Minería de Datos

UTDT

Master in Management + Analytics

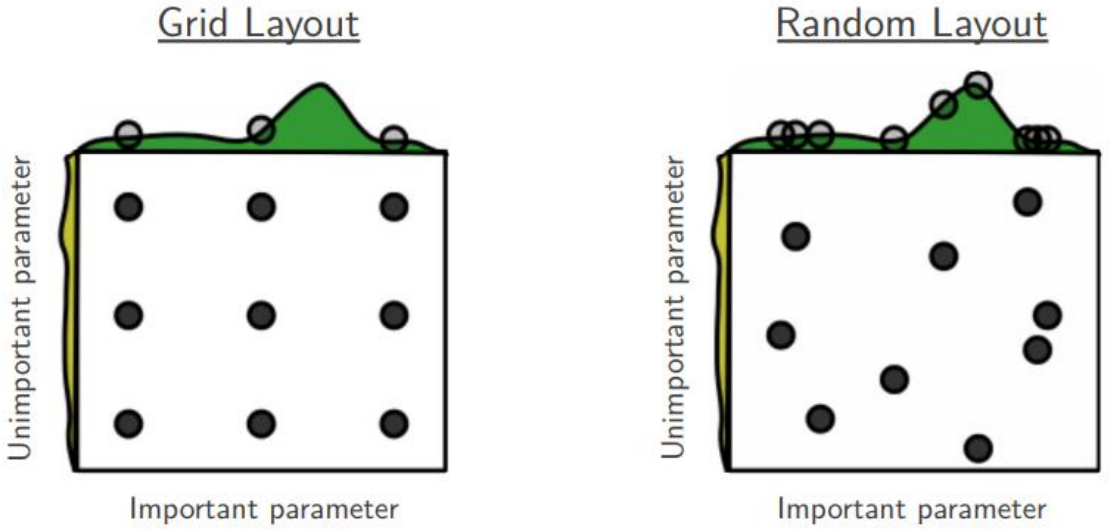
Sesión 5

¿Qué vimos hasta ahora?

- Qué es la minería de datos.
- Distintos tipos de aprendizaje automático.
- Aprendizaje supervisado.
- Naïve bayes.
- K-nearest neighbors.
- Noción de overfitting & underfitting
- Árboles de decisión
- Selección de modelos
- Introducción a Caret
- Ingeniería de atributos (peligro de data leaking)
- Métricas de performance de clasificadores
- Introducción a XGboost
- K-means

Optimización de hiperparámetros

La performance de un modelo comúnmente depende de varios hiperparámetros (por ej., en árboles α y el número mínimo de observaciones para realizar un split de un nodo. En xgboost hay 7…). Existen distintas estrategias para probar las combinaciones de los mismos.



Random search es más robusto a la presencia de hiperparámetros irrelevantes y puede acelerar el proceso de búsqueda (fuente, no es obligatorio leerla).

Optimización de hiperparámetros

Veamos el primer bloque de código.

Aprendizaje no supervisado

Sólo tendremos valores conocidos (X) y nuestro objetivo será descubrir patrones interesantes detrás de los datos.

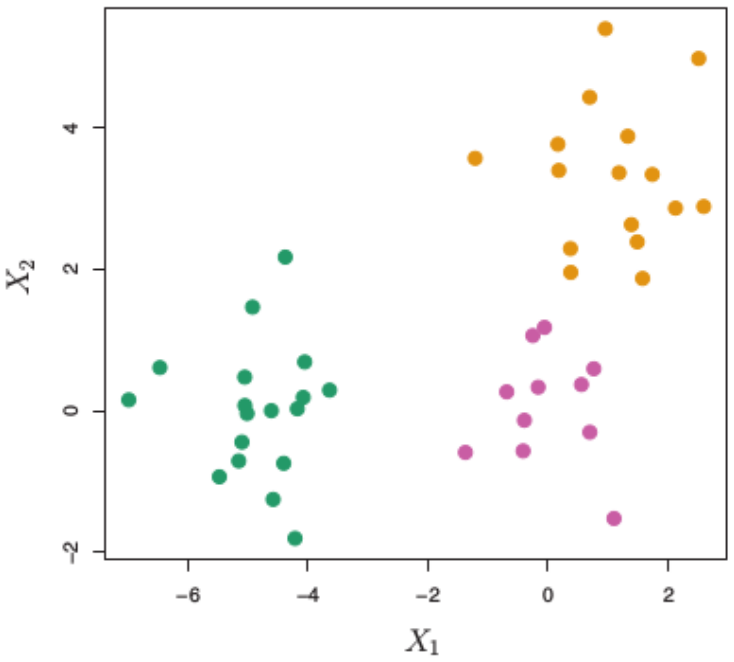
Por ejemplo:

- Existe alguna forma informativa de visualizar los datos.
- Se pueden detectar subgrupos similares de observaciones.
- Se pueden detectar subgrupos de variables relacionadas.

Clustering jerárquico

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

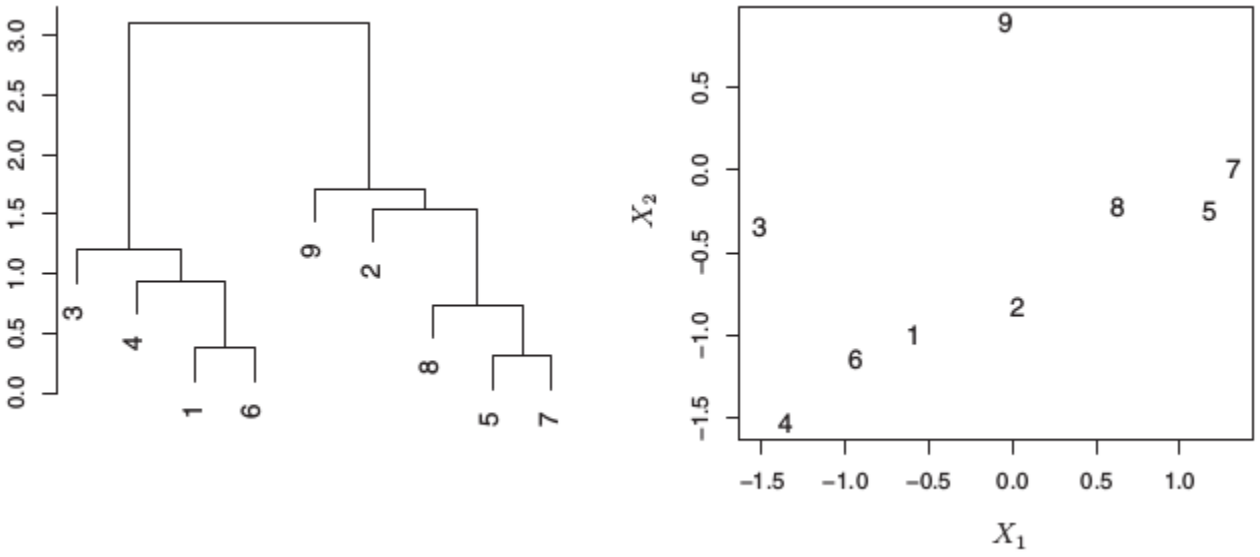
El objetivo va a ser obtener un dendrograma.



Clustering jerárquico

Dendrograma:

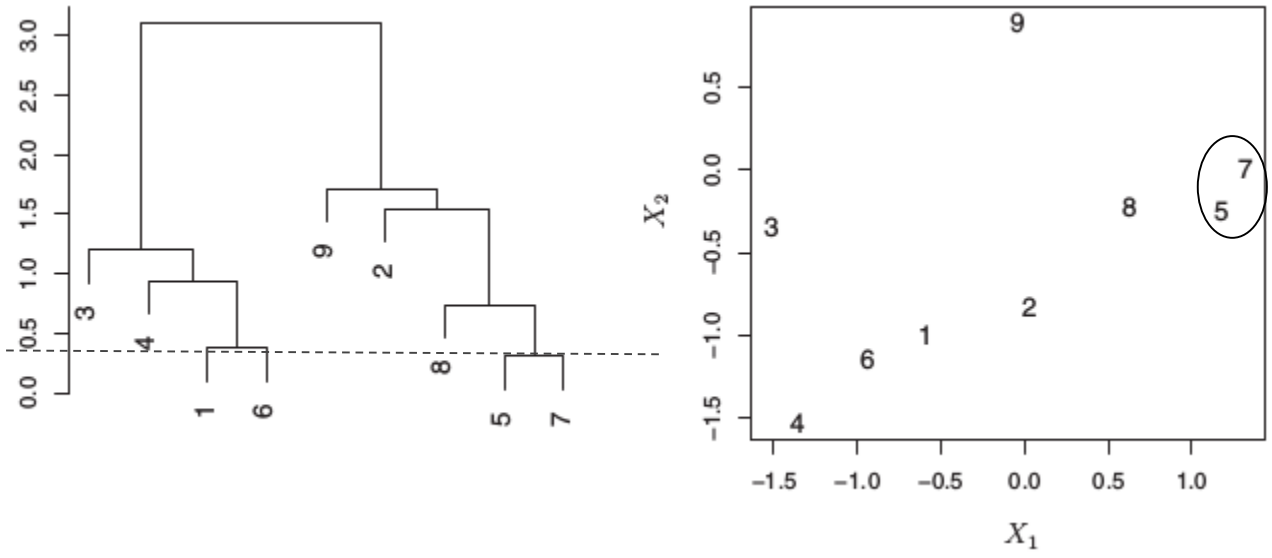
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendrograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

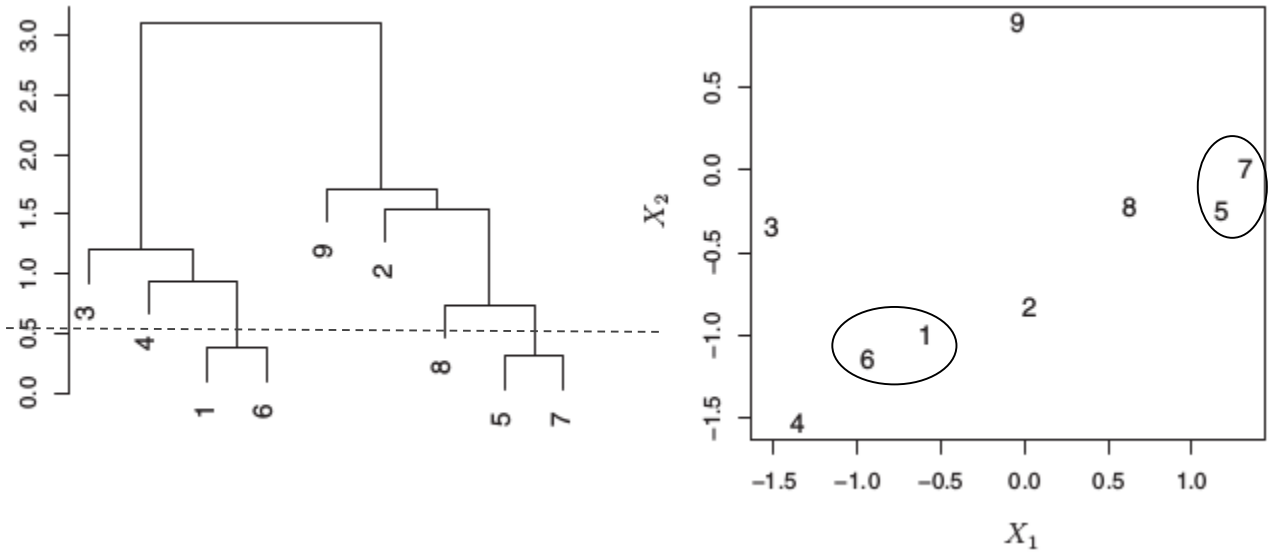
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

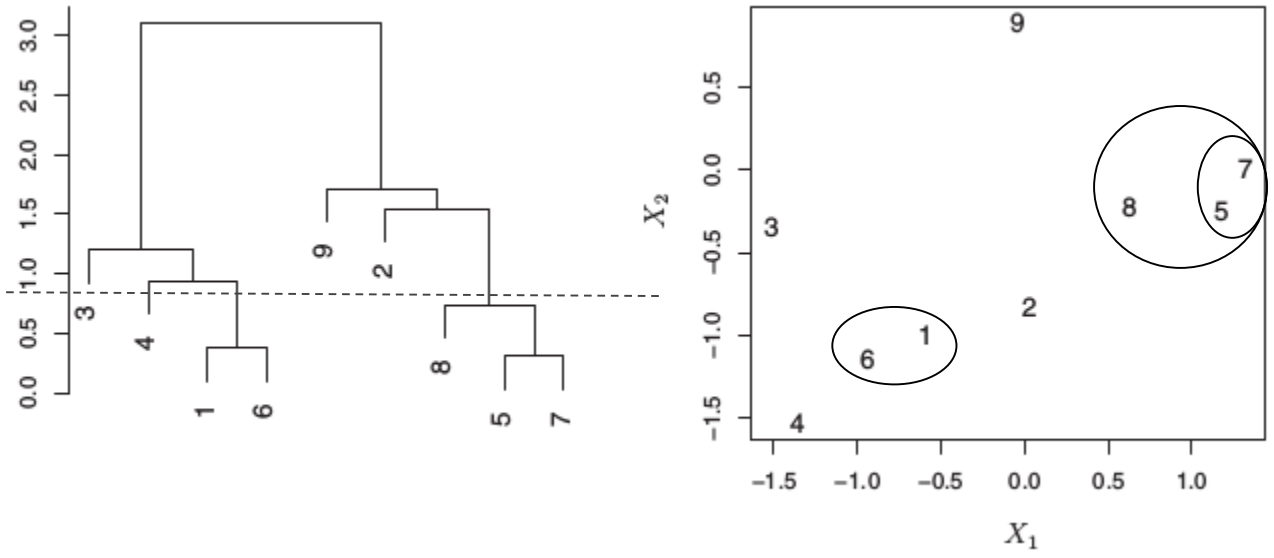
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

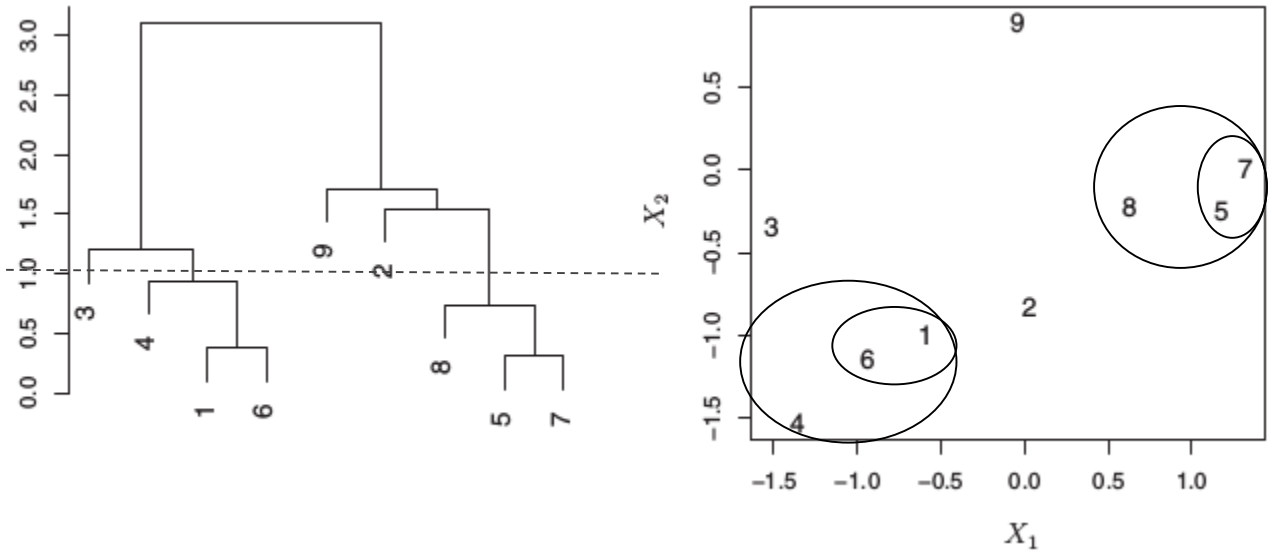
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

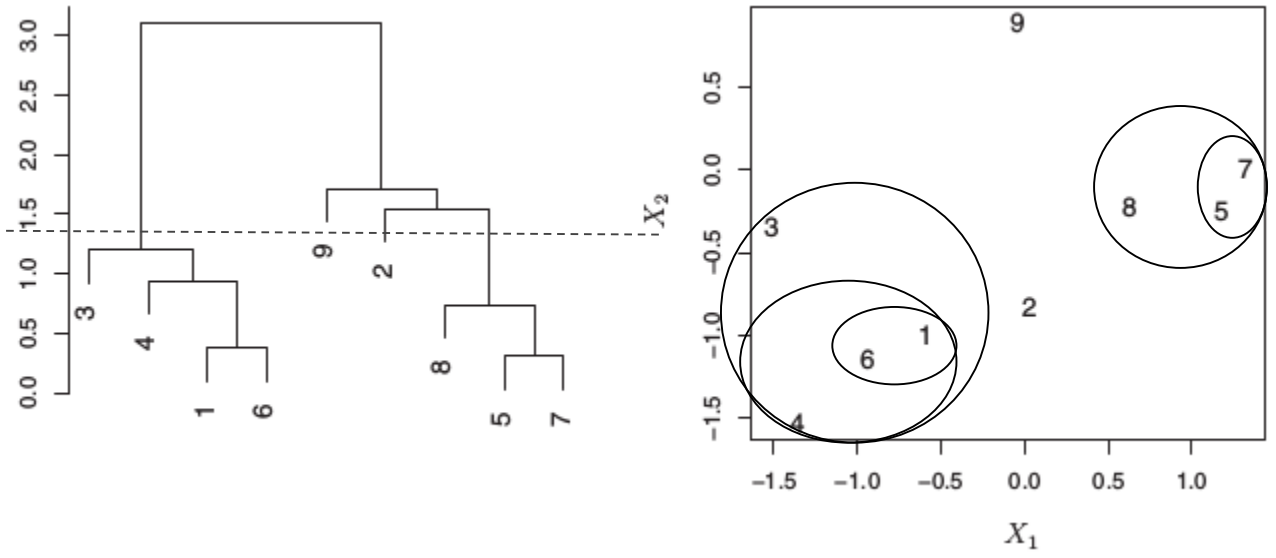
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

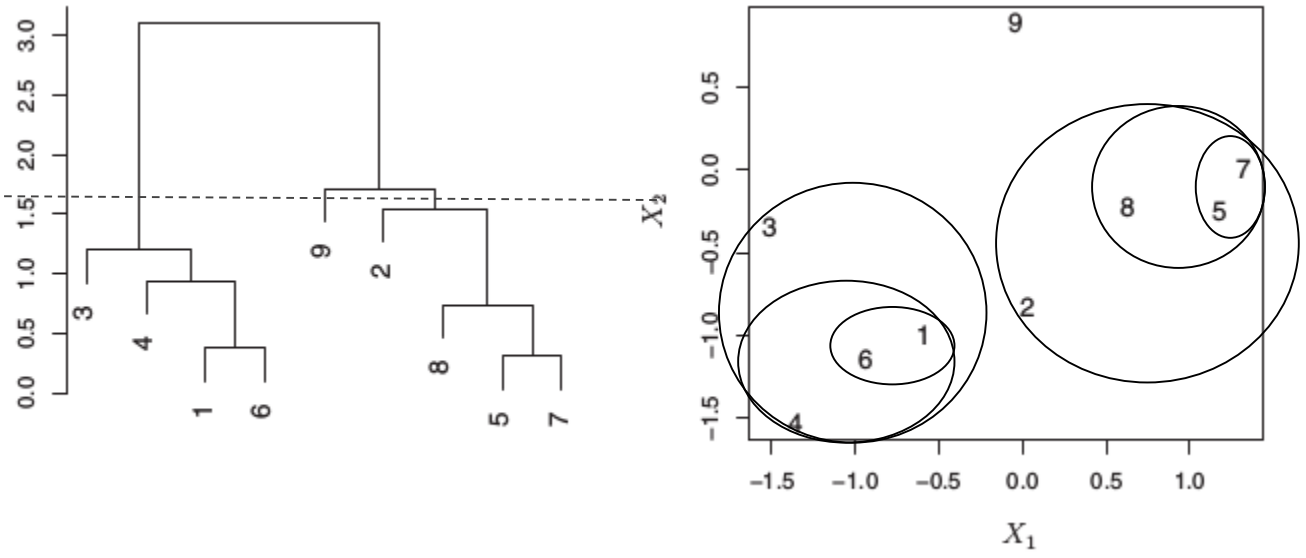
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

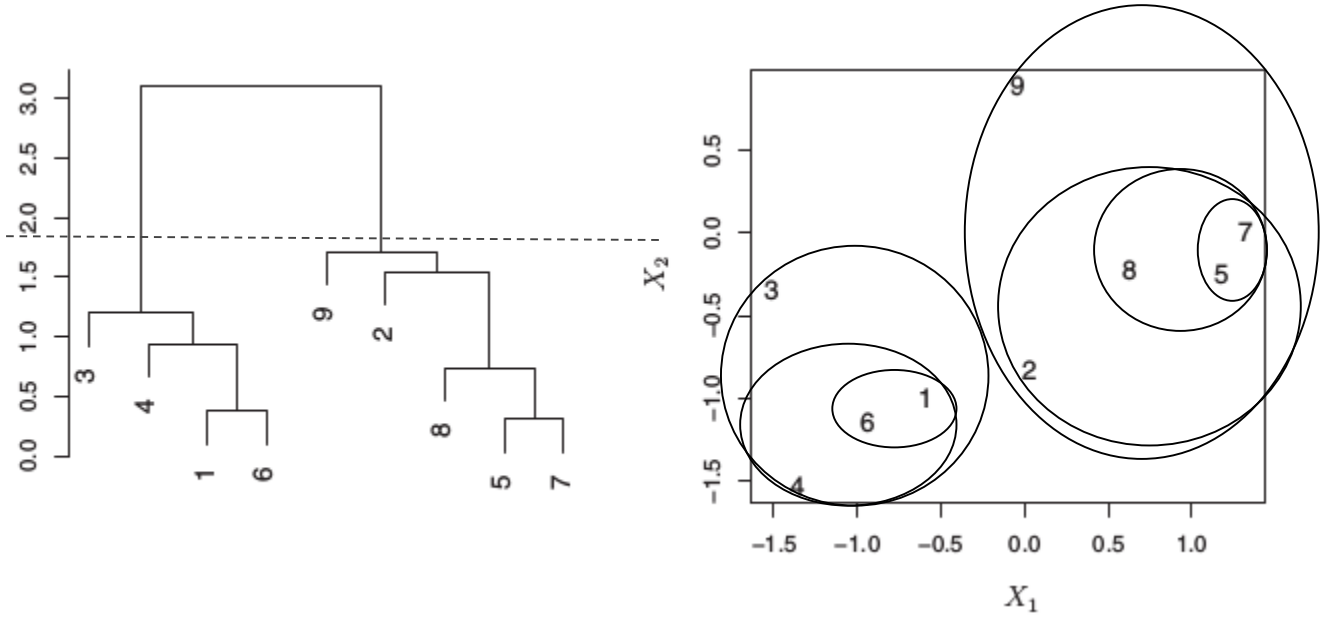
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

Dendrograma:

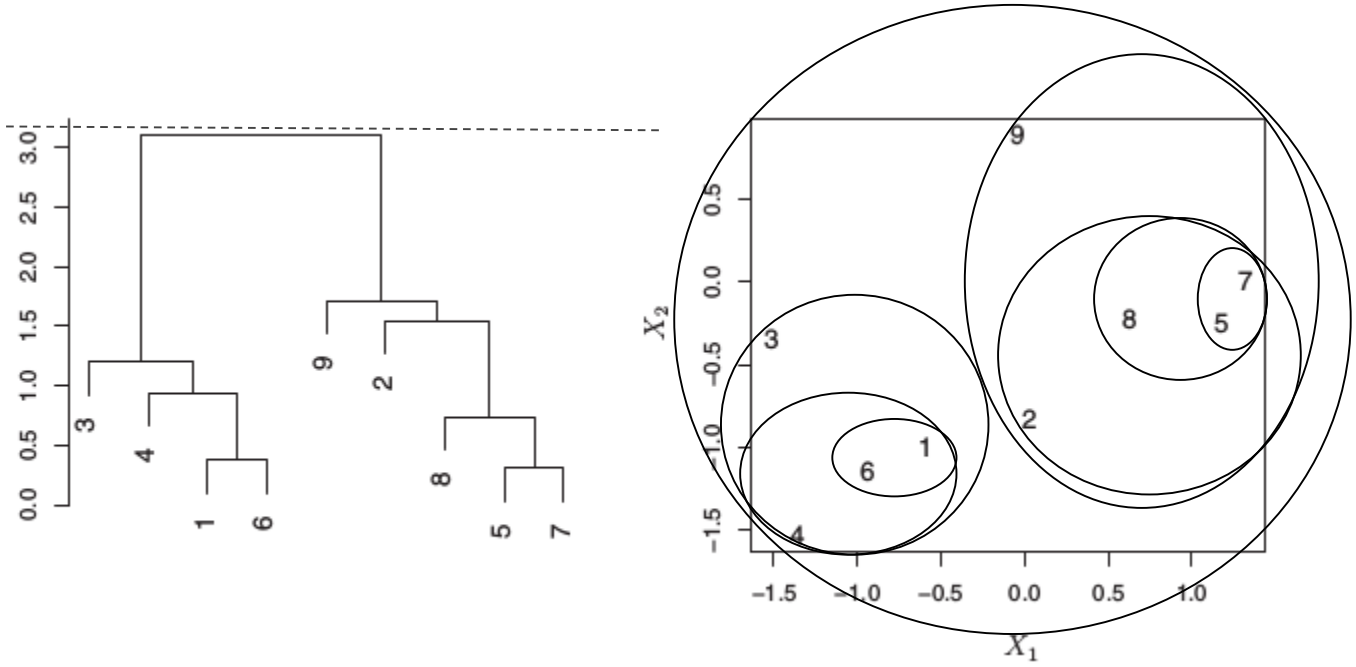
- Cada hoja representa una observación.
- Las hojas se unen un grupos similares en entre sí.
- A medida que uno sube en el dendograma los grupos se unen entre sí.
- Mientras más “bajo” se haga una unión, más cercanos son los elementos.



Clustering jerárquico

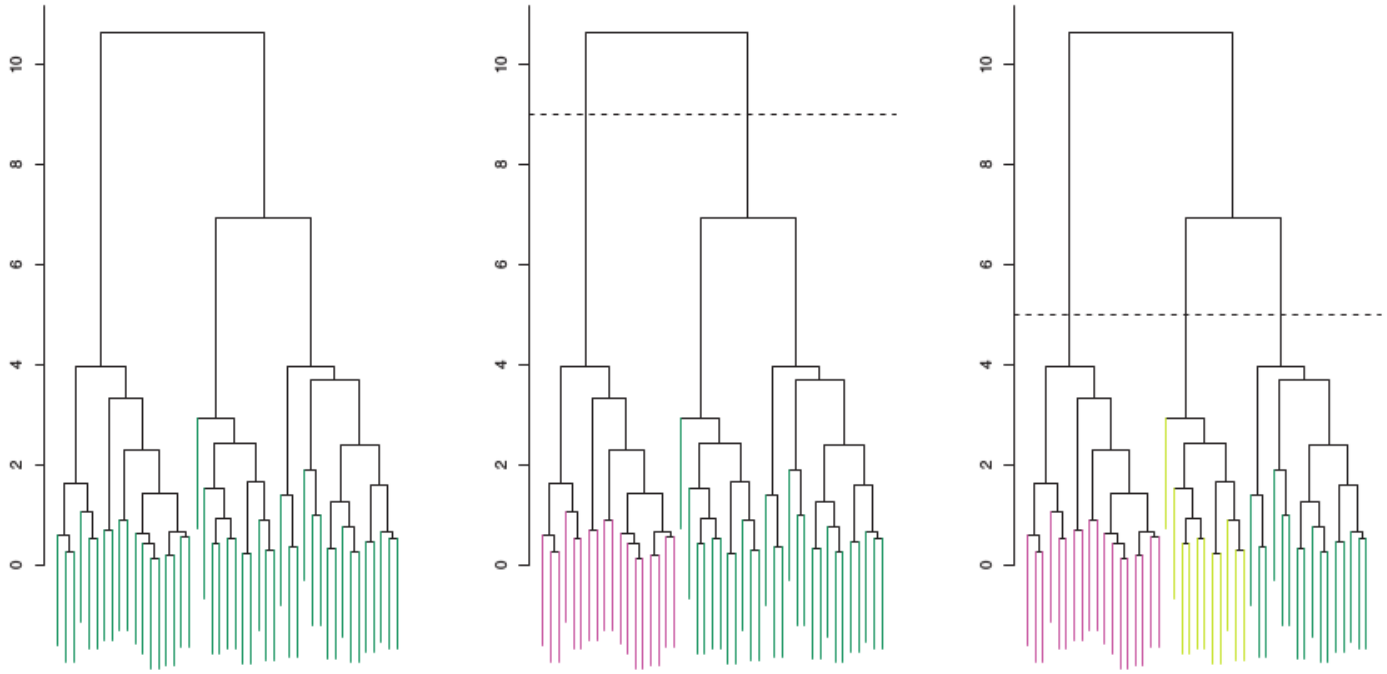
Dendrograma:

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



Clustering jerárquico

Los clusters se encuentran haciendo cortes a distintos niveles del dendrograma.



Clustering jerárquico

Algorithm 10.2 *Hierarchical Clustering*

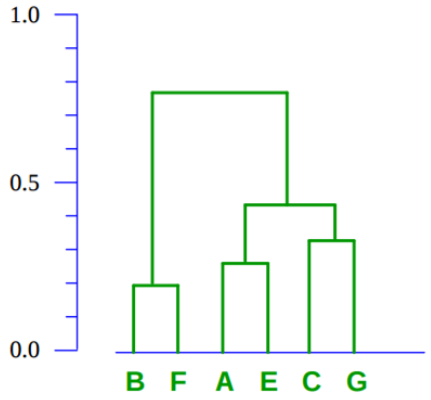
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

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.

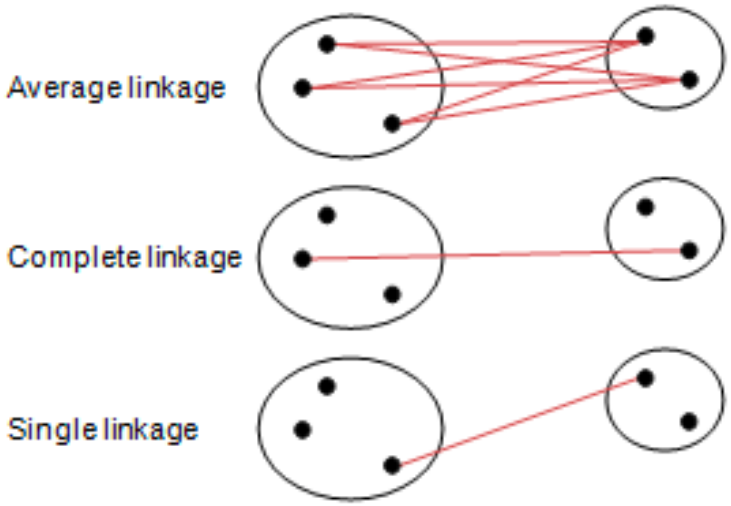
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

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?



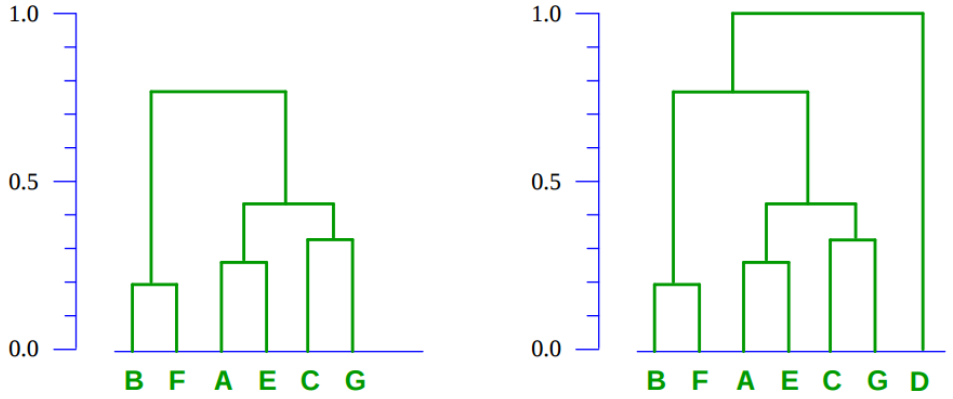
Clustering jerárquico

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 de la técnica usada para medir distancia entre grupos de observaciones, es importante en el resultado final.

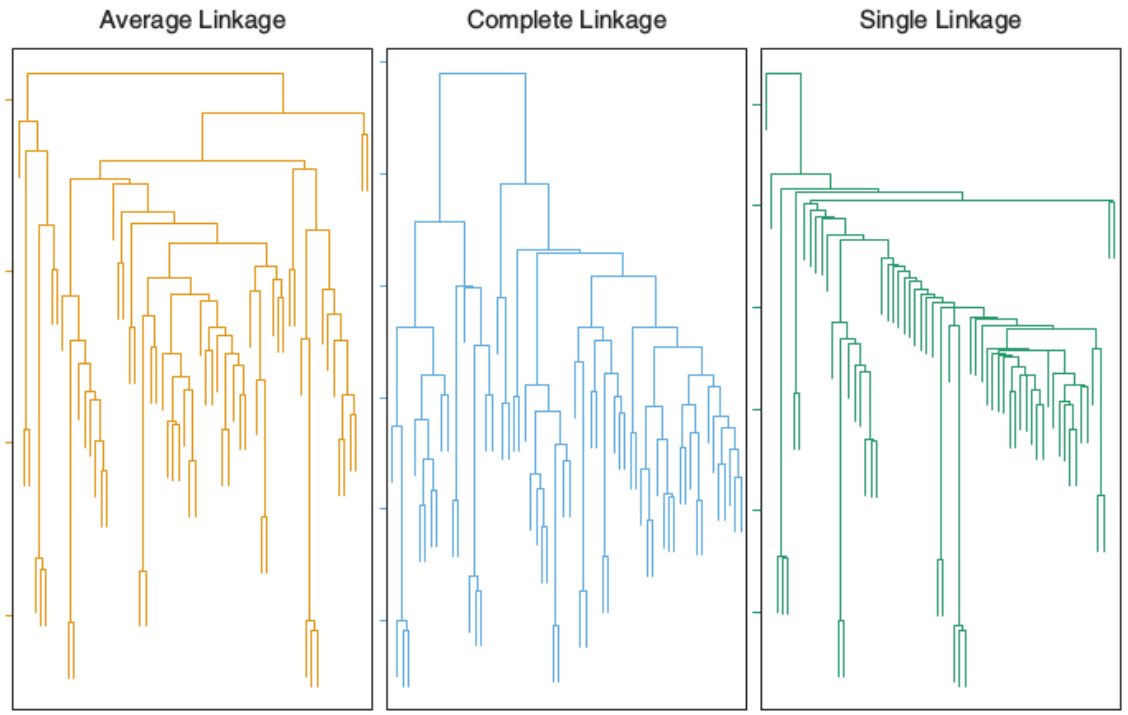


FIGURE 10.12. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

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.

Clustering jerárquico en R

Veamos el segundo bloque de código.

Load characterization

Ahora vamos a ver una aplicación interesante de clustering.

Veamos el tercer bloque de código.

Tiempo para consultas

Lecturas recomendadas para los temas vistos hoy

Clustering jerárquico:

- ISLR (Cap 10, salvo lo referido a componentes principales)

Práctica de laboratorio

Para hacer:

cargue el dataset "Hitters" (que es parte de la librería) ISLR y elimine todas las variables que no sean numéricas.

Se pide:

- Construya y grafique un dendrograma utilizando el método "complete", "single" y "average". ¿Son diferentes los dendrogramas encontrados?
- Repita el punto anterior, pero escalando las variables. ¿Son diferentes los dendrogramas encontrados utilizando diferentes métodos? ¿Son diferentes a los que se obtienen cuando las variables no están escaladas?
- Elijan el dendrograma que más le convence de los construidos hasta ahora. Elija un número de clusters que considere adecuado (llamémosle esquema A). Ahora corra el algoritmo de k-medias sobre los mismos datos (ya sea escalados o no, tal como lo haya hecho en el esquema A) y buscando el mismo número de clusters que antes consideró adecuado (llamémosle B a este esquema).
- Estudie qué es lo que mide el coeficiente silhoutte para validar esquemas de clustering y vea cómo se calcula el mismo en R. Compare el esquema A y B usando esta métrica.

Práctica teórica

1) ¿Por qué k-medias puede calcularse con grandes volúmenes de datos, pero no así los esquemas de clustering jerarquico vistos en clases? Justifique su respuesta

2) ¿Los esquemas de clusteing jerárquico se constuyen minimizando alguna función de costo? Justifique su respuesta.

3) El siguiente código escribe una matriz de distancia, y obtiene dendrogramas siguiendo tres estrategias diferentes de linkage. Calcule los tres dendrogramas a mano y revise que efectivamente obtiene los mismos resultados.

```
dist_m <- matrix(c(0.00, 0.21, 0.10, 0.15, 0.30,
                   0.21, 0.00, 0.05, 0.21, 0.32,
                   0.10, 0.05, 0.00, 0.20, 0.31,
                   0.15, 0.21, 0.20, 0.00, 0.23,
                   0.30, 0.32, 0.31, 0.23, 0.00), ncol = 5)

plot(hclust(as.dist(dist_m), method = "complete"))
plot(hclust(as.dist(dist_m), method = "single"))
plot(hclust(as.dist(dist_m), method = "average"))
```

4) Proponga usted una nueva matriz de distancia y respita el punto anterior.