TDVI: Inteligencia Artificial

Clustering

UTDT - LTD

Estructura de la clase

- Motivación
- K-medias
- Clustering jerárquico

Estructura de la clase

- Motivación
- K-medias
- Clustering jerárquico

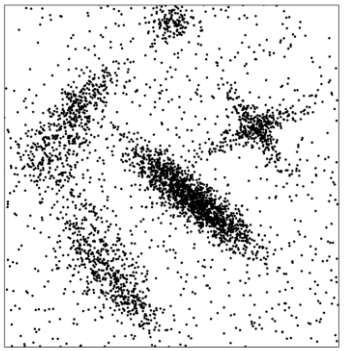
Motivación

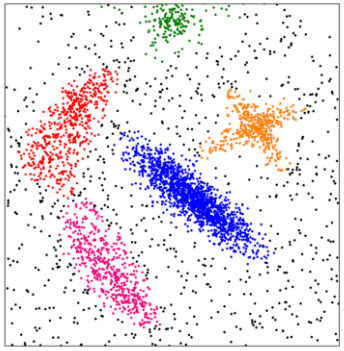
Clustering se refiere a una serie de técnicas cuyo objetivo es encontrar subgrupos o clusters en un conjunto de datos

La idea es particionar los datos de manera que:

- Observaciones que pertenecen a un grupo sean similares entre ellas
- Observaciones de grupos distintos sean distintas entre ellas

Esto plantea el "problema" de definir cuando dos observaciones son similares o distintas entre sí (algo que en gran medida puede depender del dominio en donde se esté trabajando.)





Motivación

Existen múltiples familias de algoritmos de clustering, por ejemplo:

- *Partitioning:* Divide los datos en grupos no superpuestos. (Ejemplos: K-Means, K-Medoids)
- *Hierarchical:* Crea una estructura jerárquica de clusters. (Ejemplos: AGNES, DIANA)
- *Density-based:* Agrupa puntos basándose en áreas de alta densidad. (Ejemplos: DBSCAN, OPTICS)
- *Grid-based:* Divide el espacio en una cuadrícula para formar clusters. (Ejemplos: STING, CLIQUE)
- *Model-based:* Asume un modelo estadístico para crear clusters. (Ejemplo: GMM)

Nosotros veremos los dos algoritmos más tradicionales:

- K-means clustering (*partitioning*)
- Clustering Jerárquico Aglomerativo (o AGNES, *Hierarchical*)

Estructura de la clase

- Motivación
- K-medias
- Clustering jerárquico

K-medias

El algoritmo de k-medias divide al conjunto de datos en *K* subconjuntos distintos sin solapamiento. Uno debe fijar el valor de *K* antes de correr el algoritmo de *K* medias

Los clusters deben cumplir las siguientes condiciones:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

K-medias

K-means asume que una buena asignación es aquella que dado un valor de *K* minimiza lo más posible la variabilidad intra cluster (*within-cluster variation*).

Si $W(C_j)$ es una medida que indica cuánto las observaciones de un cluster *j* difieren entre sí. El problema se puede escribir como:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Si uno usa la distancia euclídea (al cuadrado) como medida de disimilaridad $W(C_j)$ puede reescribirse de la siguiente manera:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

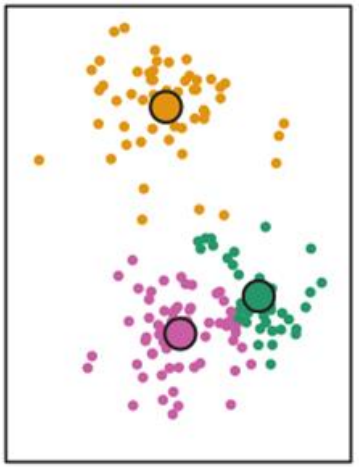
K-medias

De esta forma el problema se puede reescribir como:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

En donde $W(C_j)$ se puede expresar como la distancia a un centroide:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$



K-medias

Algoritmo de K-medias:

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).

K-medias

Algoritmo de K-medias:

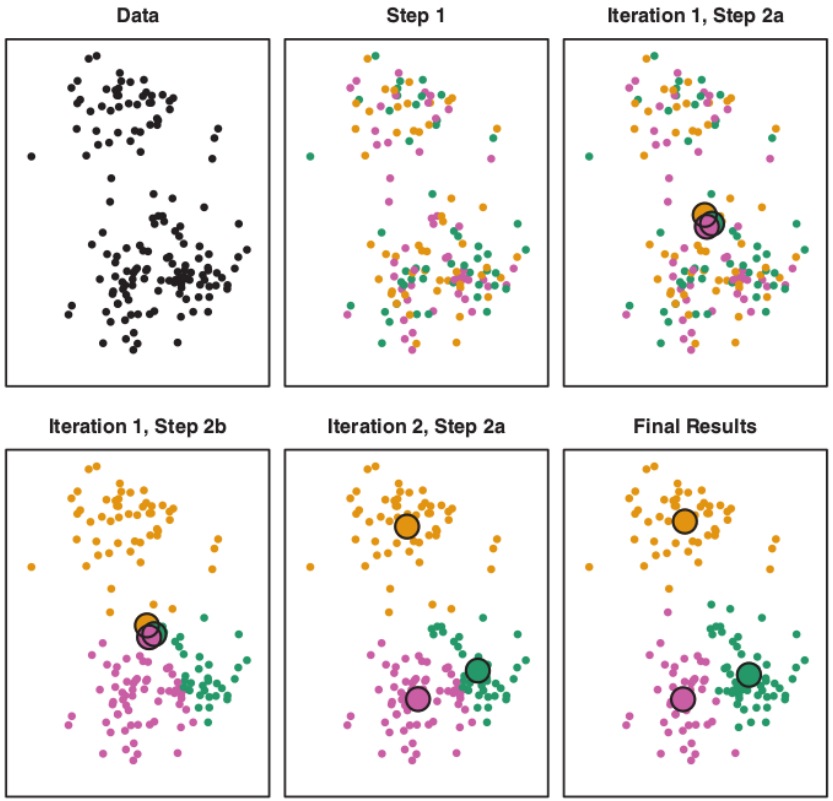
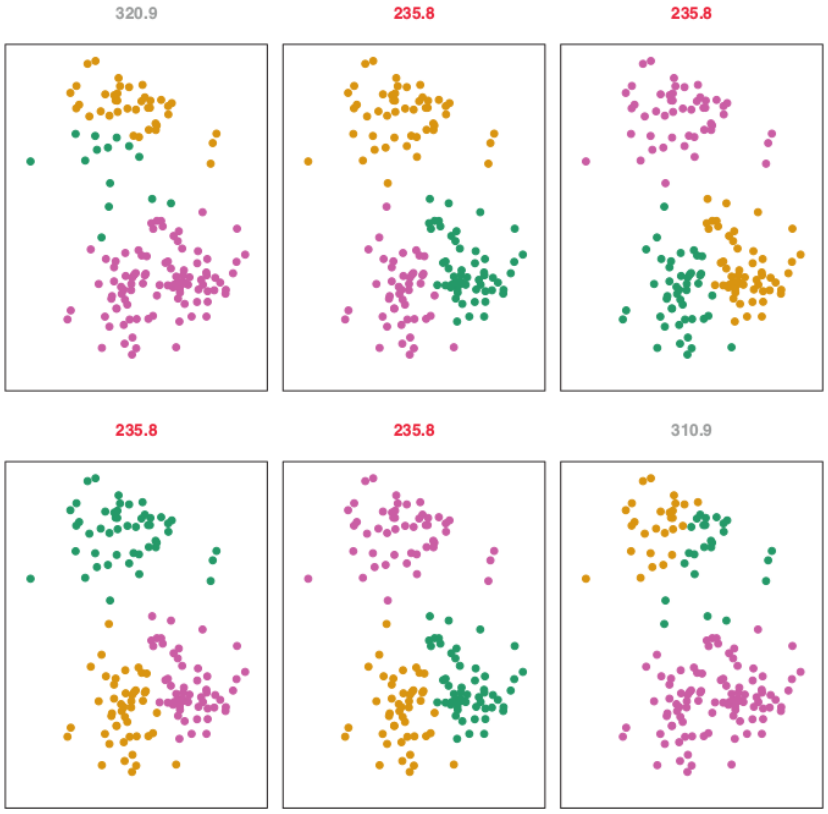


FIGURE 10.6. *The progress of the K-means algorithm on the example of Fig ure 10.5 with K=3.* Top left: *the observations are shown.* Top center: *in Step of the algorithm, each observation is randomly assigned to a cluster.* Top right *in Step 2(a), the cluster centroids are computed. These are shown as large col ored disks. Initially the centroids are almost completely overlapping because th 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) i once again performed, leading to new cluster centroids.* Bottom right: *the result obtained after ten iterations.*

K-medias

La solución obtenida por K-means depende en gran medida de los valores iniciales de asignación a clusters, tiene un componente aleatorio

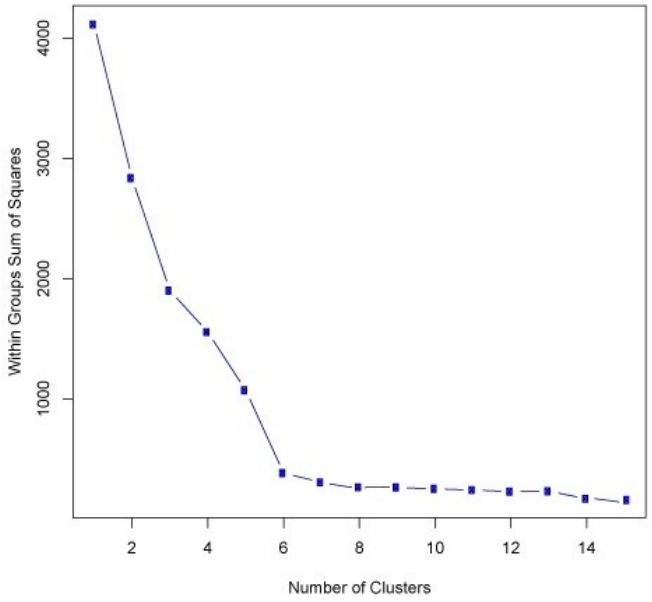
Por este motivo se suele correr el algoritmo muchas veces y quedarse con la mejor solución



K-medias

El valor óptimo a elegir de *K* es algo que no está claro cómo debe ser elegido

Algo que se suele hacer es correr el algoritmo con valores crecientes de K y evaluar cómo varía la función a minimizar, eligiendo como valor de K aquel en donde aparezca un “codo”



K-medias

¿Se modifican los resultados si se escalan las variables?

¿Qué se capta en cada caso?

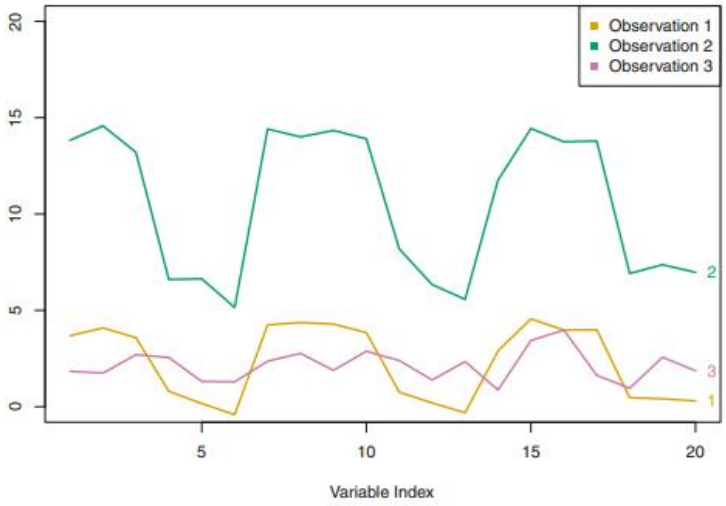


FIGURE 10.13. *Three observations with measurements on 20 variables are shown. Observations 1 and 3 have similar values for each variable and so there is a small Euclidean distance between them. But they are very weakly correlated, so they have a large correlation-based distance. On the other hand, observations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

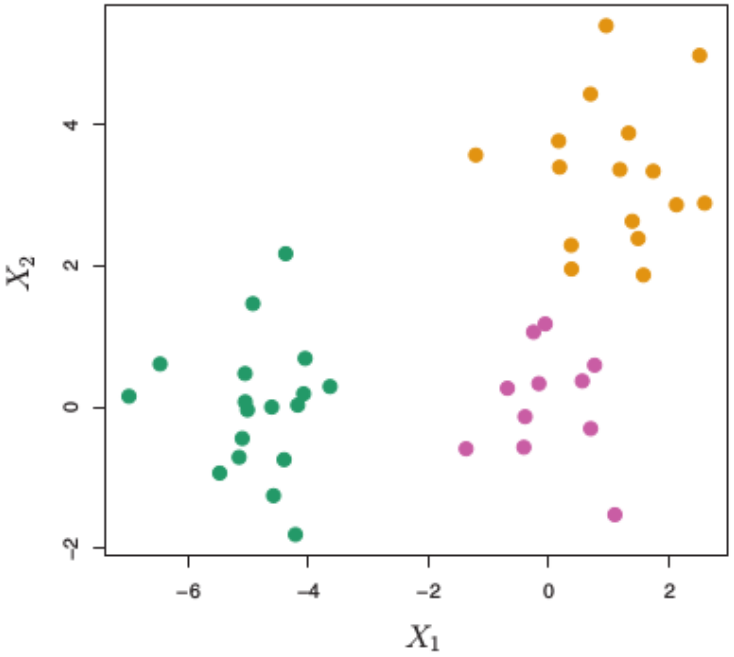
Estructura de la clase

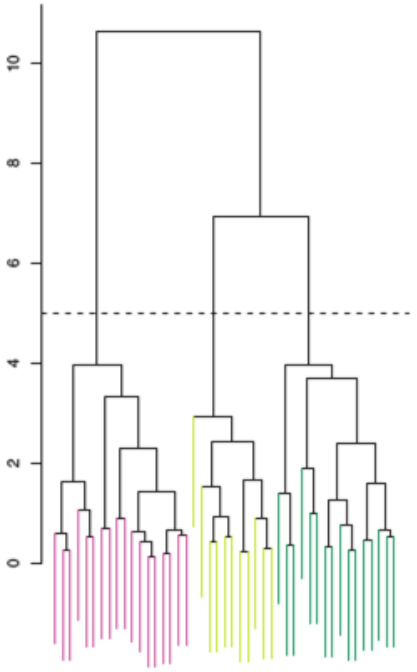
- Motivación
- K-medias
- Clustering jerárquico

Clustering jerárquico

Es un método *bottom-up* o aglomerativo. A diferencia de K-medias, uno no se debe comprometer a un número de clusters antes de ejecutar el algoritmo

El objetivo va a ser obtener un dendrograma

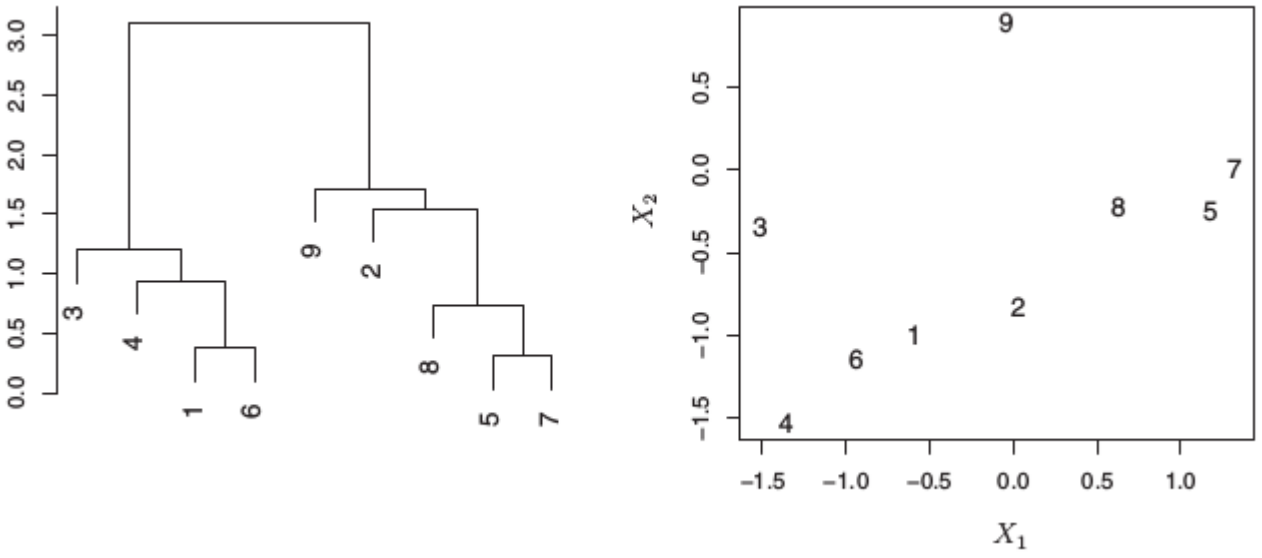




Clustering jerárquico

Dendrograma:

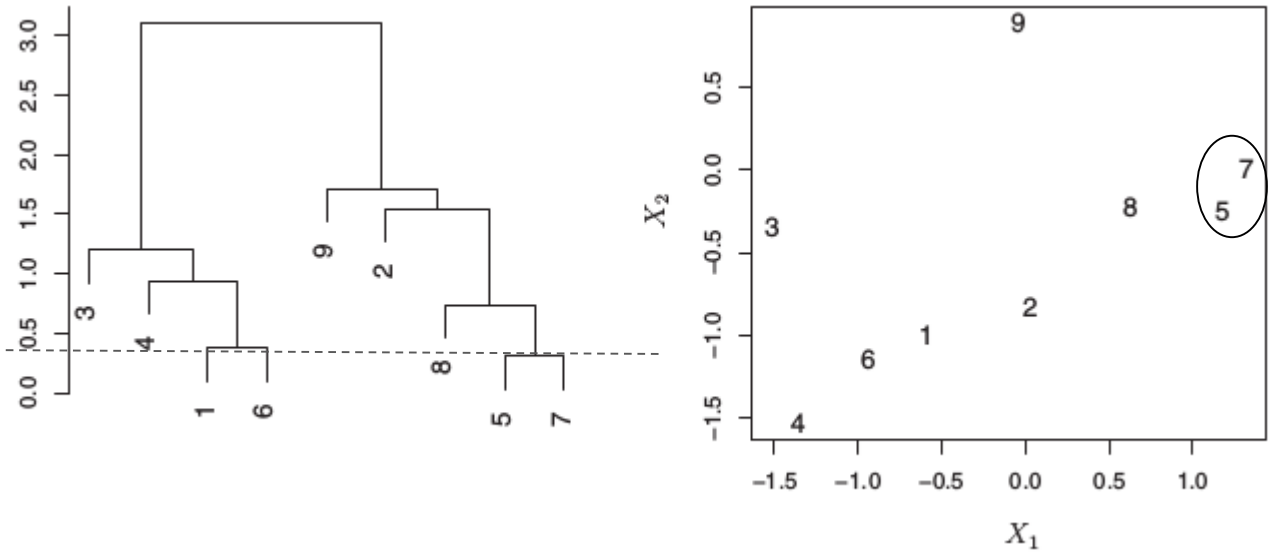
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

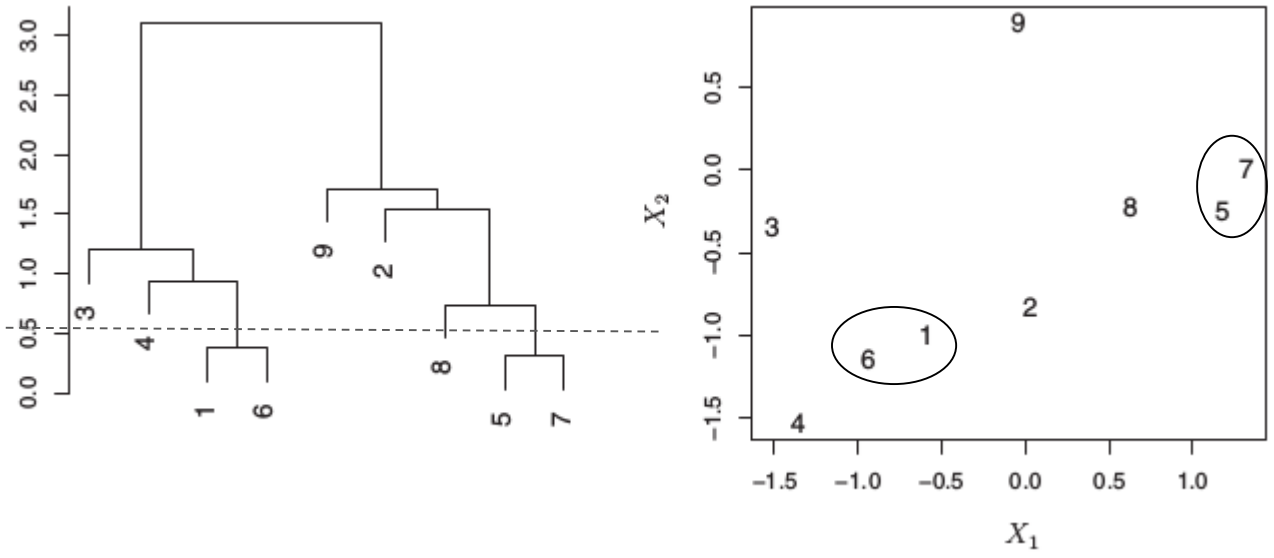
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

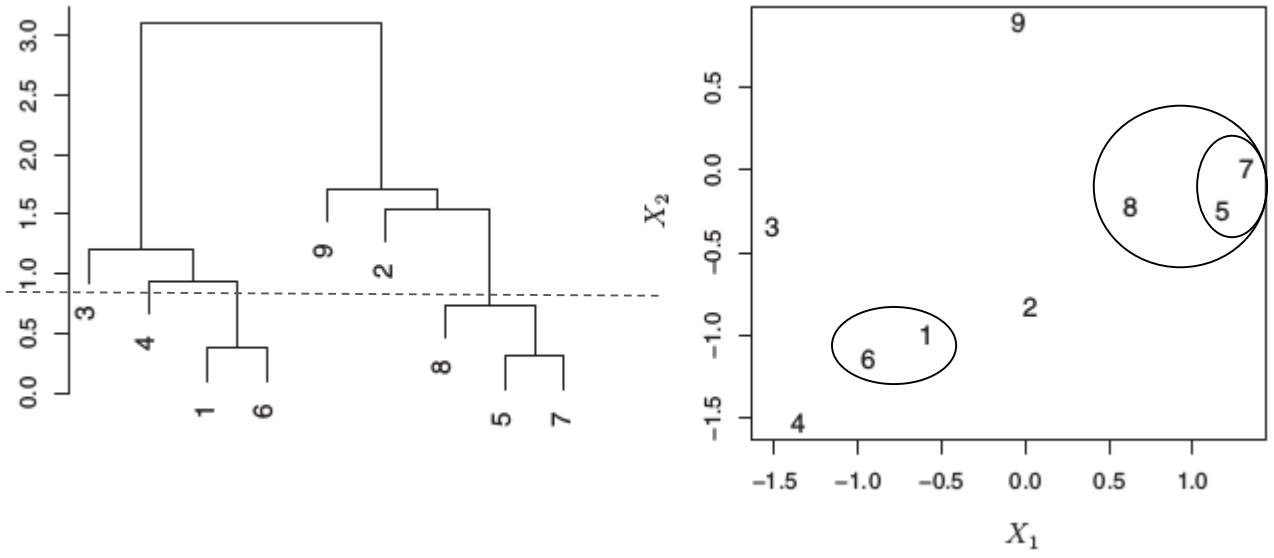
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

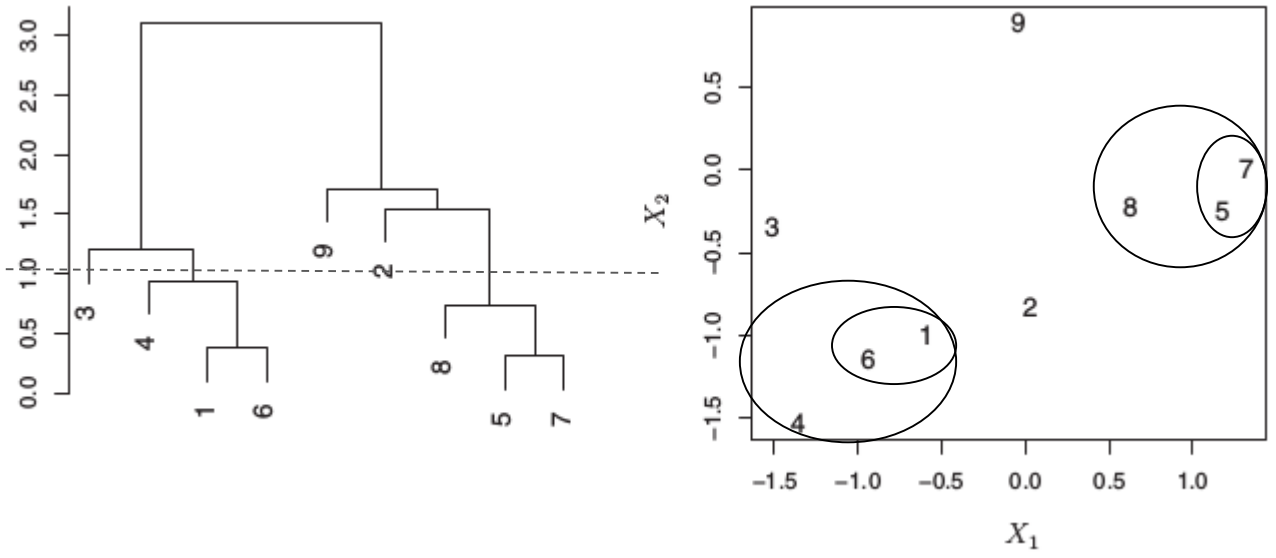
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

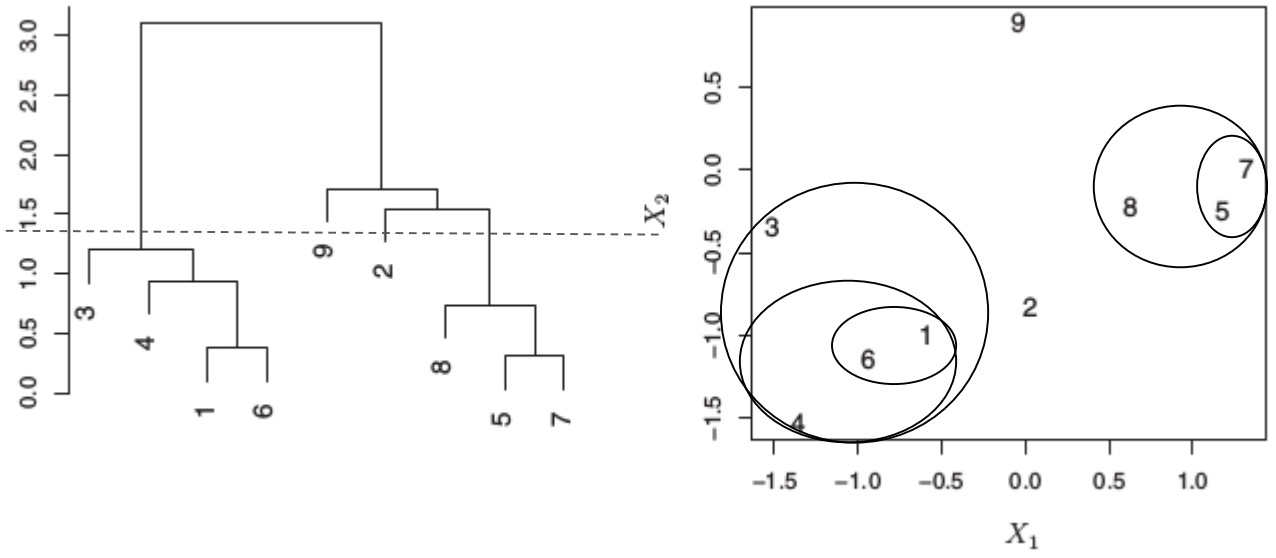
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

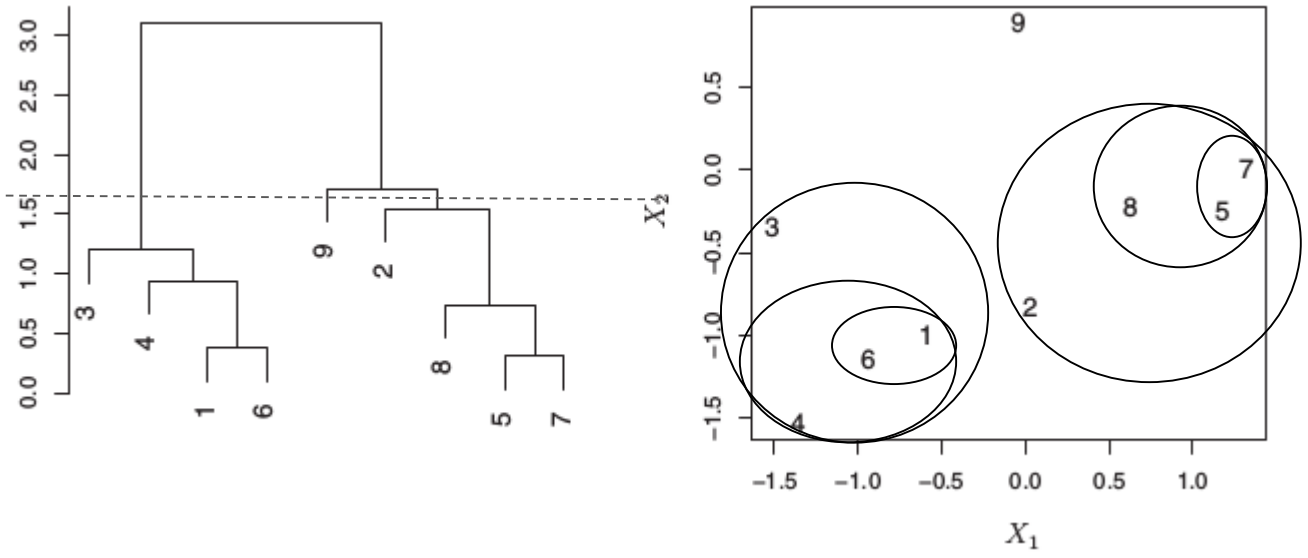
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

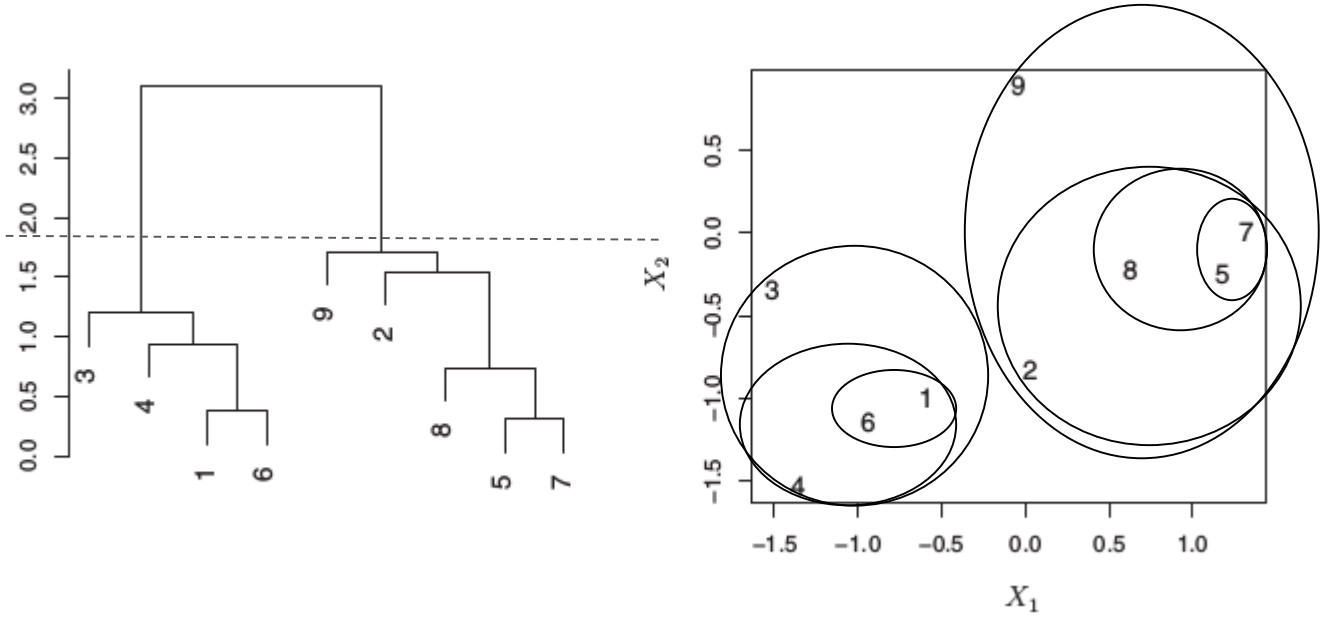
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos



Clustering jerárquico

Dendrograma:

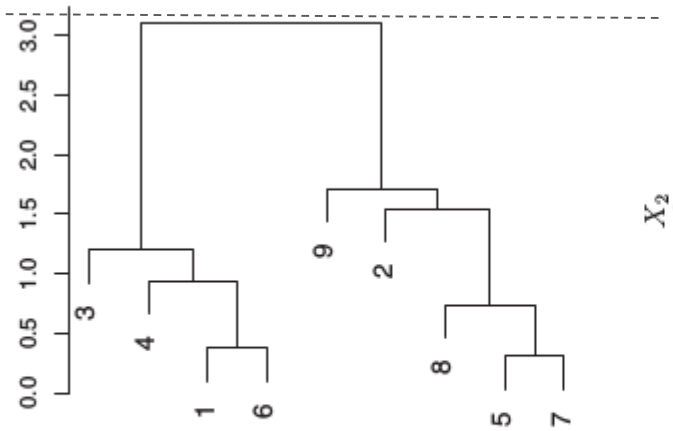
- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos

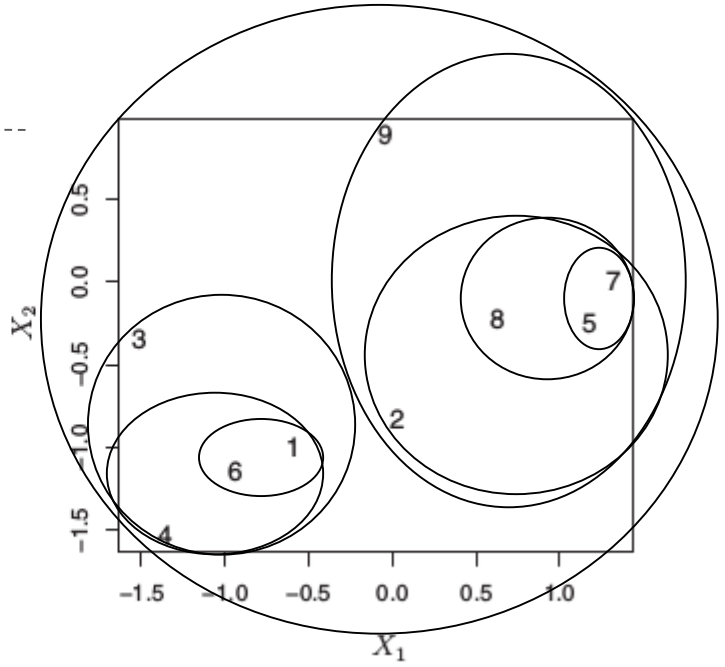


Clustering jerárquico

Dendrograma:

- Cada hoja representa una observación
- Las hojas se unen un grupos similares en entre sí
- A medida que uno sube en el dendrograma los grupos se unen entre sí
- Mientras más “bajo” se haga una unión, más cercanos son los elementos





Clustering jerárquico

Algorithm 10.2 *Hierarchical Clustering*

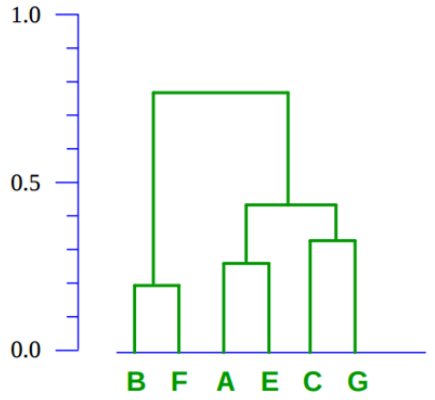
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

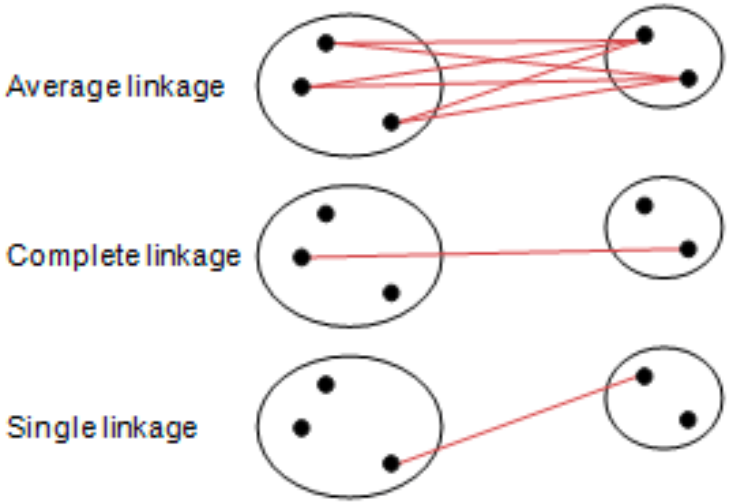
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0



Clustering jerárquico

¿Cómo se ve la distancia entre grupos de registros?



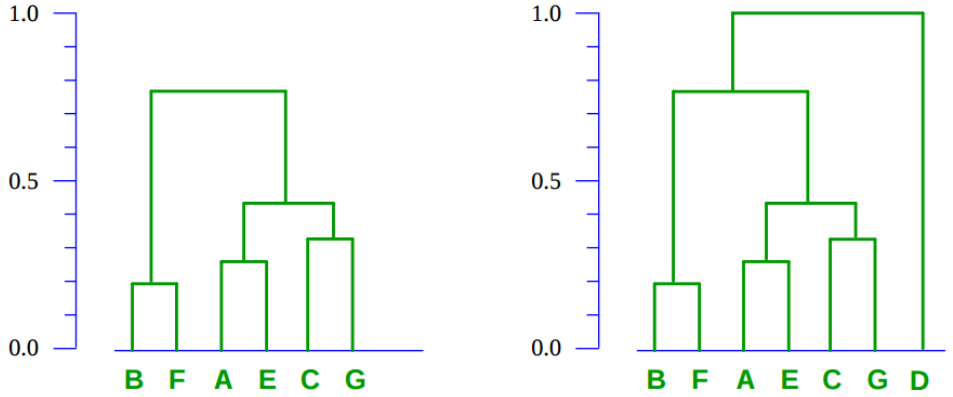
Clustering jerárquico

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0

samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0

Exhibit 7.8 The fifth and sixth steps of hierarchical clustering of Exhibit 7.1, using the 'maximum' (or 'complete linkage') method. The dendrogram on the right is the final result of the cluster analysis. In the clustering of n objects, there are $n-1$ nodes (i.e. 6 nodes in this case).



Clustering jerárquico

El efecto del enlace utilizado para medir distancia entre grupos de observaciones es importante en el resultado final

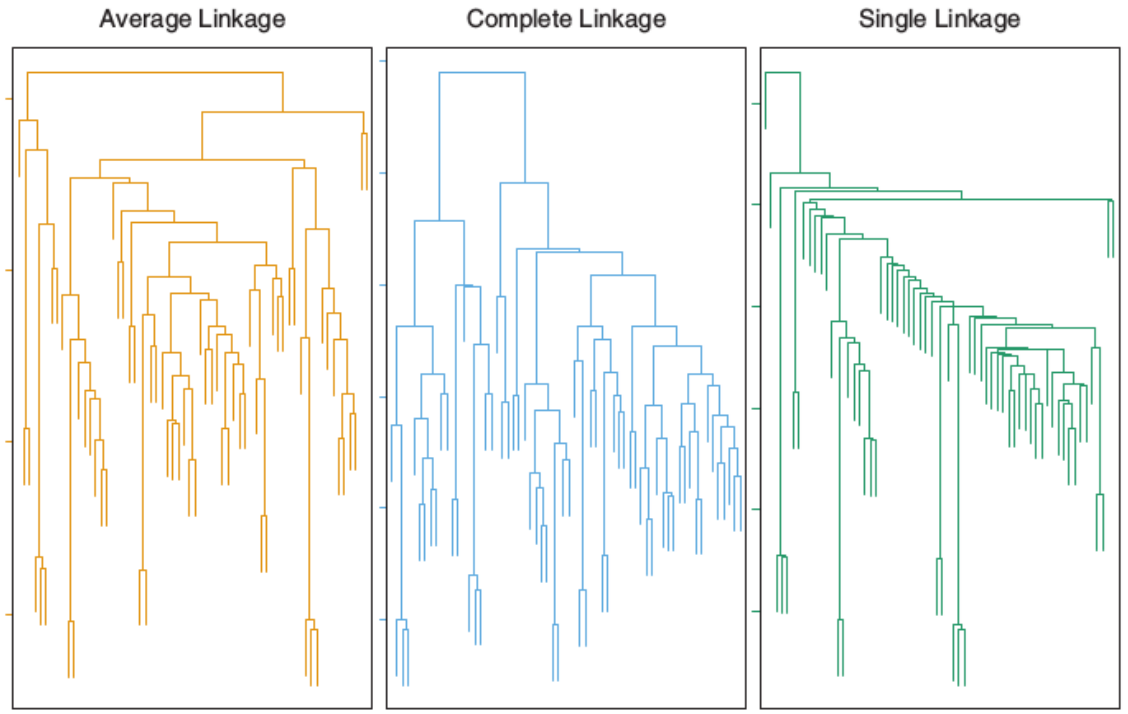


FIGURE 10.12. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

Clustering jerárquico

Consideraciones a tener en cuenta al hacer clustering:

- Pequeñas decisiones pueden tener grandes consecuencias (en la práctica se prueban muchas opciones y se analiza la robustez de los resultados)
- No existe un consenso referido a cómo validar clusters encontrados
- ¿Necesariamente una observación debe pertenecer 100% a un cluster?
- Las particiones pueden ser poco estables al quitar un pequeño subconjunto de observaciones

Bibliografía

ISLP. Sección 12.4