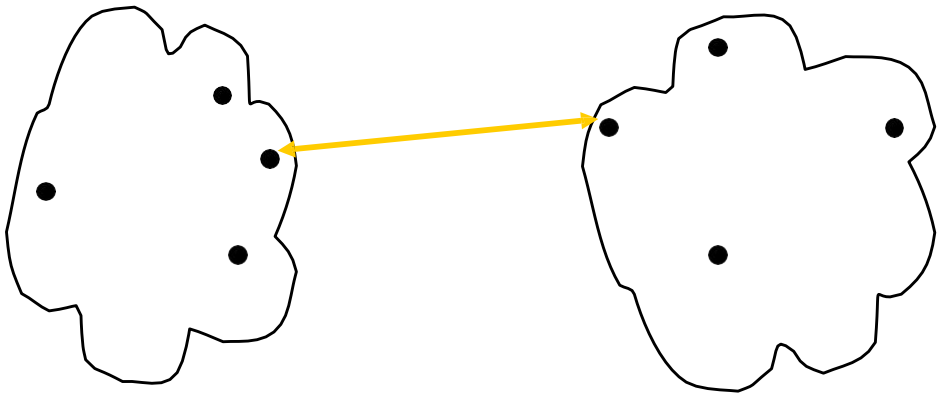


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Matriz de Proximidad

- MIN (o Single)
- MAX (o Complete)
- Promedio del grupo
- Distancia entre centroides
- Otro método basado en una función objetivo
  - Ejemplo: Método de Ward que utiliza el error cuadrado

Algoritmo de Clustering Aglomerativo
Similaridad entre clusters

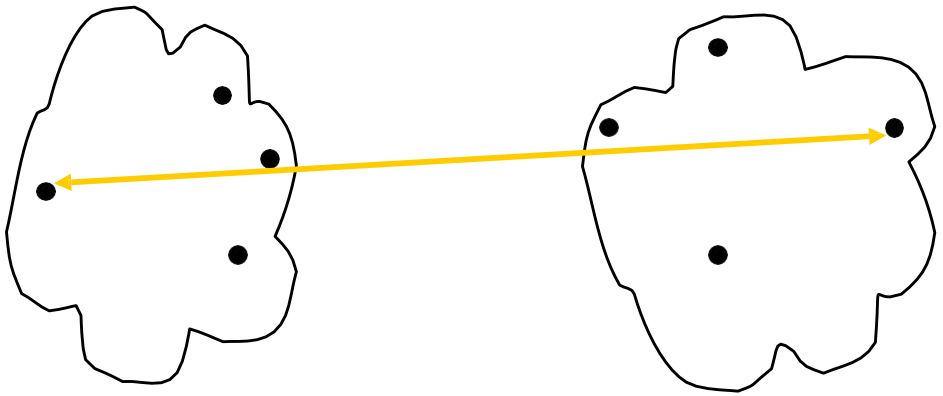


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de Proximidad

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Otro método basado en una función objetivo
  - Ejemplo: Método de Ward que utiliza el error cuadrado

Algoritmo de Clustering Aglomerativo
Similaridad entre clusters

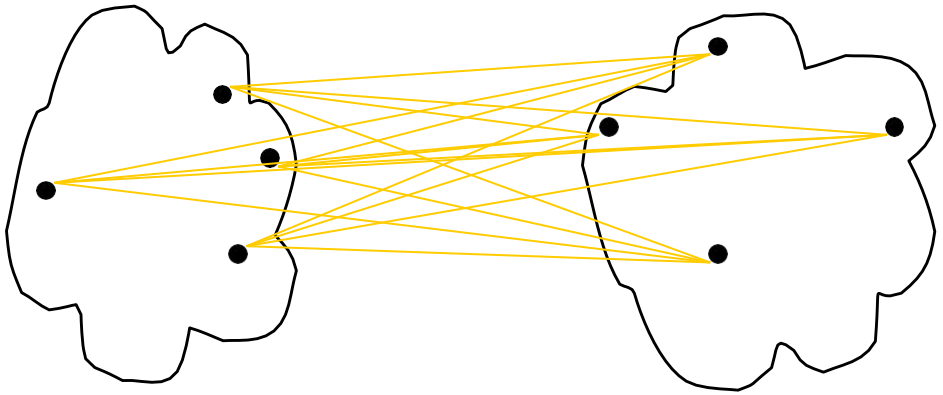


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Matriz de Proximidad

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Otro método basado en una función objetivo
  - Ejemplo: Método de Ward que utiliza el error cuadrado

Algoritmo de Clustering Aglomerativo
Similaridad entre clusters

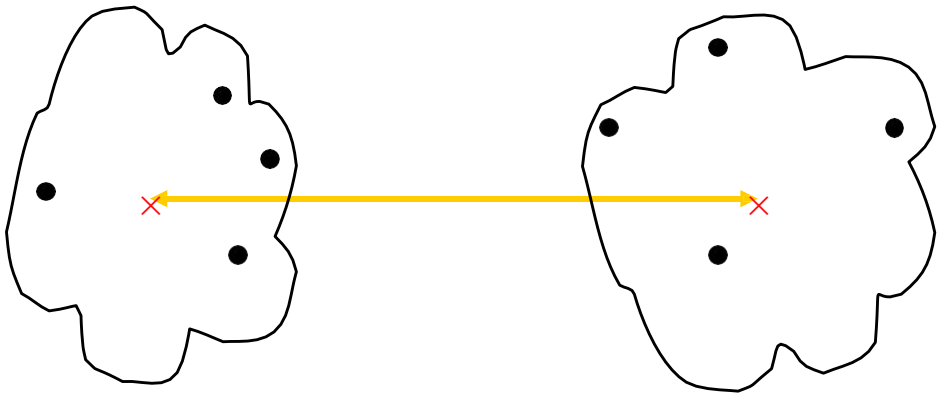


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de Proximidad

- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Otro método basado en una función objetivo
  - Ejemplo: Método de Ward que utiliza el error cuadrado

Algoritmo de Clustering Aglomerativo
Similaridad entre clusters

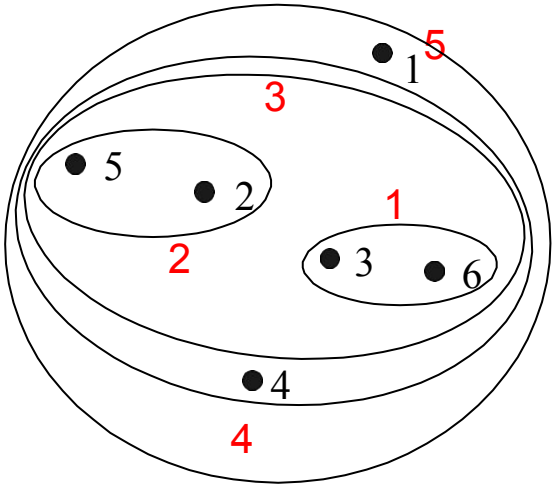


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de Proximidad

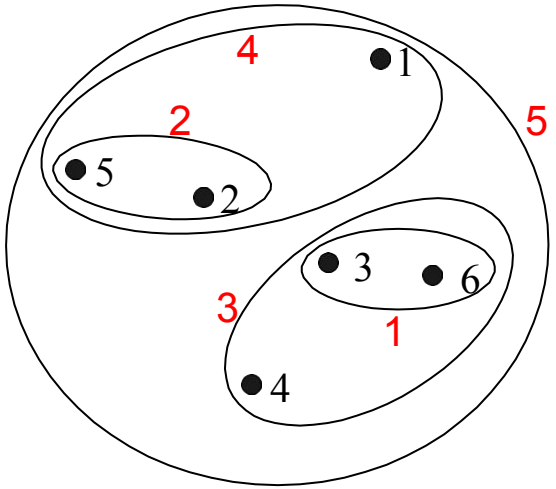
- MIN
- MAX
- Promedio del grupo
- Distancia entre centroides
- Otro método basado en una función objetivo
  - Ejemplo: Método de Ward que utiliza el error cuadrado

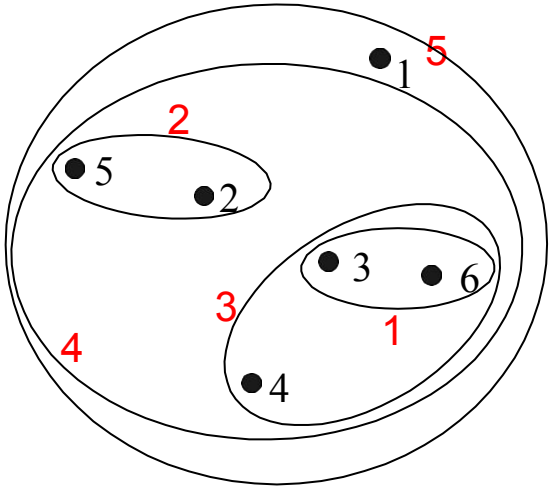
Clustering Jerárquico – Comparación



MIN

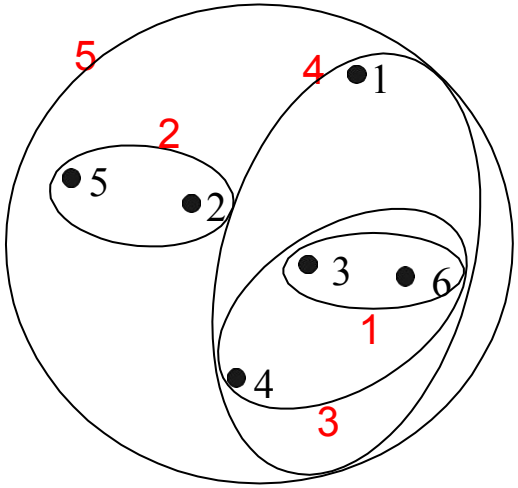
MAX





Promedio
del grupo

Ward's Method



Ejemplo de datos a agrupar

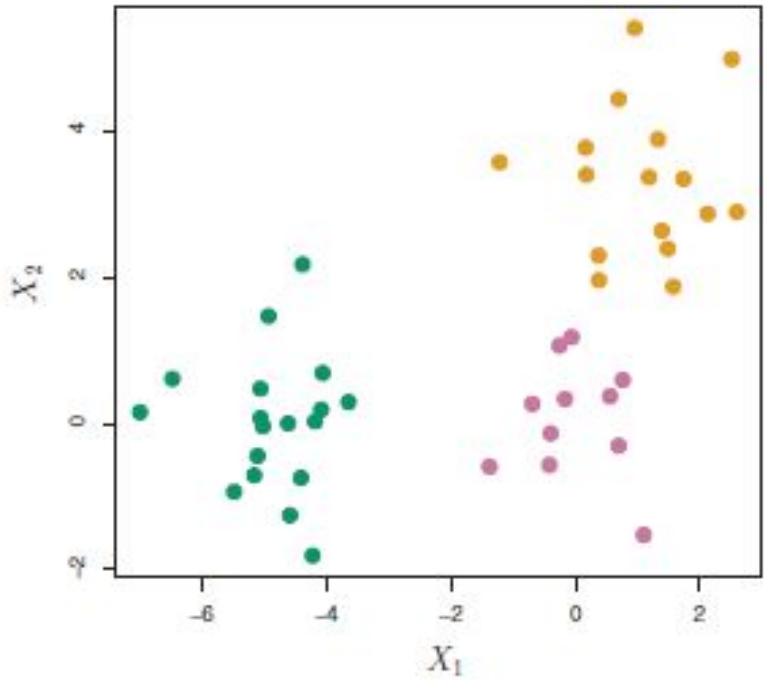


FIGURE 10.8. *Forty-five observations generated in two-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.*

Ejemplo de agrupamiento jerárquico

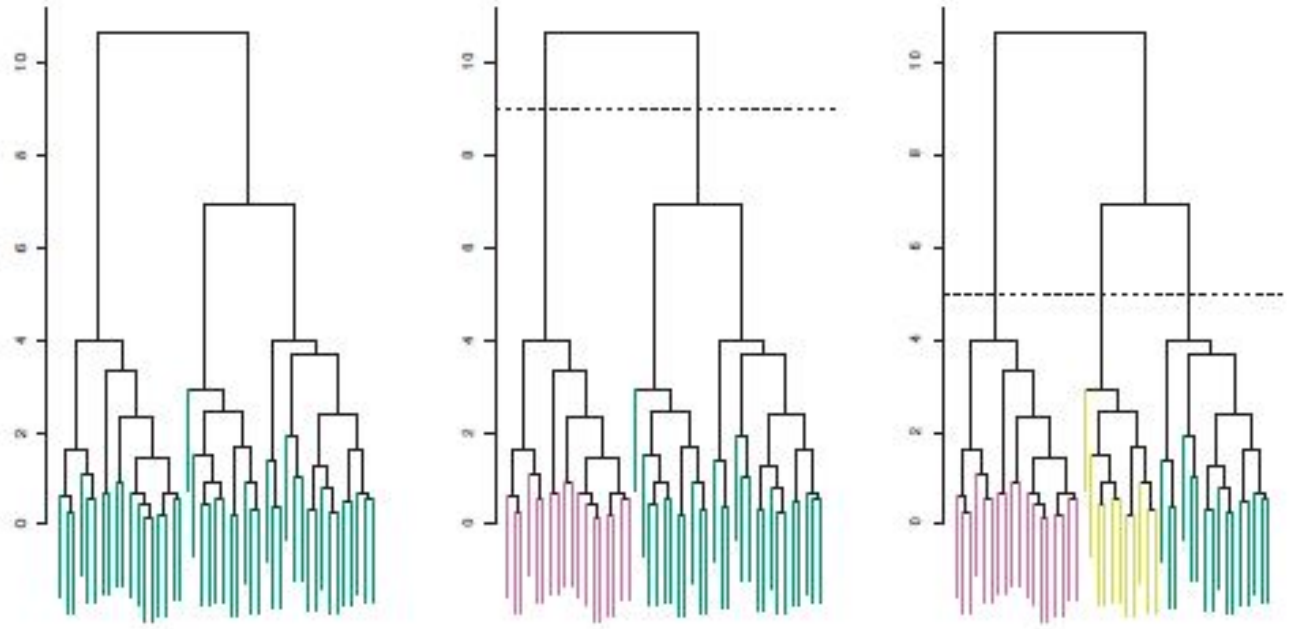


FIGURE 10.9. Left: *dendrogram obtained from hierarchically clustering the data from Figure 10.8 with complete linkage and Euclidean distance.* Center: *the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.* Right: *the dendrogram from the left-hand panel, now cut at a height of five. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure.*

Diferentes formas de calcular la distancia entre clusters

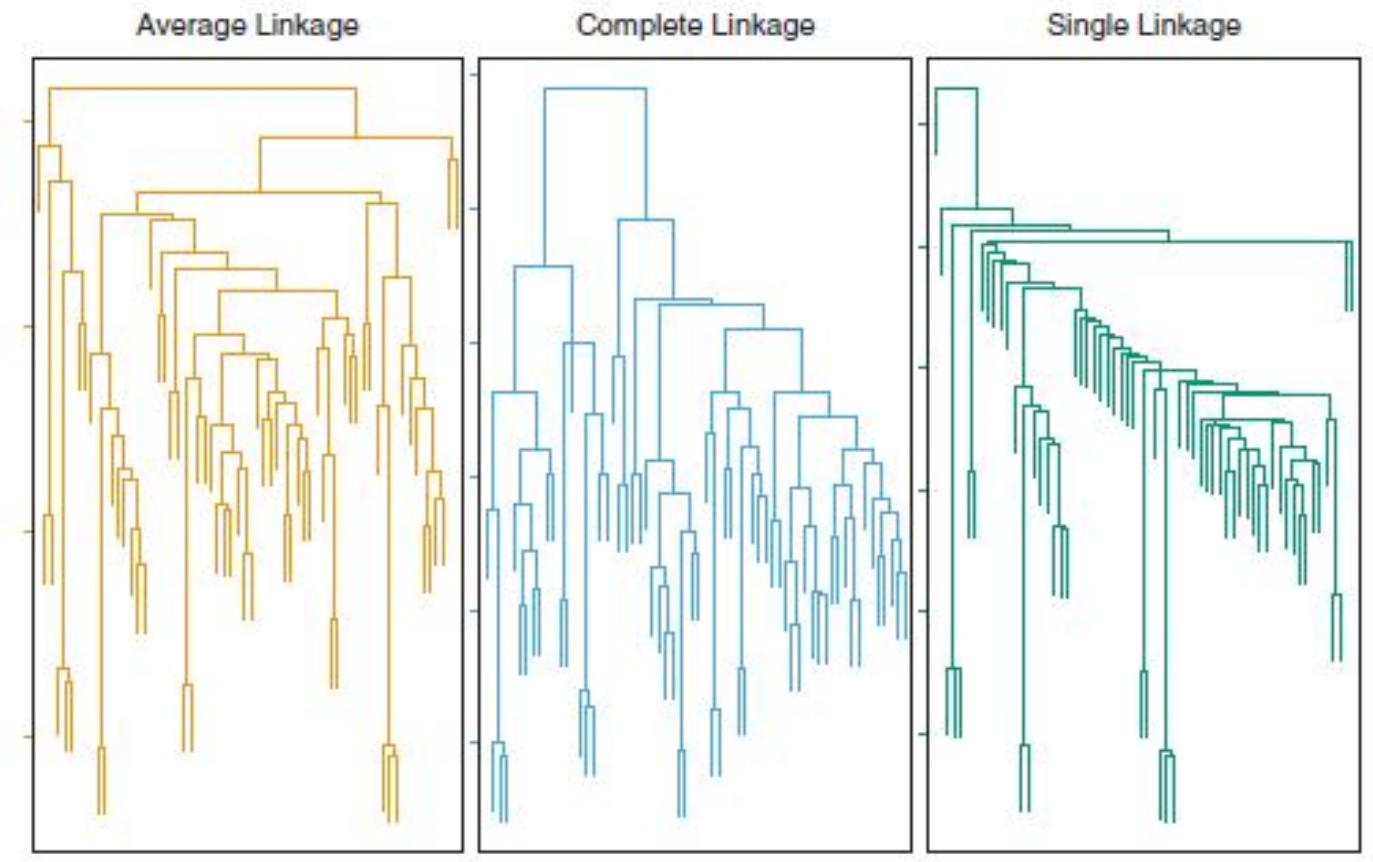


FIGURE 10.12. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

Diferentes medidas de distancia entre observaciones

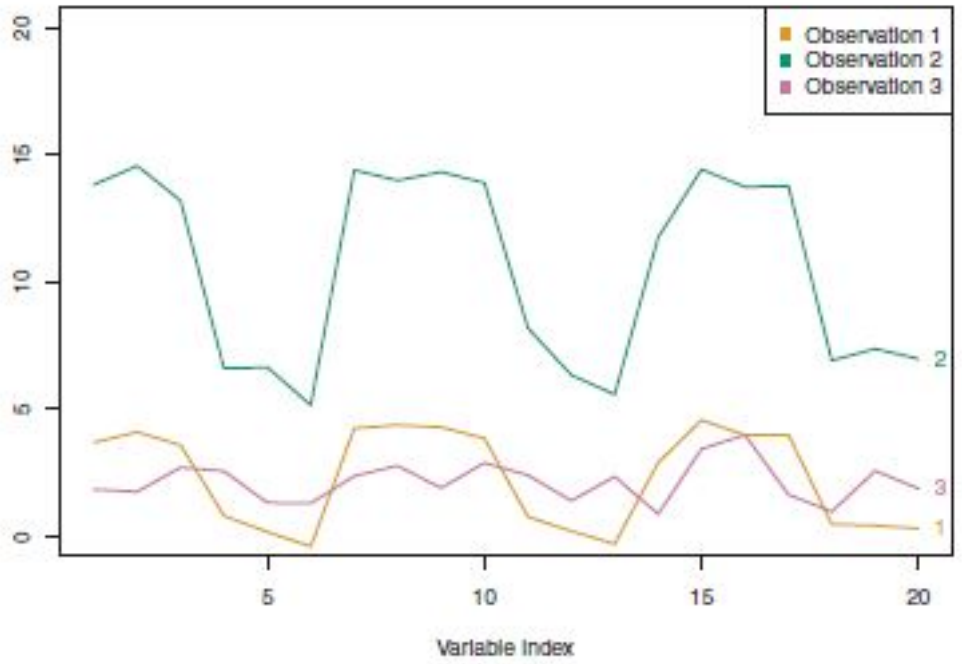


FIGURE 10.13. *Three observations with measurements on 20 variables are shown. Observations 1 and 3 have similar values for each variable and so there is a small Euclidean distance between them. But they are very weakly correlated, so they have a large correlation-based distance. On the other hand, observations 1 and 2 have quite different values for each variable, and so there is a large Euclidean distance between them. But they are highly correlated, so there is a small correlation-based distance between them.*

Efecto de escalar datos en el valor de las distancias

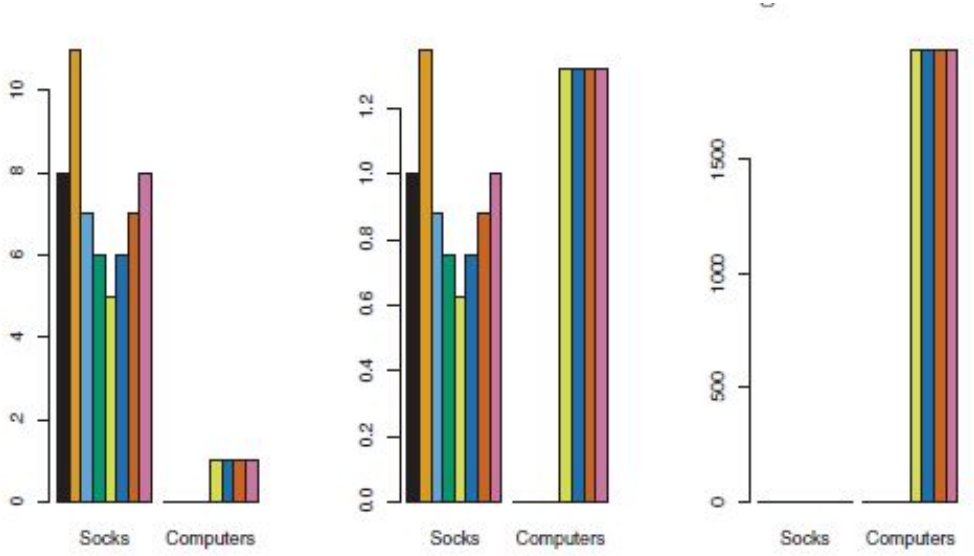


FIGURE 10.14. *An eclectic online retailer sells two items: socks and computers.* Left: *the number of pairs of socks, and computers, purchased by eight online shoppers is displayed. Each shopper is shown in a different color. If inter-observation dissimilarities are computed using Euclidean distance on the raw variables, then the number of socks purchased by an individual will drive the dissimilarities obtained, and the number of computers purchased will have little effect. This might be undesirable, since (1) computers are more expensive than socks and so the online retailer may be more interested in encouraging shoppers to buy computers than socks, and (2) a large difference in the number of socks purchased by two shoppers may be less informative about the shoppers' overall shopping preferences than a small difference in the number of computers purchased.* Center: *the same data is shown, after scaling each variable by its standard deviation. Now the number of computers purchased will have a much greater effect on the inter-observation dissimilarities obtained.* Right: *the same data are displayed, but now the y-axis represents the number of dollars spent by each online shopper on socks and on computers. Since computers are much more expensive than socks, now computer purchase history will drive the inter-observation dissimilarities obtained.*

Clustering Jerárquico: Problemas y limitaciones

- Muy costoso en espacio ($O(N^2)$ para la matriz de proximidad) y tiempo ($O(N^3)$)
- Una vez combinados dos clusters no se puede deshacer
- No minimiza directamente una función objetivo
- Los diferentes esquemas tienen problemas con uno o más de los siguientes puntos:
  - Sensibilidad al ruido y outliers
  - Dificultad para manejar clusters de diferentes tamaños y formas convexas
  - Particiones de clusters grandes

Clustering No Jerárquico – k-medias (*K-means*)

- Cada cluster está asociado a un centroide (punto central)
- Cada punto es asignado al cluster del centroide más cercano
- Debe especificarse K, el número de clusters

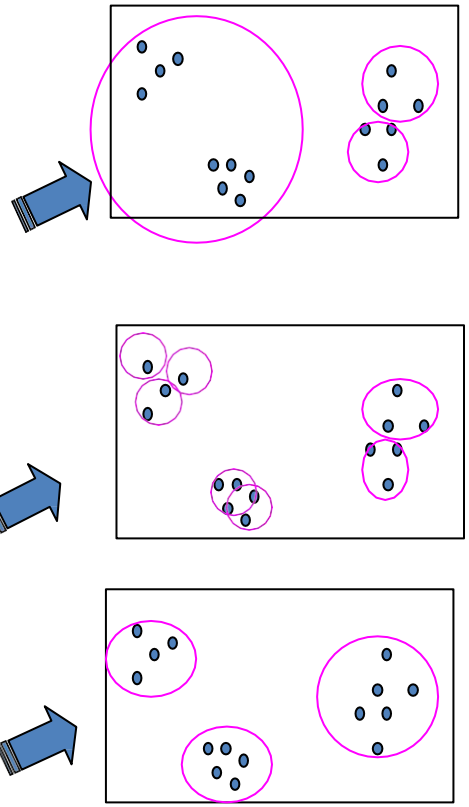
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-medias – Detalles

- Los centroides iniciales suelen elegirse al azar
  - Los resultados varían en las distintas corridas
- El centroide es típicamente la media de los puntos del cluster.
- La cercanía se mide por distancia Euclídea, de Manhattan, similitud de coseno, correlaciones, etc
- K-medias converge para las medidas de similitud mencionadas
- Generalmente converge en pocas iteraciones.
  - Frecuentemente la condición de parada se cambia por 'hasta que relativamente pocos puntos cambian de cluster'
- La complejidad es del O(n * K * I * d)
  - n = número de elementos, K = número de clusters, I = número de iteraciones, d = número de atributos

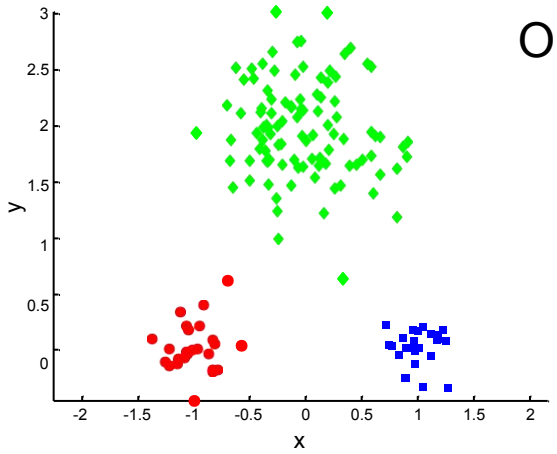
K-medias

- El valor de *k* se suele determinar heurísticamente.
- Problemas:
  - Si k es muy pequeño puede que haya grupos que queden sin centro.
    O incluso si se sabe que hay *n* clases, hacer *k=n*
    puede resultar en que algún grupo use dos centros
    y dos grupos separados tengan que compartir centro.

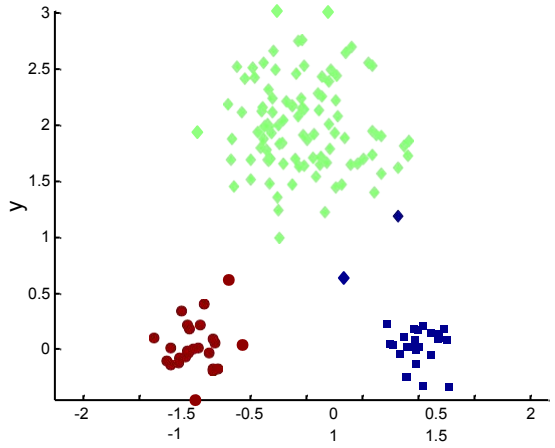


  - Si *k* se elige muy grande, la generalización es pobre y las agrupaciones futuras serán malas.

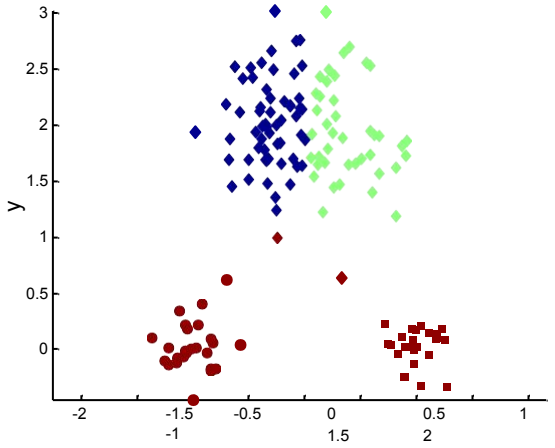
Diferentes agrupamiento con k-medias



Original Points

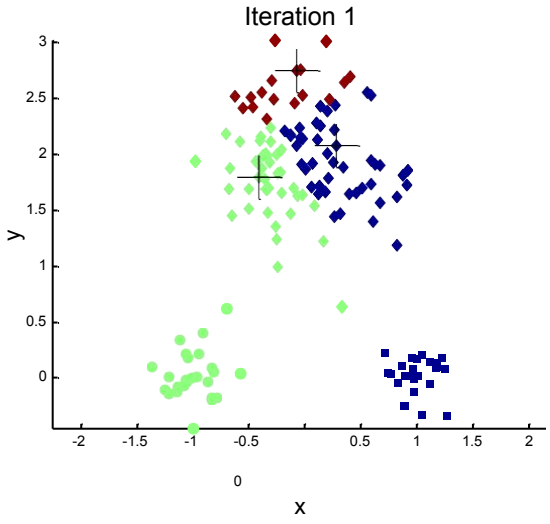


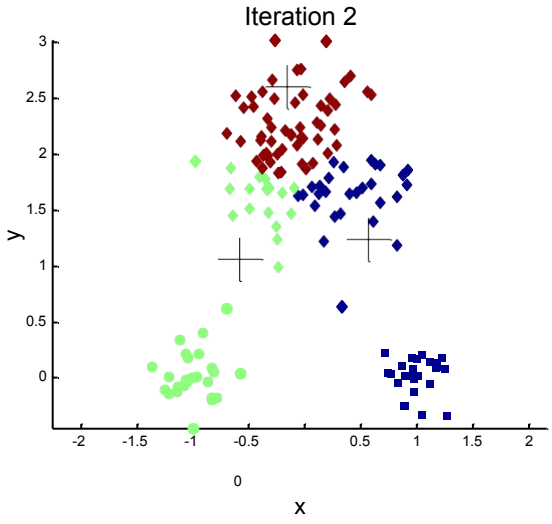
Optimal Clustering

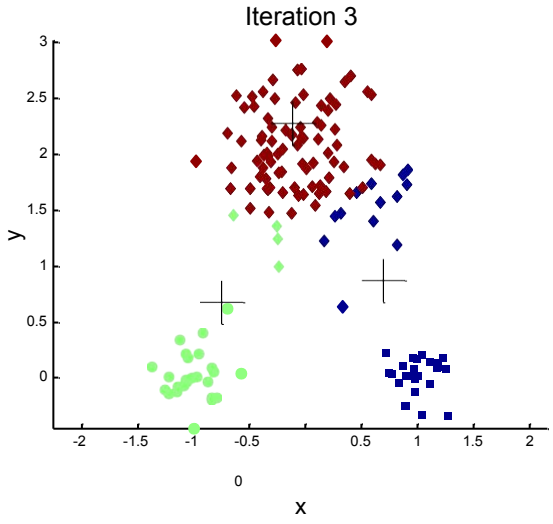


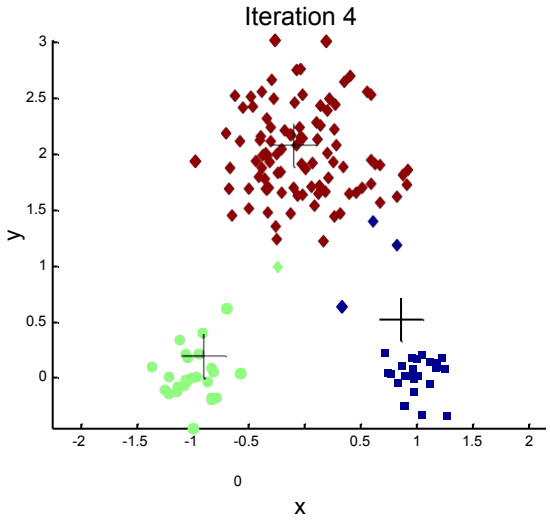
Sub-optimal Clustering

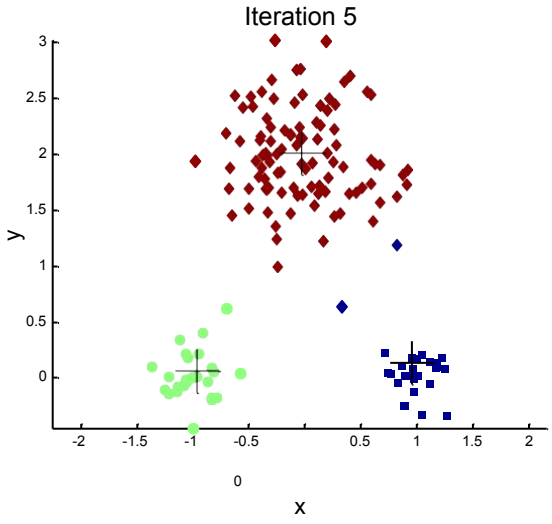
Importancia de la elección inicial de los centroides

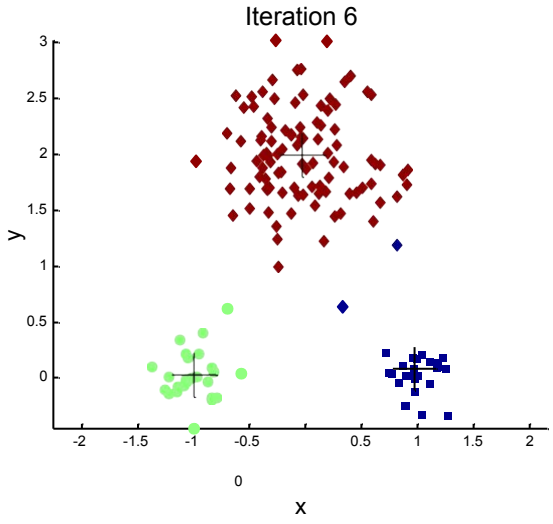




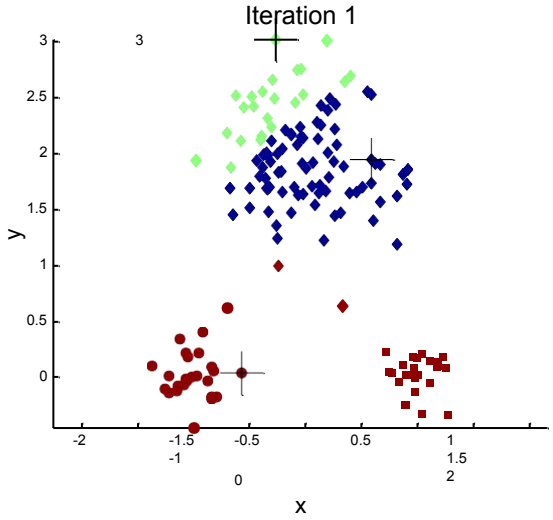


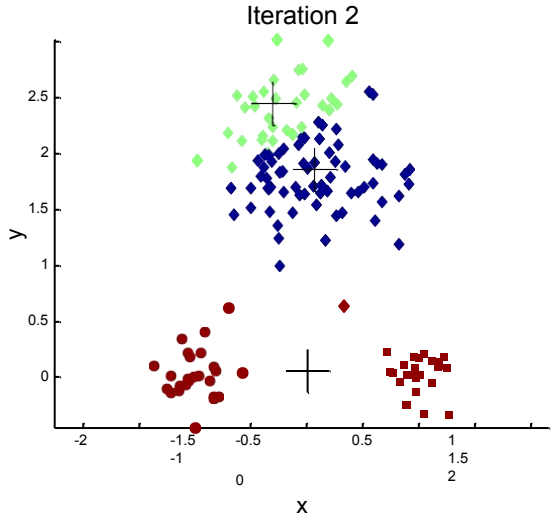


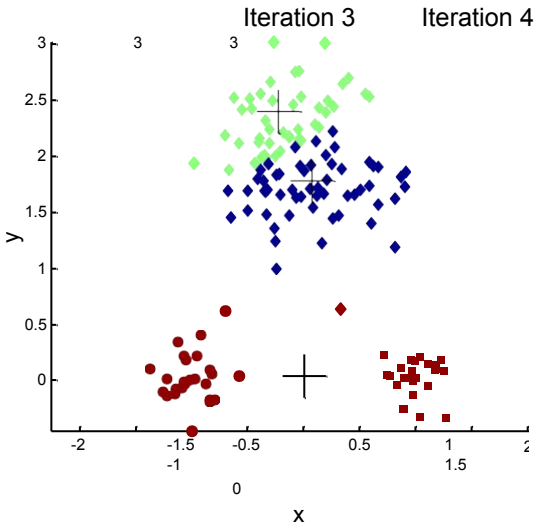


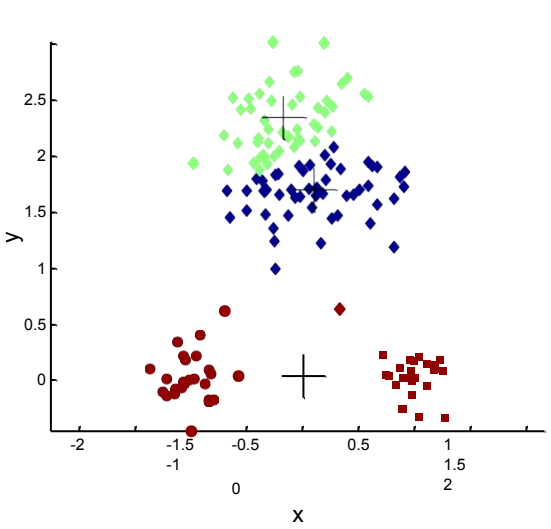


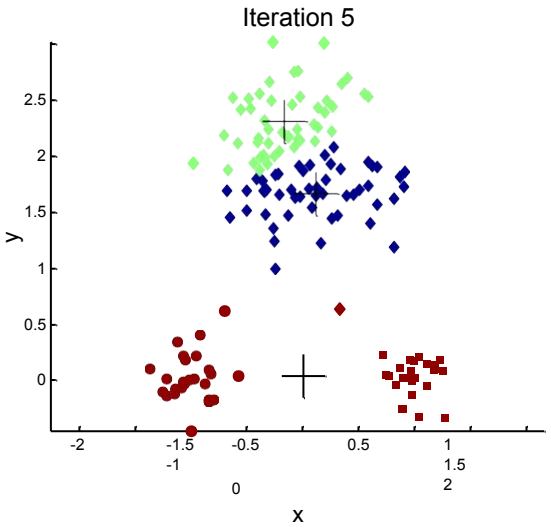
Importancia de la elección inicial de los centroides











Ejemplo de agrupamientos con diferentes valores de k

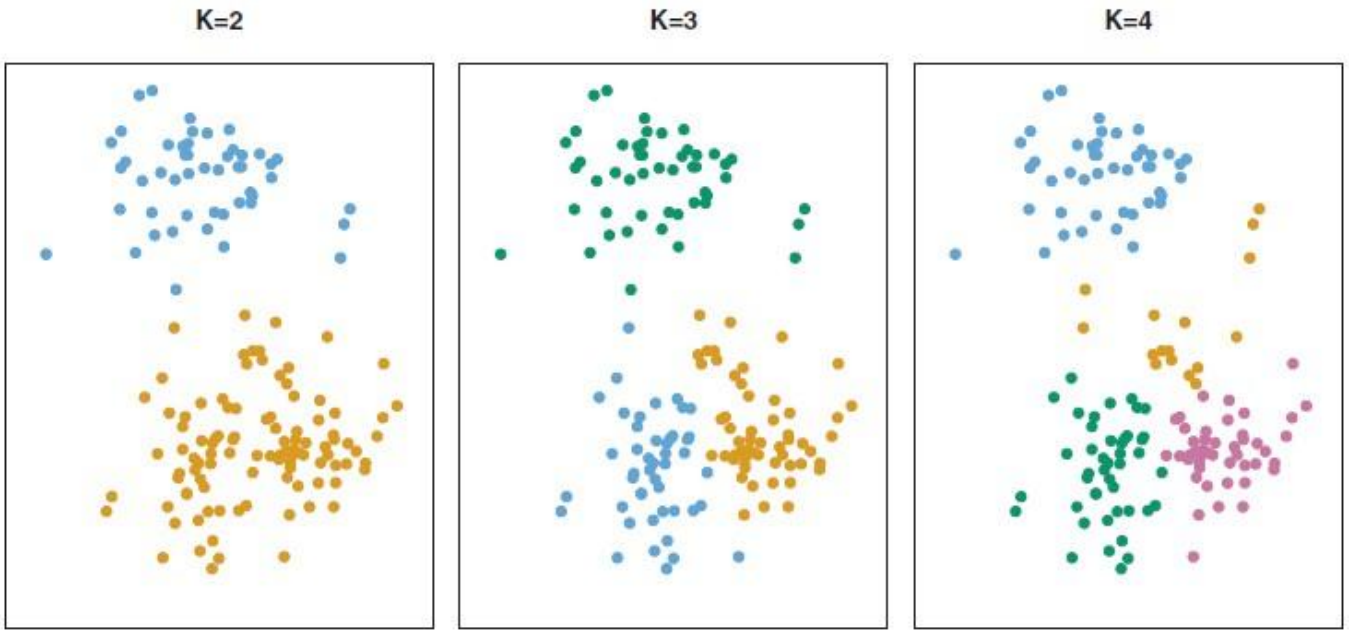


FIGURE 10.5. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.*

K medias en R

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Ejemplo del proceso de K-medias

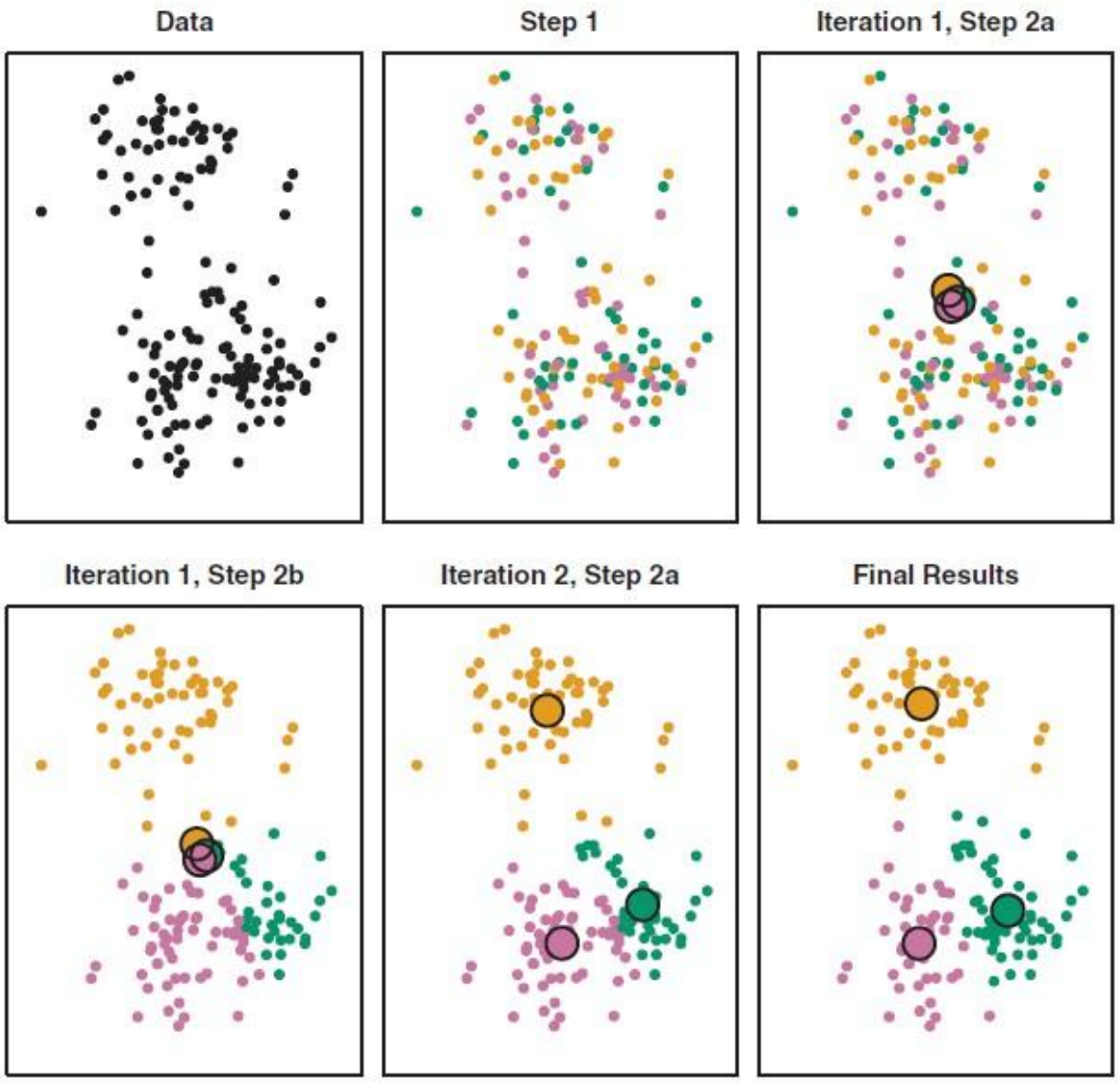


FIGURE 10.6. ***The progress of the K-means algorithm on the example of Figure*** *10.5 with K=3. Top left: the observations are shown. Top center: in Step 1 of the algorithm, each observation is randomly assigned to a cluster. Top right: in Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random. Bottom left: in Step 2(b), each observation is assigned to the nearest centroid. Bottom center: Step 2(a) is once again performed, leading to new cluster centroids. Bottom right: the results obtained after ten iterations.*

Distintas corridas, distintos resultados

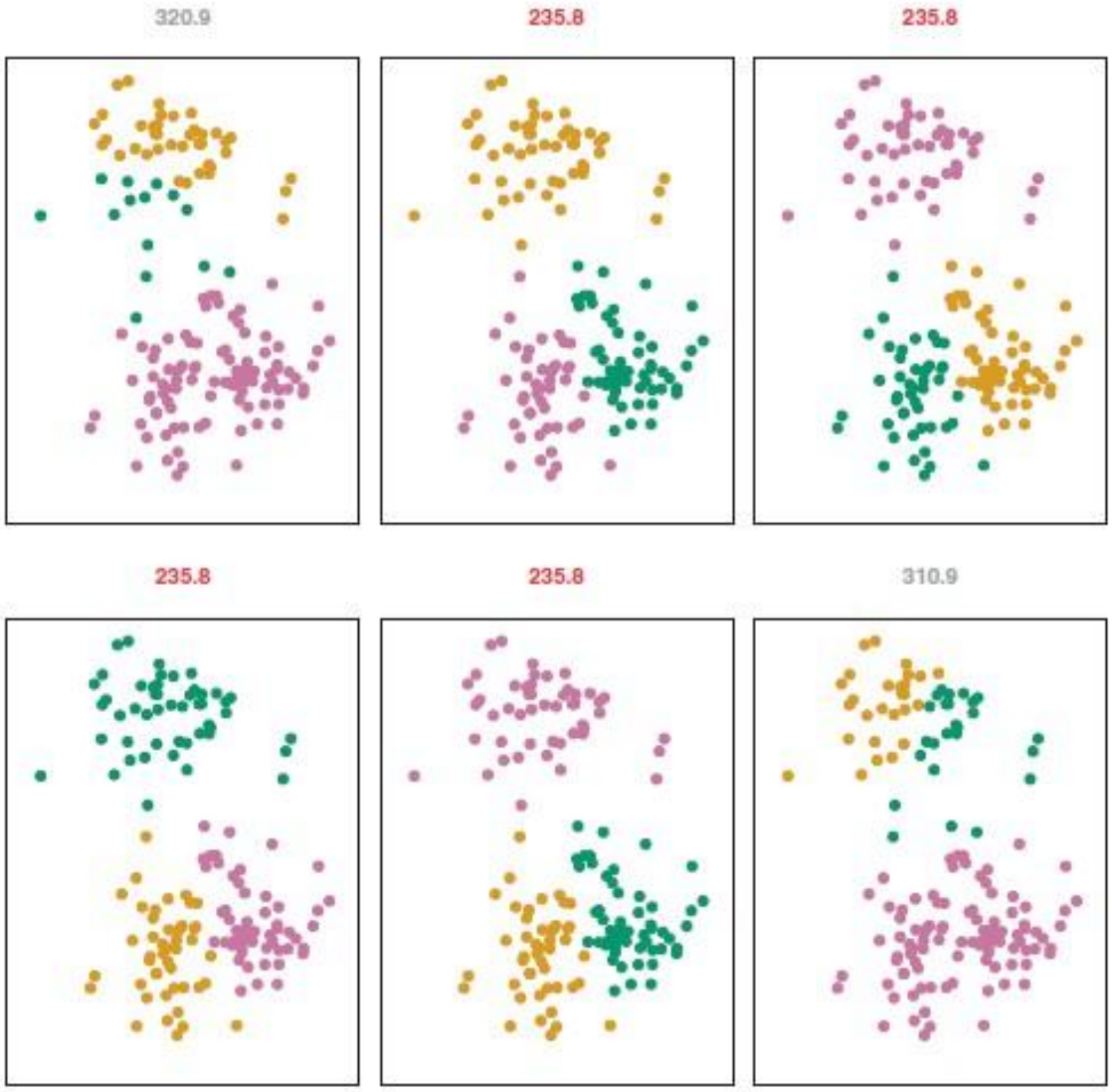


FIGURE 10.7. ***K-means clustering performed six times on the data from Figure*** *10.5 with K = 3, each time with a different random assignment of the observations in Step 1 of the K-means algorithm. Above each plot is the value of the objective (10.11). Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters. Those labeled in red all achieved the same best solution, with an objective value of 235.8.*

Evaluación de clusters con k-medias

- La medida más común es la suma del error cuadrado (*Sum of Squared Error - SSE*)
  - Para cada punto, el error es la distancia al centroide más cercano

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

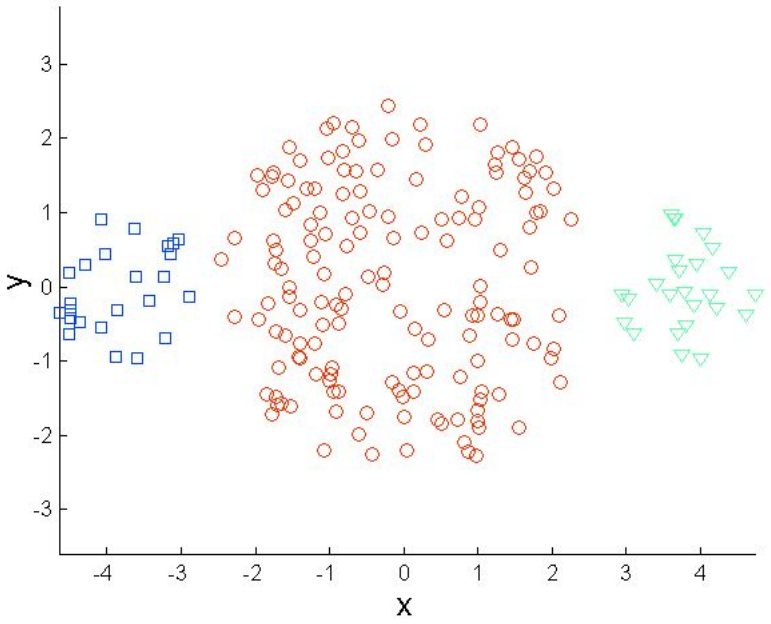
  - X es un elemento de datos en el cluster C_i y m_i es el punto representativo C_i (centroide)
  - Dados dos agrupamientos se puede elegir el que tenga error más pequeño.
  - Una forma sencilla de reducir el SSE es incrementar K

Limitaciones de K-medias

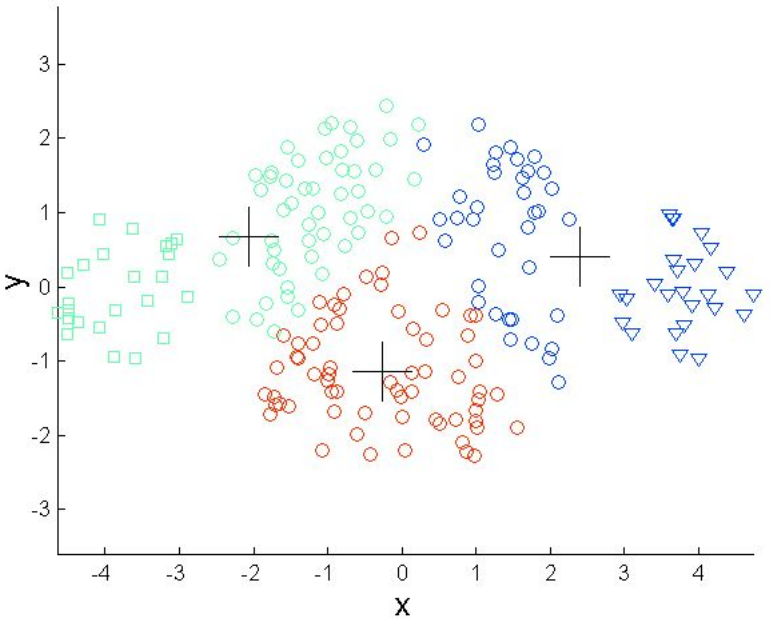
- K-medias tiene problemas cuando los clusters difieren en:
  - Tamaño
  - Densidad
  - Figuras no redondeadas

- K-medias tiene problemas con outliers.

Limitaciones de K-medias – Diferentes tamaños

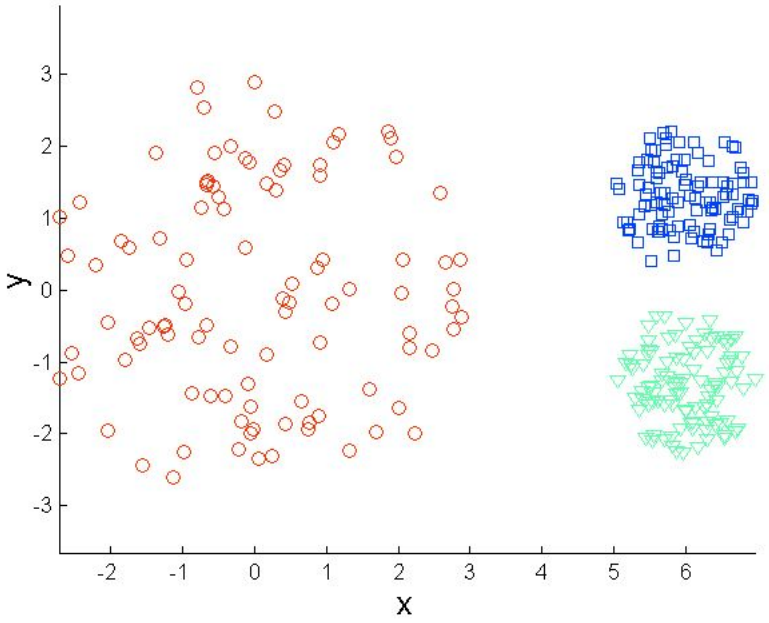


Original Points

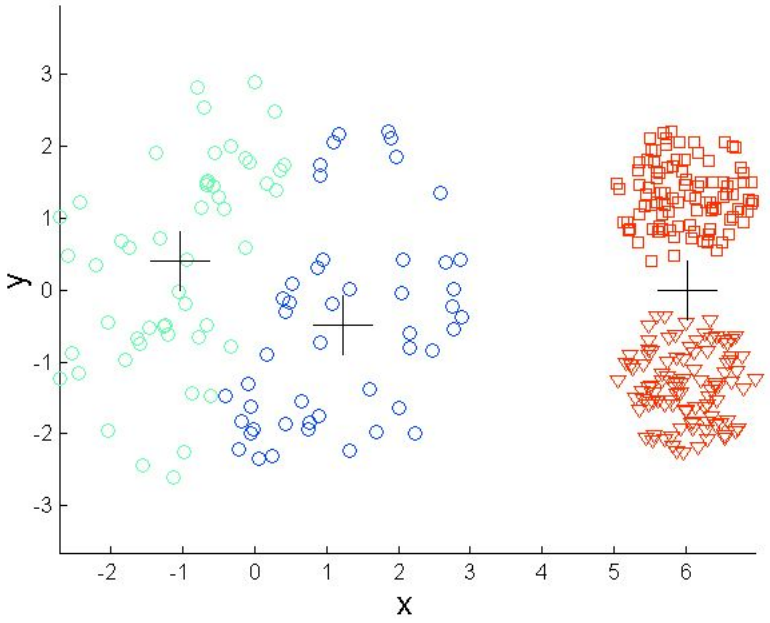


K-means (3 Clusters)

Limitaciones de K-medias – Diferente densidad

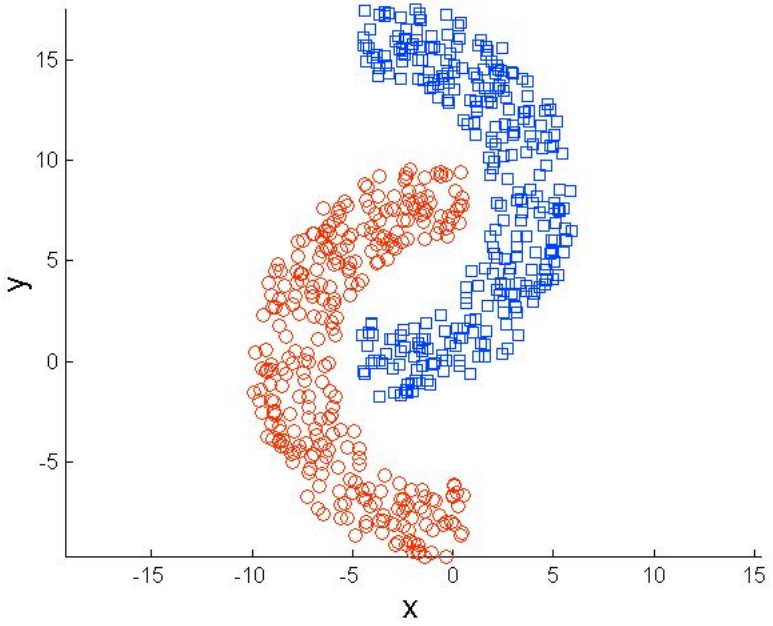


Original Points

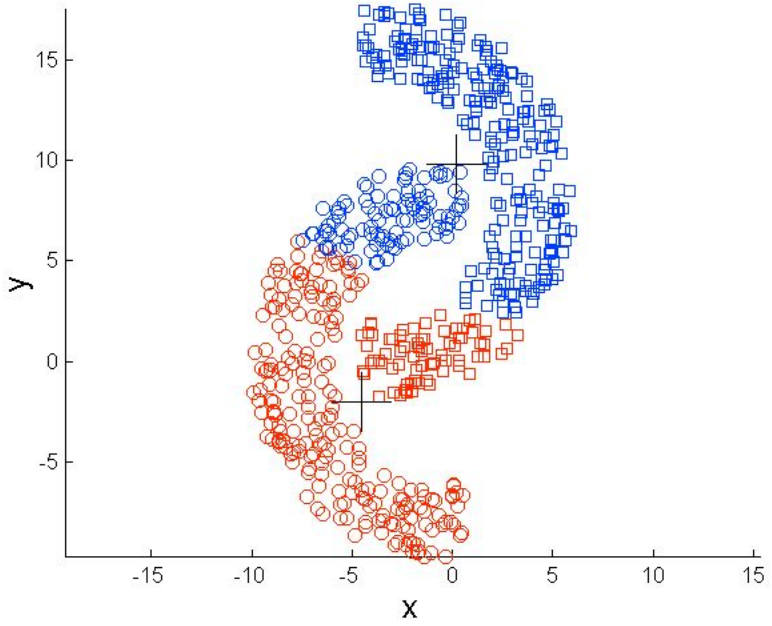


K-means (3 Clusters)

Limitaciones de K-medias – Formas no redondeadas



Original Points



K-means (2 Clusters)

Validación de los agrupamientos

- De la misma forma que se tienen medidas para evaluar que tan bueno es un modelo de clasificación supervisado se quiere evaluar que tan bueno son los agrupamientos resultantes
- “los clusters están en el ojo del observador”
- Para qué evaluarlos?
  - Para evitar ruido en los patrones
  - Para comparar resultados con distintos algoritmos de clustering (distintos métodos y distancias)
  - Para comparar dos conjuntos de clusters para ver cuál agrupamiento resulta mejor
  - Para comparar dos clusters
  - Para ver los grupos que se forman naturalmente. Y si se corresponden con algún agrupamiento externo
  - Para determinar el número de grupos adecuado (k)

Medidas para validación de agrupamientos

- Las medidas numéricas se clasifican en:
  - Externas: Usadas para medir la correspondencia entre el alcance del grupo y las etiquetas de clases provistas externamente.
    - Entropía
  - Internas: Usadas para medir la bondad de la estructura del cluster sin tener en cuenta información externa.
    - Sum of Squared Error (SSE)
  - Relativas: Usadas para comparar dos agrupamientos o dos grupos diferentes.
    - Generalmente se usa alguno de los índices internos o externos para esta

      función, ejemplo SSE o entropía

Medidas Internas: Cohesión y Separación

- Cohesión: Mide que tan estrechamente relacionados están los objetos dentro de un cluster
  - Cohesión se mide con la suma de cuadrados dentro (Within-cluster sum of squares)

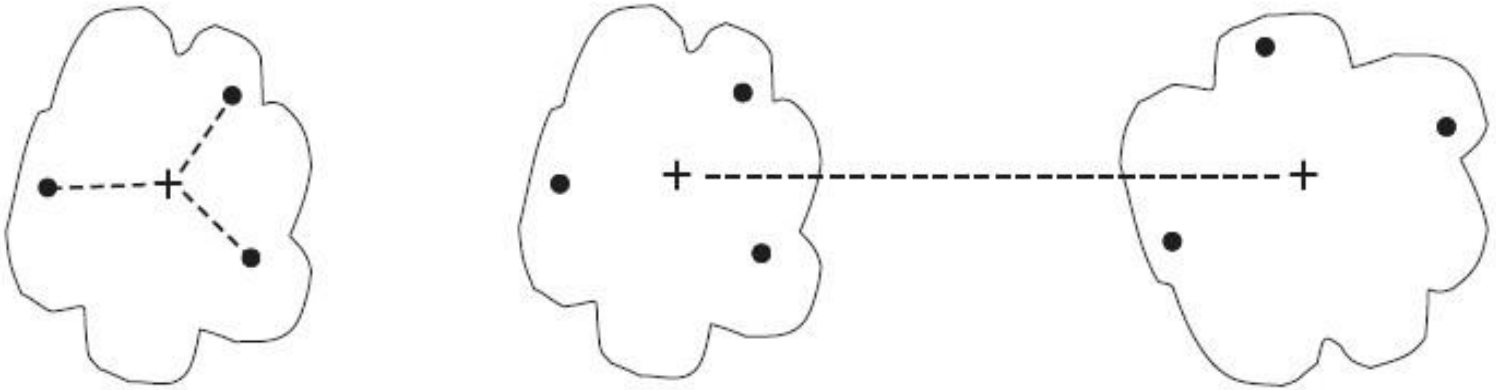
$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

- Separación: Mide que tan distintos son o bien separados está un cluster de los otros
  - Separación se mide a través de la suma de cuadrados entre clusters (Between- cluster Sum of Squared)

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

donde $|C_i|$ es el tamaño del cluster i , m_i es el centroide del cluster C_i, y m es el "centroide global"

Medidas Internas: Cohesión y Separación

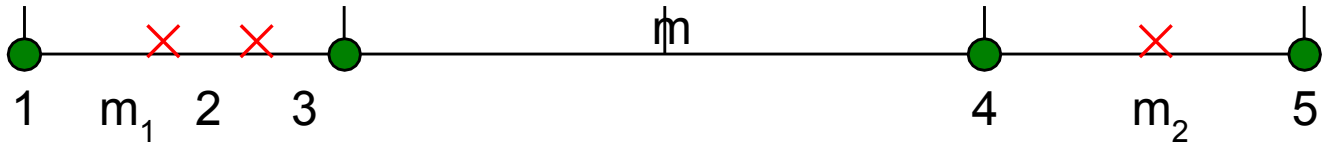


(a) Cohesion. (b) Separation.

Figure 8.28. Prototype-based view of cluster cohesion and separation.

Medidas Internas: Cohesión y Separación

- Ejemplo: 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:

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

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

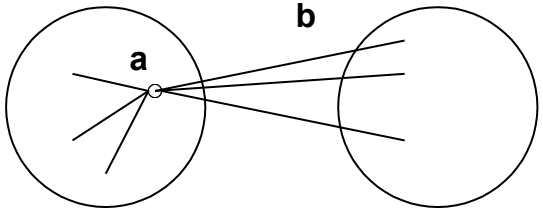
$$Total = 1 + 9 = 10$$

Medida interna: Coeficiente Silhouette

El coeficiente *silhouette* mide cuan buena es la asignación de un elemento o dato a su grupo. Combina las ideas de cohesión y separación. Para esto compara las distancias de este elemento respecto a todos los demás elementos del grupo al que pertenece, contra las distancias respecto a los grupos vecinos. El coeficiente del elemento *i* se denota *s(i)*.

Sea i un elemento perteneciente al grupo A, y C cualquier otro grupo distinto de A, ent.

a (i) = promedio de distancia de *i* a todos los otros objetos de *A*
d (i, C) = promedio de distancia de *i* a todos los objetos de C
b(i) = min *d (i, C) , C distinto de A*



$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

$$-1 \leq s(i) \leq 1$$

si *s(i)* ≈ −1, el dato i esta mal agrupado
si *s(i)* ≈ 0, el dato i está entre dos clusters
si *s(i)* ≈ 1, el dato i esta bien agrupado

Medida interna: Coeficiente Silhouette

El promedio de los *s* de los elementos dentro un cluster, da una idea de la calidad de ese cluster. El promedio de los *s* de todos los elementos dan una idea de que tan bien están agrupados todos los datos; si el clustering realizado es bueno o no.

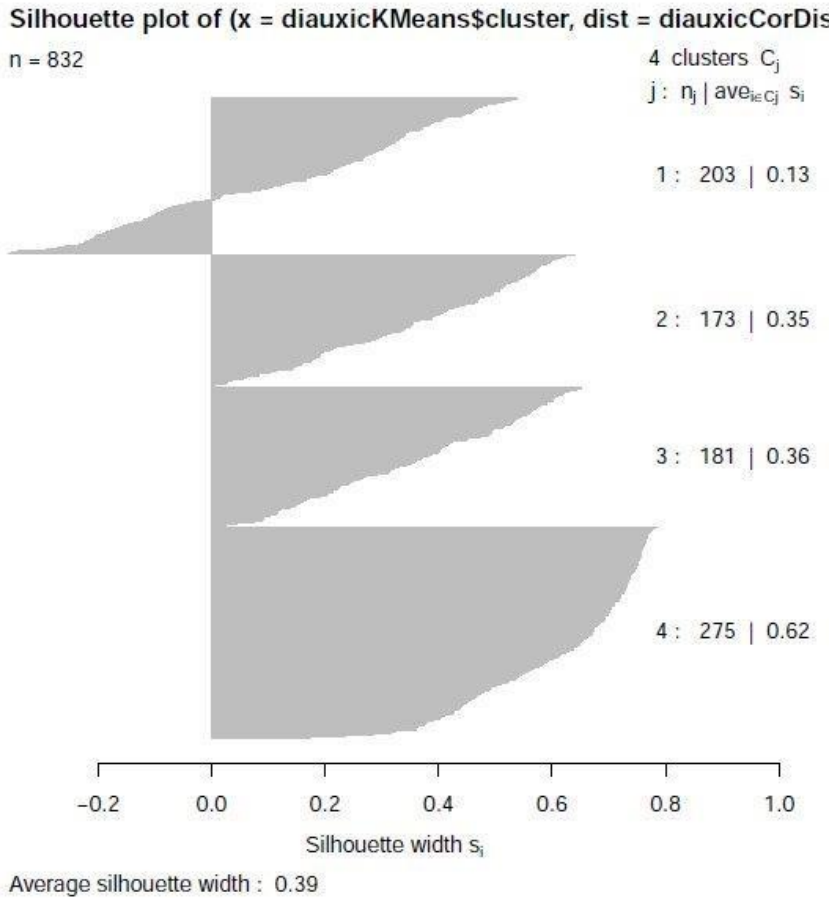
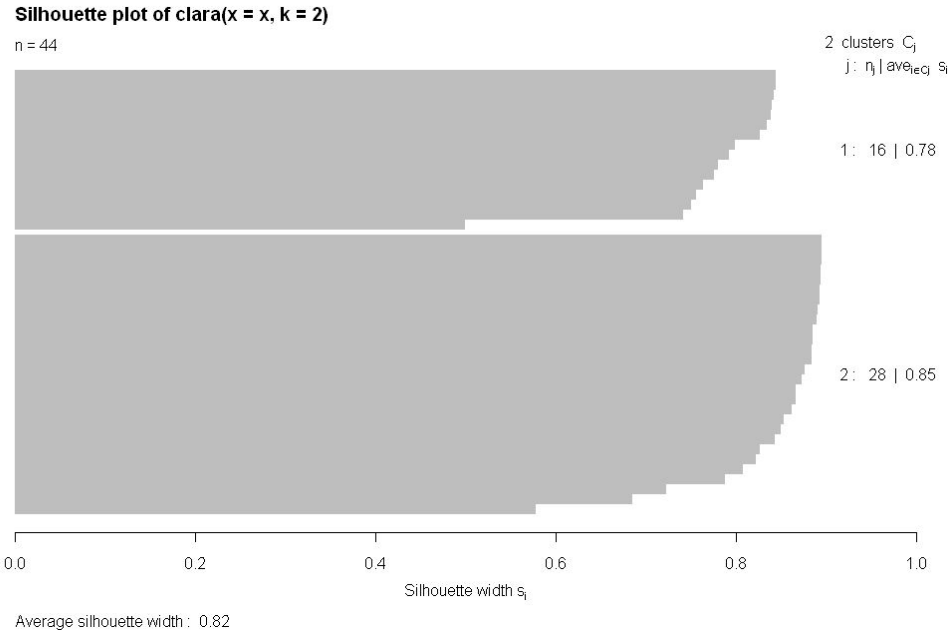
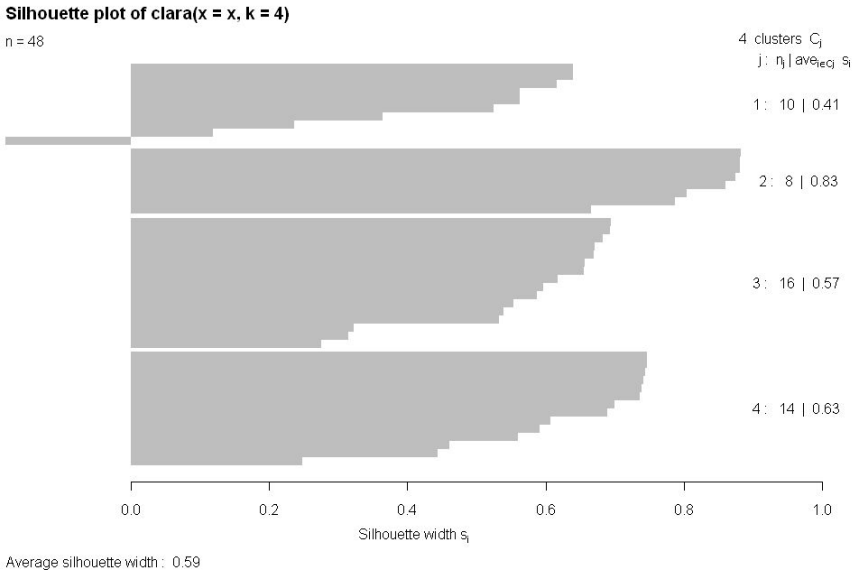


Tabla de interpretación de valores SC	
Rango de SC	**Interpretación**
0.71-1.0	Se ha encontrado una estructura fuerte
0.51-0.70	Se ha encontrado una estructura razonable
0.26-0.50	La estructura es débil y puede ser artificial.
<=0.25	No se ha encontrado ninguna estructura substancial

Medida interna: Coeficiente Silhouette

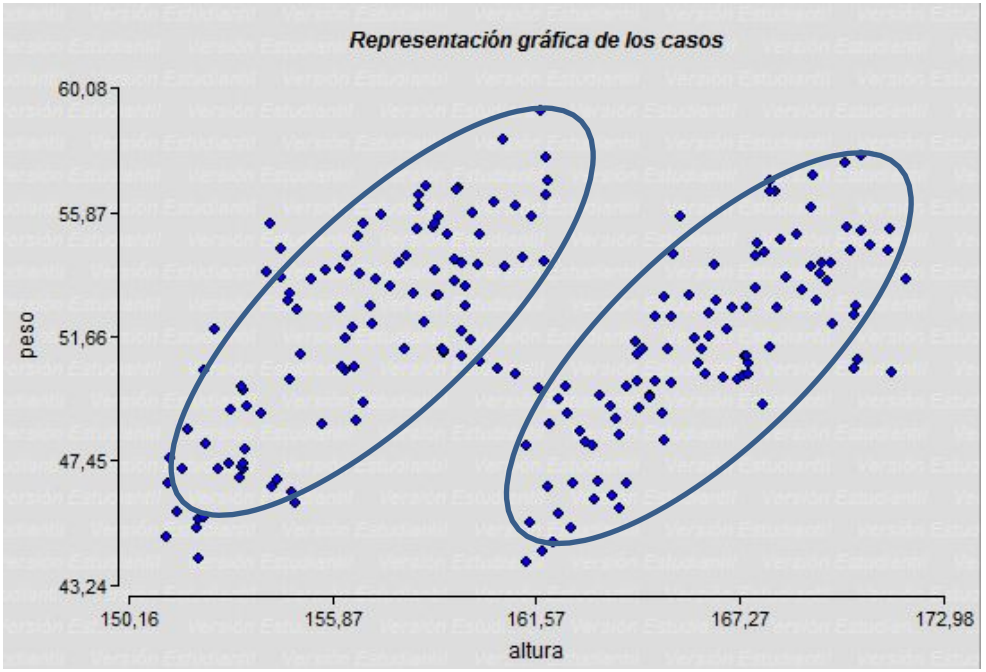




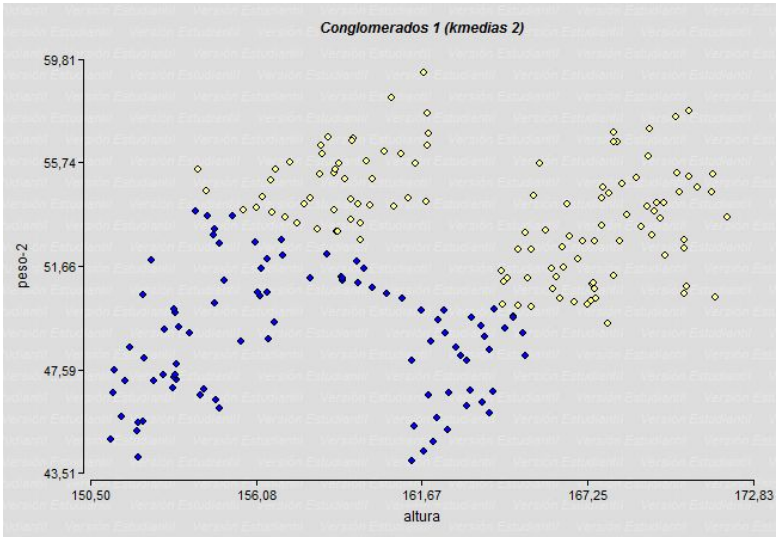
Comparación de agrupamientos: ¿k=2 o k=4?

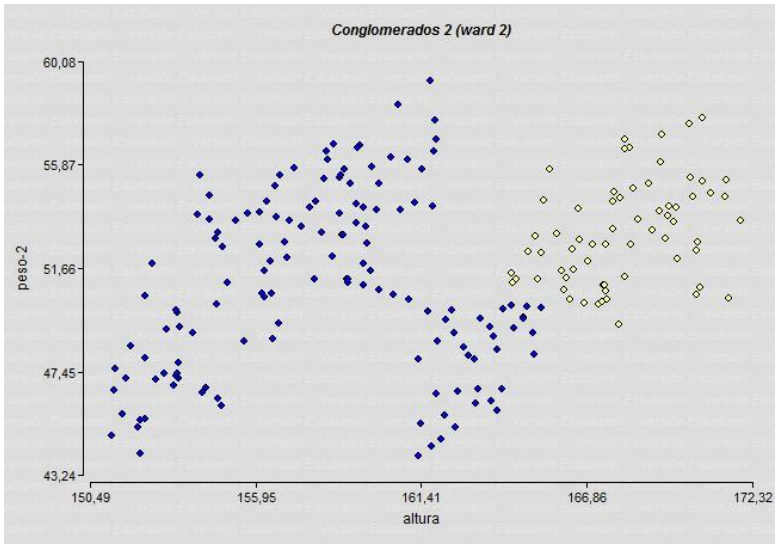
Clustering - Ejemplo

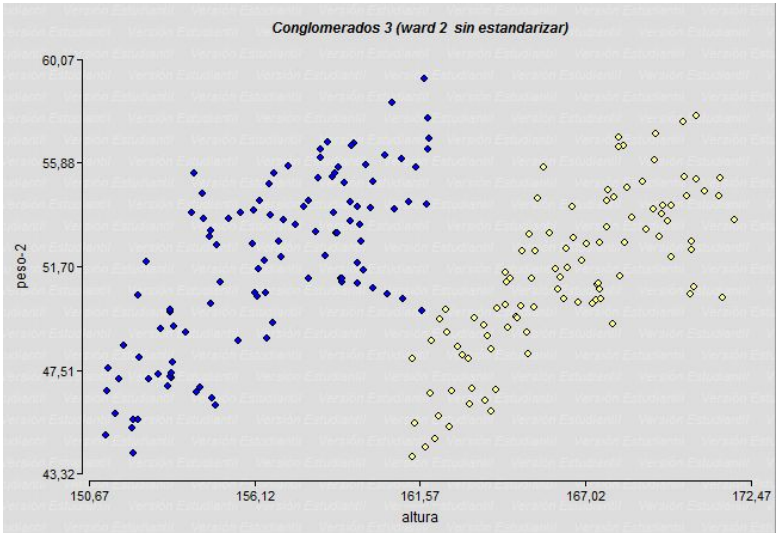
ID	altura (en cm)	peso (en kg)
1	151,26	46,68
2	152,32	48,01
3	154,71	46,38
4	154,12	55,47
5	156,09	53,95
6	156,59	55,03
7	158,3	56,10
8	158,96	51,22
9	160,01	55,11
10	160,64	58,32
…	…	…

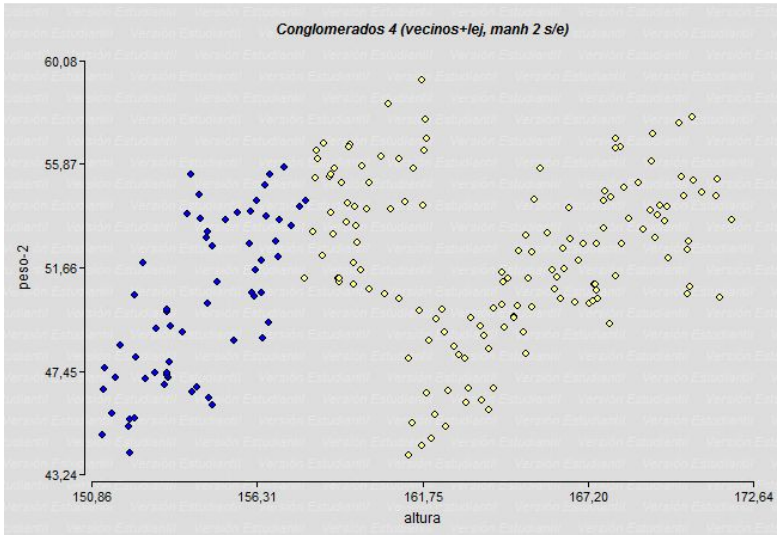


Ejemplo: Diferentes métodos y distancias

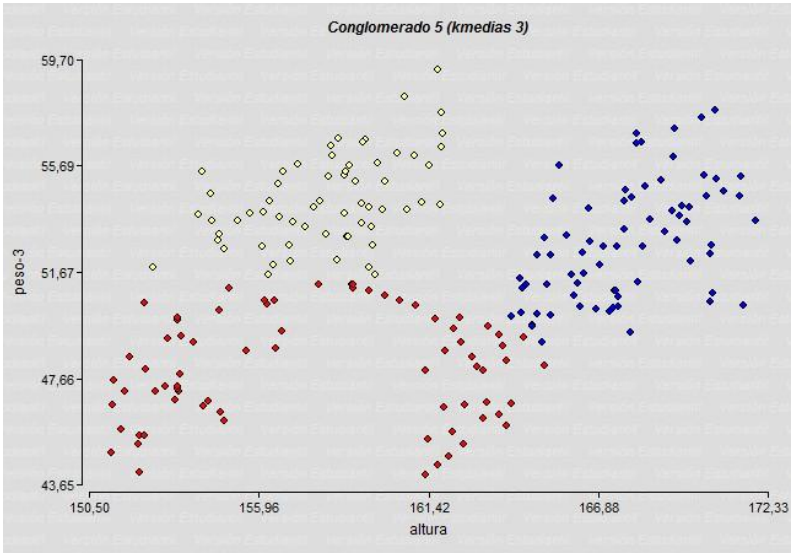


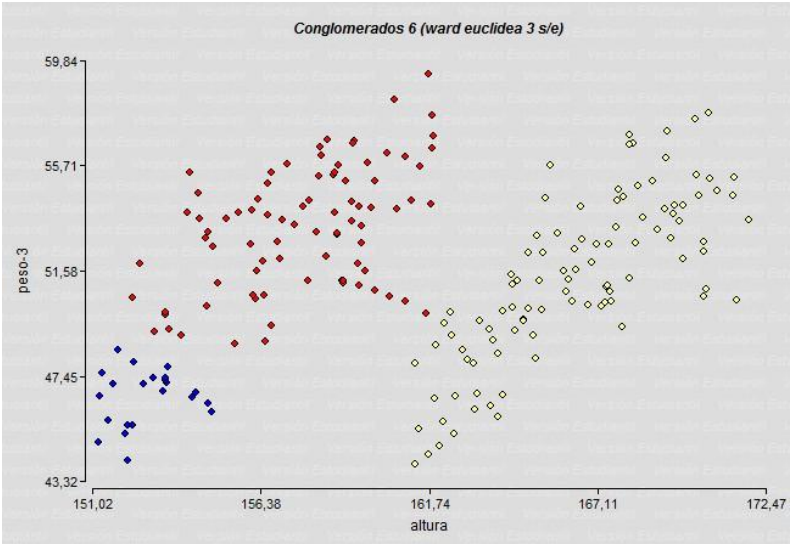


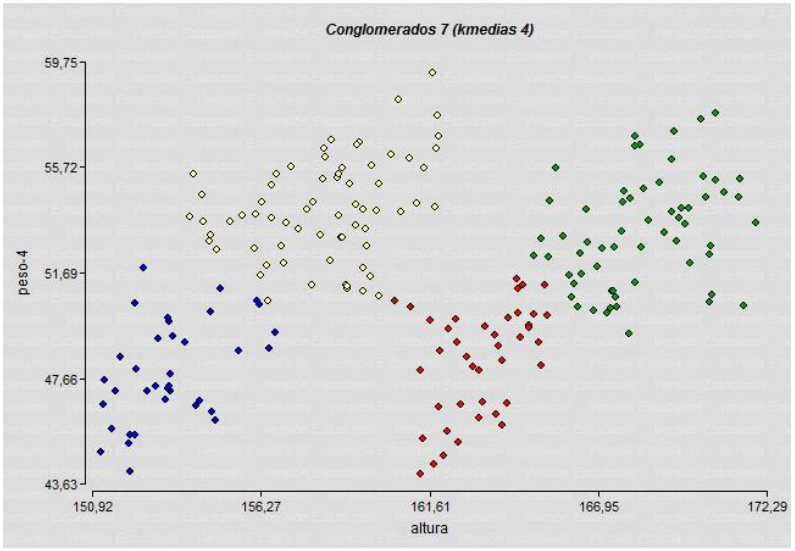


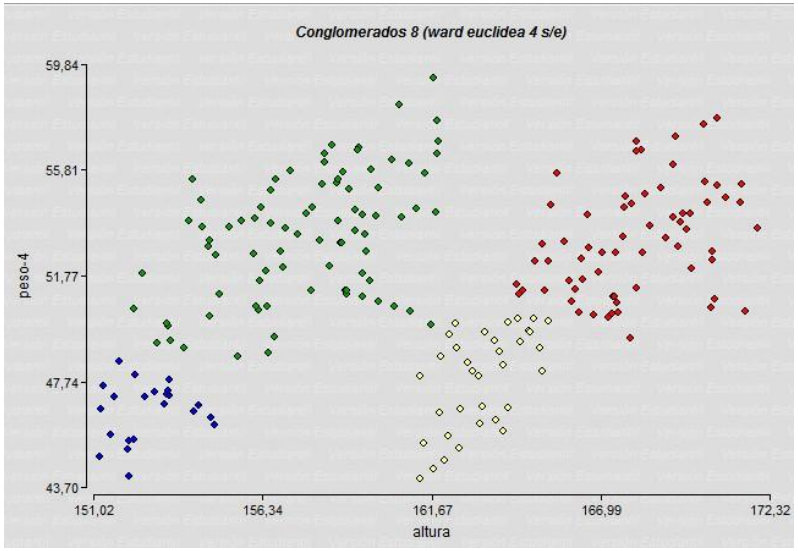


Ejemplo: Diferentes métodos y cantidades de grupos 3 o 4 grupos









Validación de clusters – Comentario final

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

Bibliografía y material utilizado

- Gareth James , Daniela Witten, Trevor Hastie, Robert Tibshirani (2013) “An Introduction to Statistical Learning – With Applications in R”, Ed.Springer