CLUSTERING

Bases de Datos Masivas

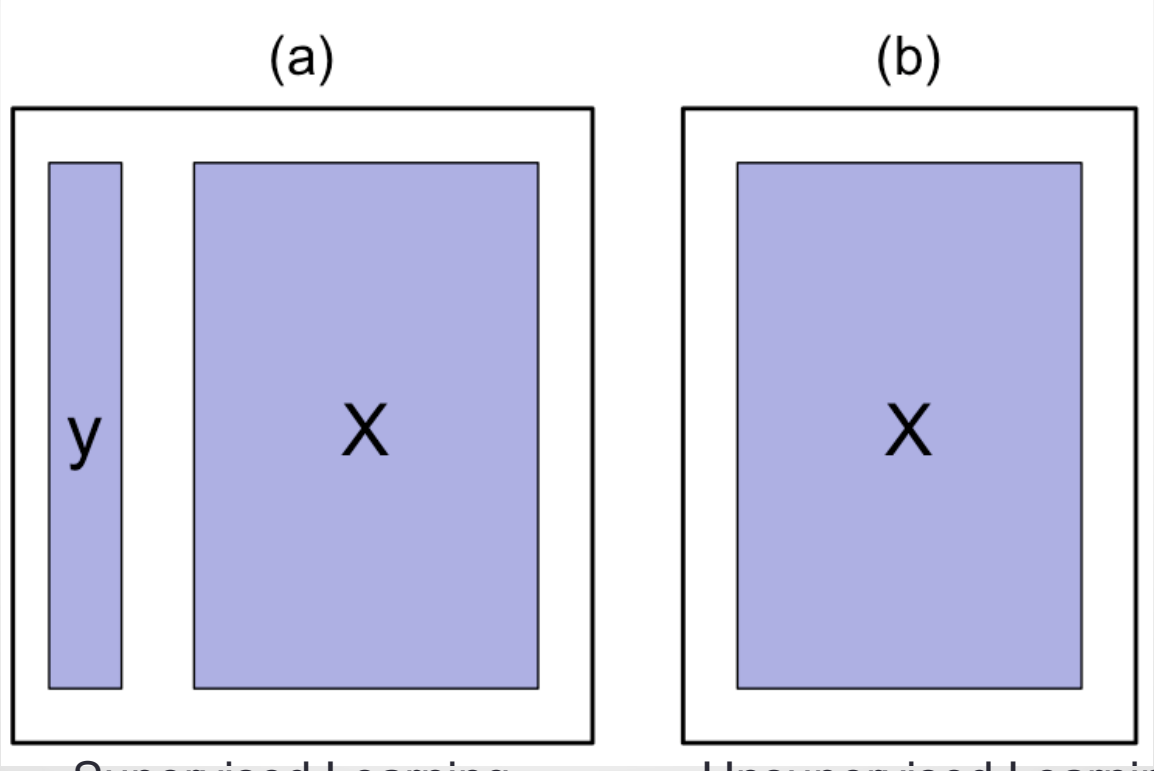
Temas

- ¿Qué es clustering?
- K-Means Clustering
- Hierarchical Clustering

¿QUÉ ES CLUSTERING?

Aprendizaje Supervisado vs. No Supervisado

- Aprendizaje Supervisado: tanto X como Y son conocidos
- Aprendizaje No Supervisado solo conozco X



Problema del agrupamiento

Dada una nube de **puntos** en un espacio, queremos entender cuál es la estructura de esa nube.

También podemos tener:

- **Vectores**
- **Conjuntos**

Problema del agrupamiento

Dado un **conjunto de puntos**, con una noción de **distancia** entre puntos, se agrupan los puntos en un **número determinado de *clusters***, de modo que:

- Los miembros de un grupo están cerca (**similares** entre sí).
- Los miembros de los otros grupos son bien diferentes.

Generalmente:

- Los puntos están en un espacio de dimensión alta.
- La similitud se define utilizando una medida de distancia.

  Euclidiana, coseno, Jaccard, distancia de edición, ...

Clustering: Problema Hard ¿Por qué?

- El agrupamiento en dos dimensiones parece fácil
- El agrupamiento de pequeñas cantidades de datos parece fácil
- Y en la mayoría de los casos, las apariencias **no engañan**

Pero la mayoría de los problemas no tienen 2 dimensiones, sino que pueden tener 10, 100 o miles de dimensiones

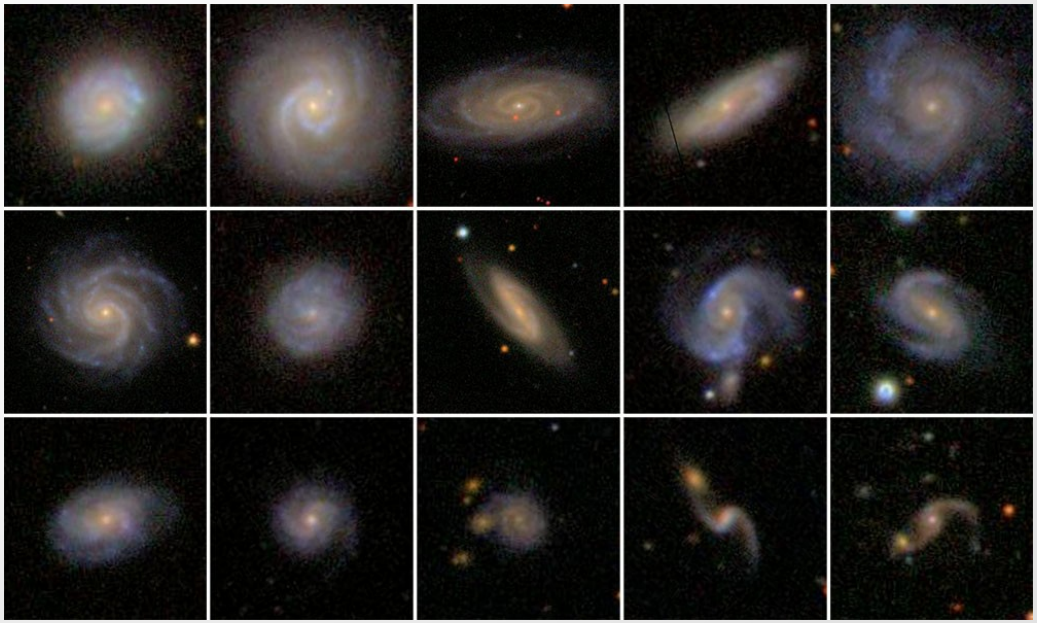
En espacios de alta dimensionalidad todo se ve diferente:

- **La Maldición de la Dimensionalidad**:
  - Una manifestación de la "maldición" es que en dimensiones altas, casi todos los pares de puntos están igualmente alejados unos de otros.

Problemas de Clustering: Galaxias

- **Datos:** Un catálogo de 2 mil millones de "cuerpos celestes" representados por su radiación en 7 dimensiones (bandas de frecuencia)
- **Problema**: Cluster en objetos similares, por ejemplo: galaxias, las estrellas cercanas, cuásares, etc.

Proyecto:
Sloan Digital Sky Survey



Problemas de Clustering: CDs de música

- **Intuitivamente**: La música se divide en categorías, los clientes prefieren unas pocas categorías.
  - ¿Pero qué son realmente las categorías?
- Representar un CD por el conjunto de clientes que lo compraron:
- CDs similares tienen un conjunto de clientes similares y viceversa.

Problemas de Clustering: CDs de música

Espacio de todos los CDs:

- Piense en un espacio con una dimensión para cada cliente.
- Los valores en una dimensión pueden ser 0 o 1 solamente.
- Un CD es un punto en este espacio (x_1, x_2, ..., x_k), donde x_i = 1 si y sólo si el i-ésimo cliente compró el CD

	C_1	C_2	C_3	...	C_k
CD_1	1	0	1		1
CD_2	0	1	0		0
CD_3	1	1	1		1
...					
CD_n	0	1	0		0

Para Amazon, la dimensión es de decenas de millones

Tarea: Buscar ***clusters*** de CDs similares

Clustering

- Clustering hace referencia al conjunto de técnicas para buscar subgrupos (o clusters) en un conjunto de datos.
- Un buen clustering permite que las observaciones **dentro de un grupo sean similares** pero **entre los grupos sean bien diferentes.**
- El clustering también es llamado **segmentación de datos** en algunas aplicaciones.
  - Se realizan particiones de grandes conjuntos de datos, en clusters de acuerdo a su similitud.

Clustering

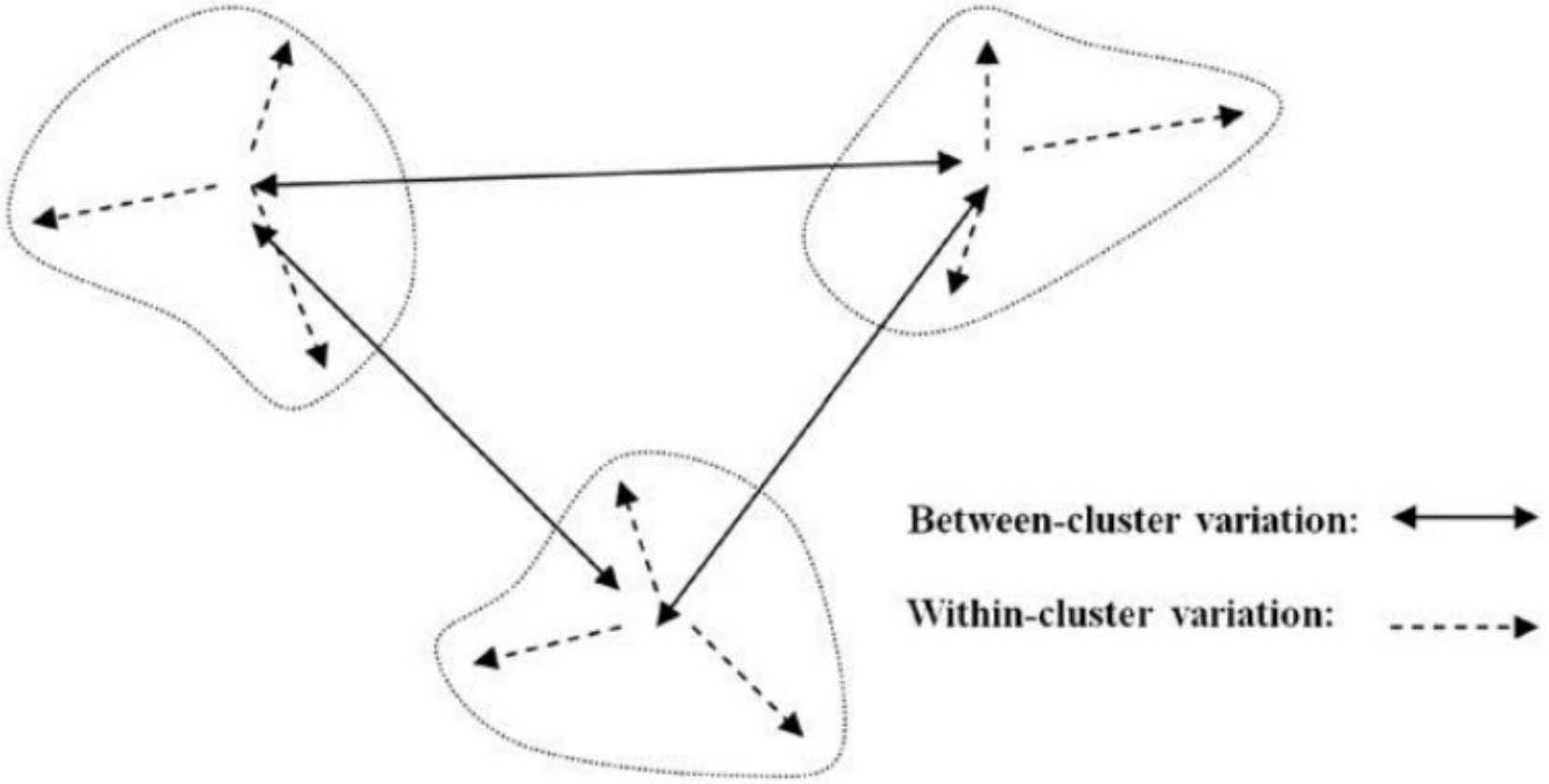
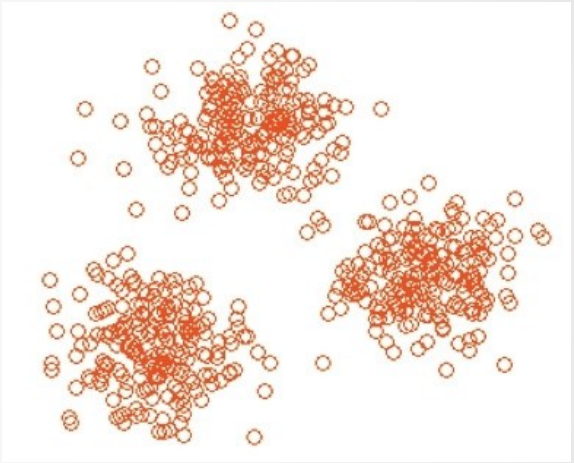


Figure 10.1 Clusters should have small within-cluster variation compared to the between-cluster variation.

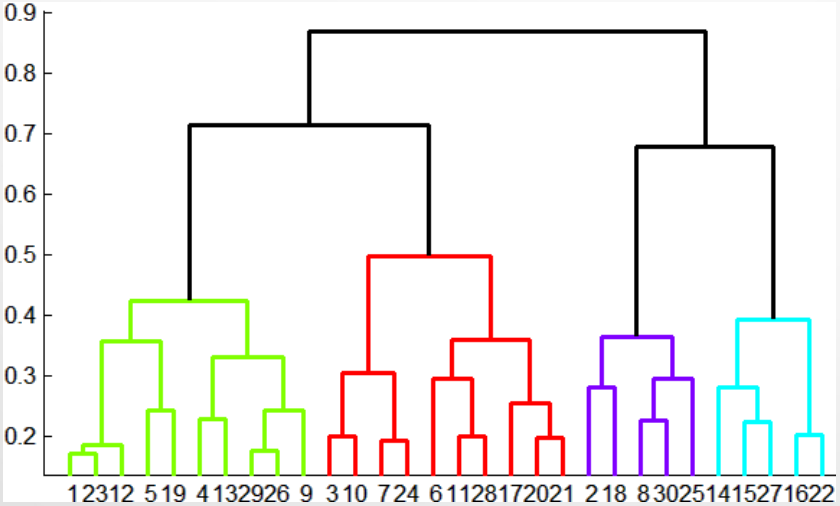
(Larose, 2014)

Métodos de Clustering

- Existen muchos tipos de métodos de agrupamiento diferentes.

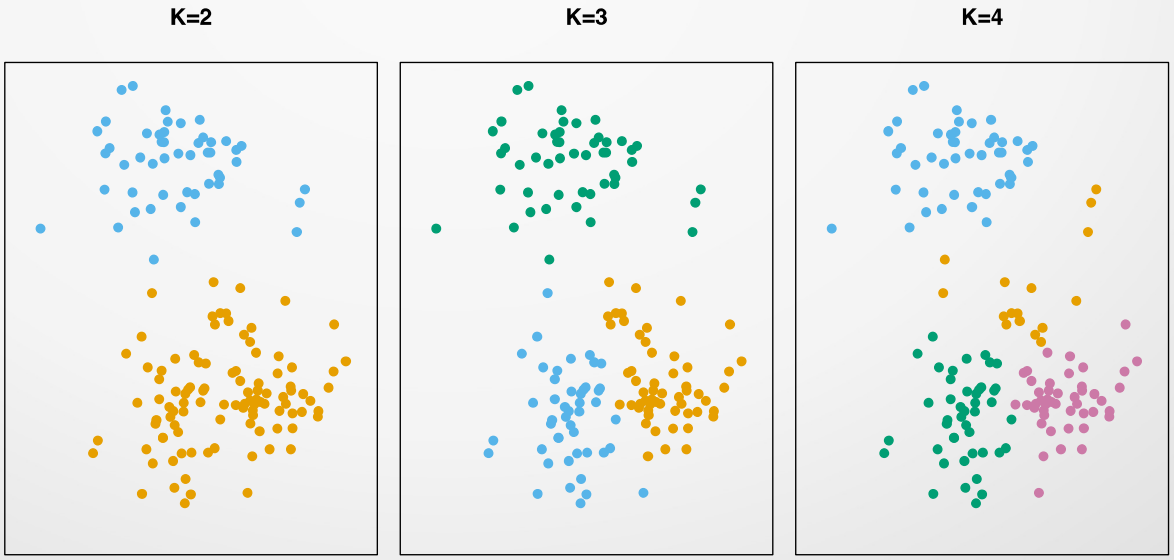


- Dos de los más utilizados son:
  - K-Means Clustering
  - Hierarchical Clustering



K-Means Clustering

- Para realizar un agrupamiento por **K-Means**, debemos especificar en primer lugar la cantidad de clusters (**K**) deseados.
- Entonces el algoritmo de K-means asignará cada una de las observaciones exactamente a uno solo de los K clusters



Métodos de *Clustering*

Jerárquico:

- Aglomerativo (de abajo hacia arriba):
  - Inicialmente, cada punto es un cluster
  - Combina repetidamente los dos grupos **"más cercanos"** en uno.
- Divisivo (de arriba abajo):
  - Comienza con un *cluster* y de forma recursiva lo va dividiendo.

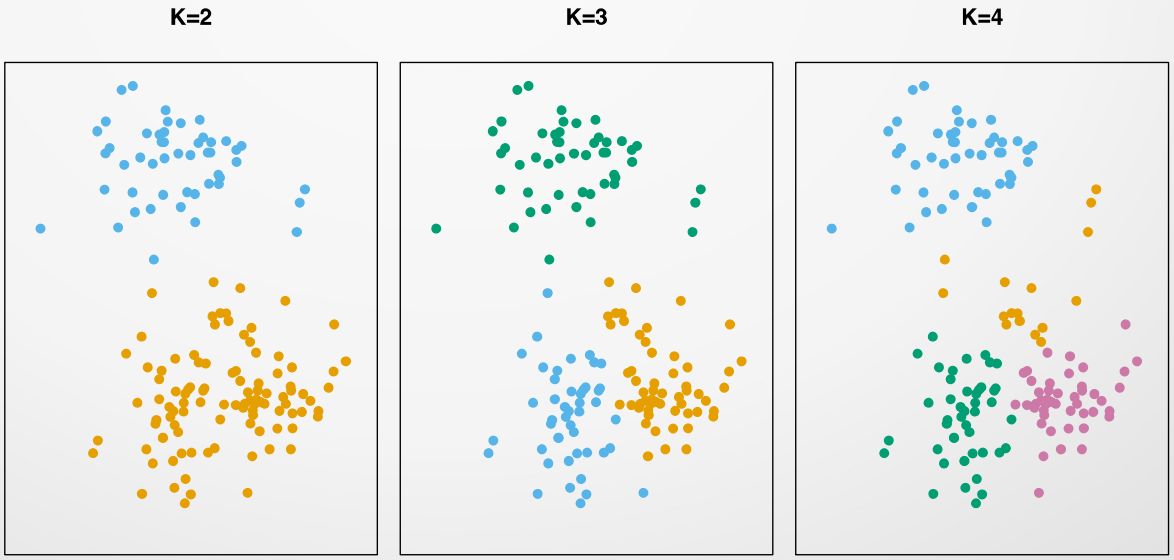
Asignación de puntos (k-means):

- Se mantiene un conjunto de *clusters*
- Cada punto pertenece al grupo de **"más cercano"**

K-MEANS CLUSTERING

K-Means Clustering

- Para realizar un agrupamiento por **K-Means**, debemos especificar en primer lugar la cantidad de clusters (**K**) deseados.
- Entonces el algoritmo de K-means asignará cada una de las observaciones exactamente a uno solo de los K clusters



¿Cómo funciona K-medias?

- Nos gustaría particionar ese conjunto de datos en K grupos

$$C_1 ... C_k$$

  - Cada una de las observaciones pertenecen al menos a uno de los K clusters
  - Los clusters no están solapados (***non-overlapping***), es decir, no hay observaciones que pertenezcan a más de un *cluster*

- El objetivo es tener una mínima **variación dentro del *cluster*,** es decir, los elementos dentro de un *cluster* deberían ser tan similares como sea posible.
- Una forma de lograr esto es reducir al mínimo la suma de todas las distancias euclidianas por pares al cuadrado entre las observaciones en cada *cluster*.

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

Algoritmo K-Means

- **Paso Inicial**: De forma **aleatoria** asignar cada observación a uno de los K clusters

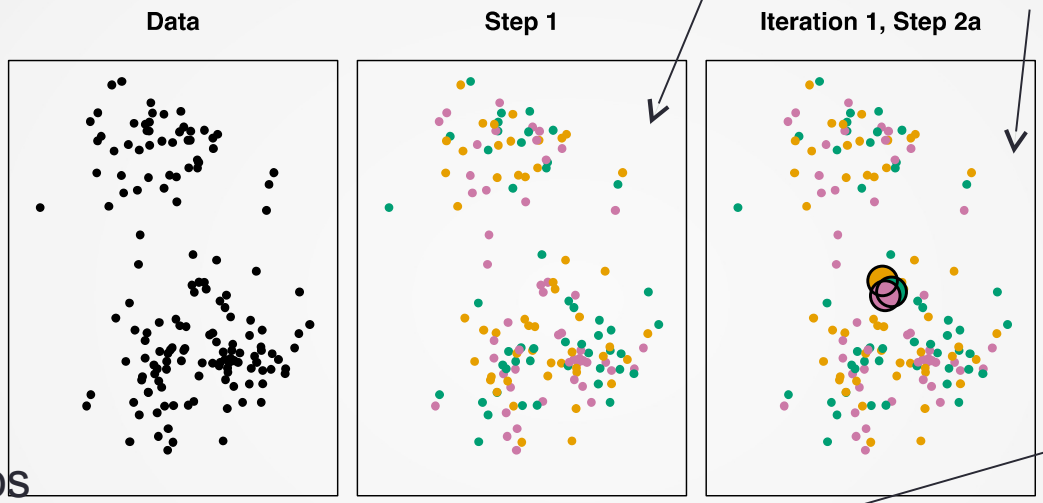
- Iterar hasta que la asignación de clusters deje de modificarse:

  - Para cada uno de los K clusters, **calcular el centroide**. El centroide del k-ésimo cluster es la **media** de las observaciones etiquetadas como K.
  - Asigna cada una de las observaciones al cluster cuyo centroide es más cercano (donde “cercano” se define utilizando la distancia Euclidea)

Ejemplo del algoritmo K-means

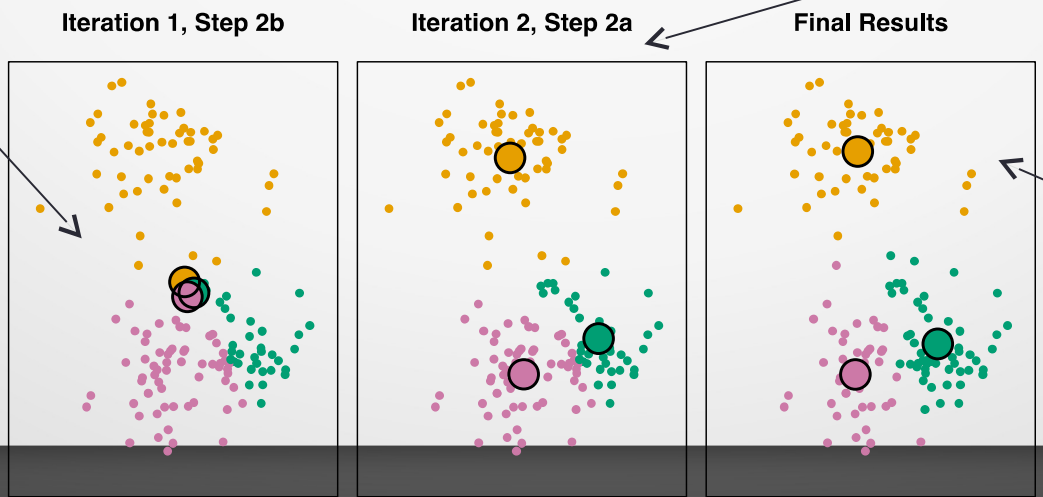
Asigna los puntos al azar

Calcula los centroides para los clusters actuales



Calcula los nuevos centroides para cada cluster

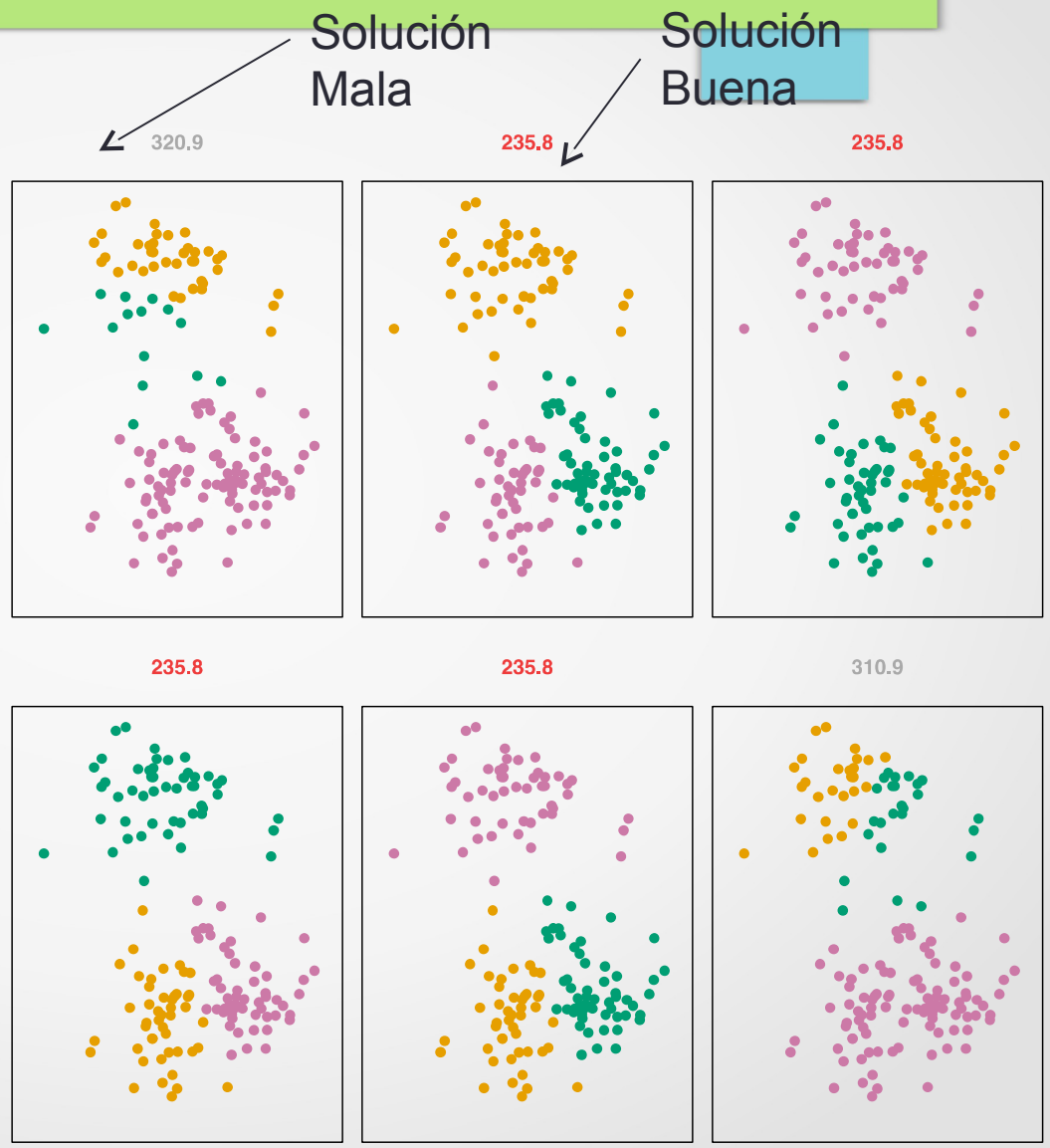
Asigna los puntos al cluster más cercano



No hay más cambios entonces frena

Óptimos locales

- El algoritmo K-means puede atascarse en **"óptimos locales"** y no encontrar la mejor solución
- El resultado va a estar atado a la imputación inicial
- Para encontrar una buena solución tenemos que correr varias veces con distintas configuraciones iniciales.



HIERARCHICAL CLUSTERING

Clustering Jerárquico

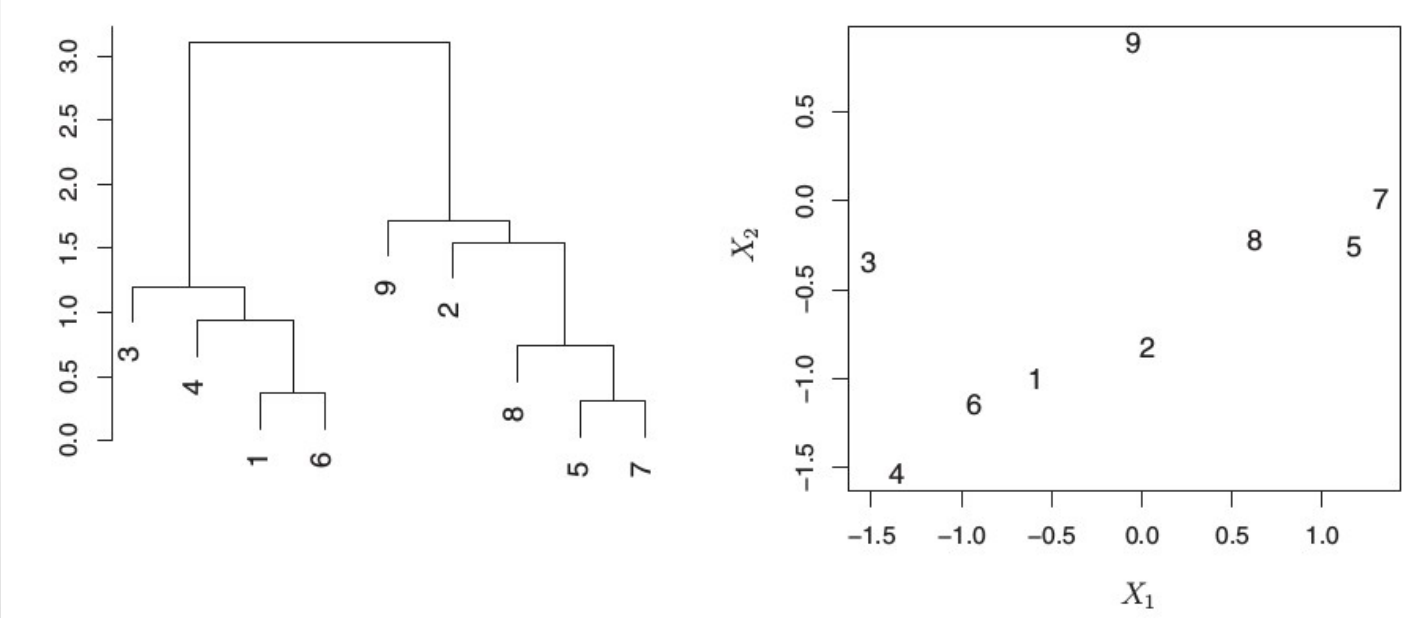
- K-Means requiere elegir un número de clusters
- El problema de seleccionar un K está lejos de ser un tema simple.
- Ahora, si no queremos pasar por el proceso de selección de K podemos utilizar **Clustering Jerárquico**
- Los clusters jerárquicos son construidos a partir de una representación gráfica basada en árboles llamada **Dendrograma**

Clustering Jerárquico

- Clustering **aglomerativos** (*Bottom-Up*).
  - Comienzan el análisis con tantos clusters como observaciones se tienen.
  - A partir de las unidades iniciales se forman grupos, de forma ascendente hasta que queda un solo cluster.
- Clustering **disociativos** (*Top-Down*).
  - Comienzan con un único cluster que engloba a todas las observaciones.
  - Y se realizan repetidas divisiones hasta llegar a tantas agrupaciones como casos tratados.

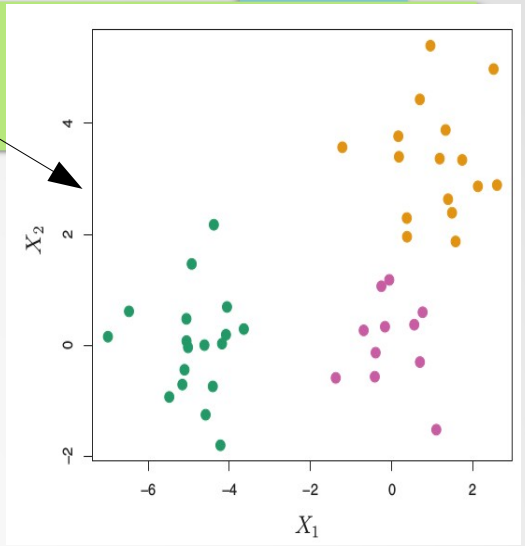
Dendrograma

- Juntamos los puntos más cercanos Ejemplo: 5 y 7
- La altura de fusión (en el eje vertical) indica cuan similares son los puntos
- Luego de ser fusionados, se los trata como una simple observación y el algoritmo continúa.

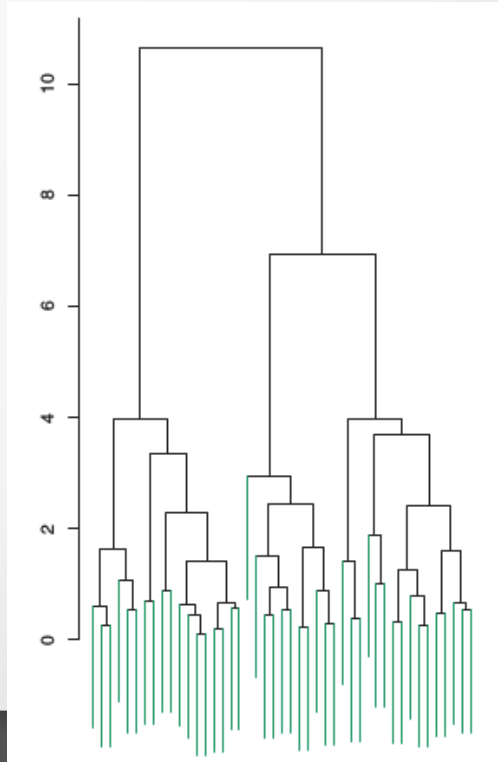


Interpretación

Tenemos 45 observaciones

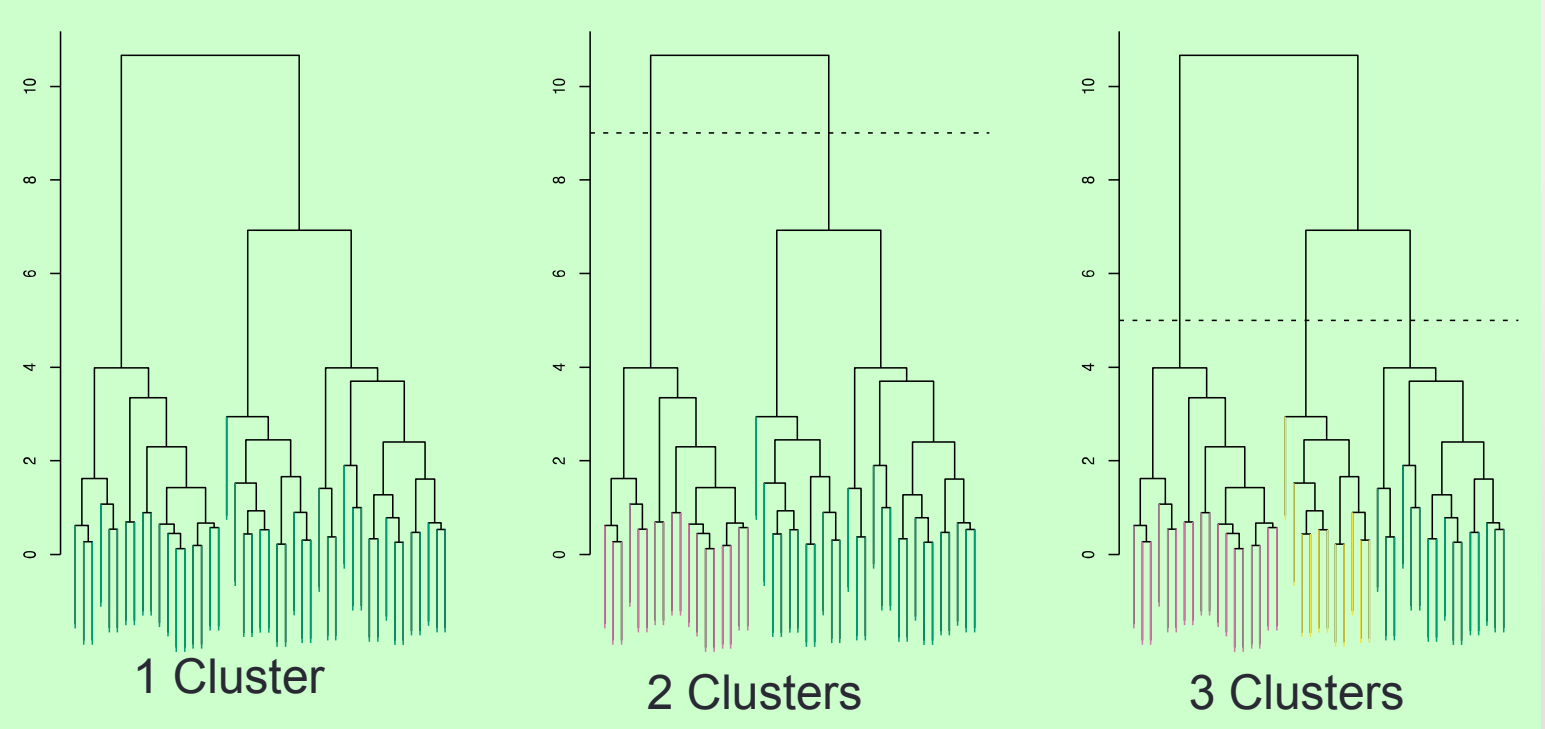


- **Cada una de las hojas** del dendrograma representa una de las 45 observaciones.
- En la parte de **abajo del dendrograma** cada una de las observaciones es una hoja distinta. Sin embargo, cuando nos movemos hacia arriba **las hojas se van fusionando**. Esto corresponde a observaciones que son similares unas a otras.
- A medida que avanzamos más arriba en el árbol, un creciente número de observaciones han fusionado.
- Las dos observaciones que se unieron previamente (más abajo) son las más similares entre si.
- Las observaciones que se fusionan al final, son muy diferentes de las primeras en fusionarse.



Elegir los Clusters

- Para elegir los clusters podemos trazar una línea que cruce el dendrograma.
- Podemos obtener un número de clusters dependiendo de las intersecciones de la línea.

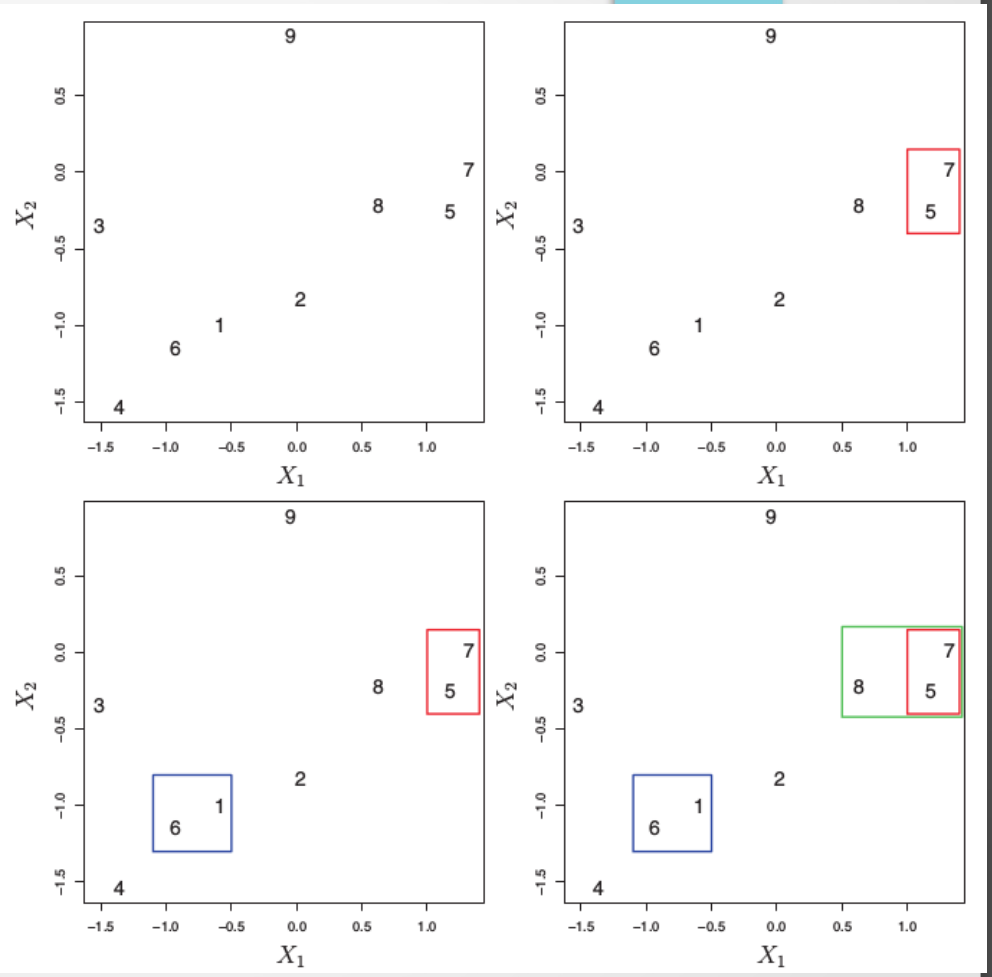


Algoritmo (Enfoque aglomerativo)

- El dendrograma se genera de la siguiente manera:
  - Comienza con cada uno de los puntos como un cluster separado (n clusters)
  - Calcula una medida de **disimilitud** entre los **puntos / clusters**
  - Fusiona dos clusters que son más similares de manera que ahora tenemos n – 1 clusters.
  - A continuación fusiona los siguientes más similares entonces nos quedan ahora n – 2 clusters.
  - Continúa hasta que nos queda un solo cluster.

Un ejemplo

- Inicia con 9 clusters
- Une 5 y 7
- Une 6 y 1
- Une el cluster (5,7) con 8.
- Continua hasta que todas las observaciones están unidas.

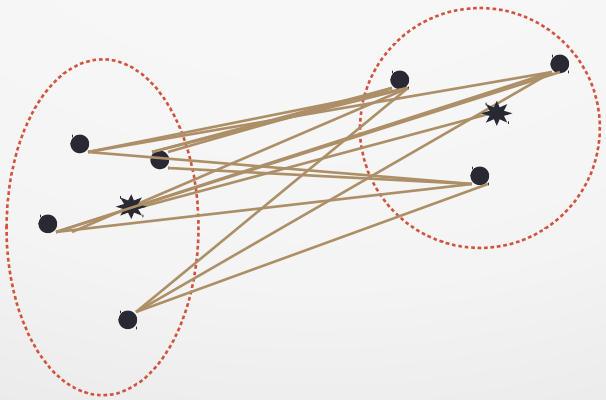


¿Cómo definimos disimilitud?

- La implementación de clustering jerárquico implica resolver un problema obvio que es...
- ¿Cómo hacer para definir la disimilitud o amalgamiento (o *linkage*) entre un cluster cluster (5 , 7) y el punto 8?
- Hay 4 opciones:
  - Complete Linkage
  - Single Linkage
  - Average Linkage
  - Centriod Linkage

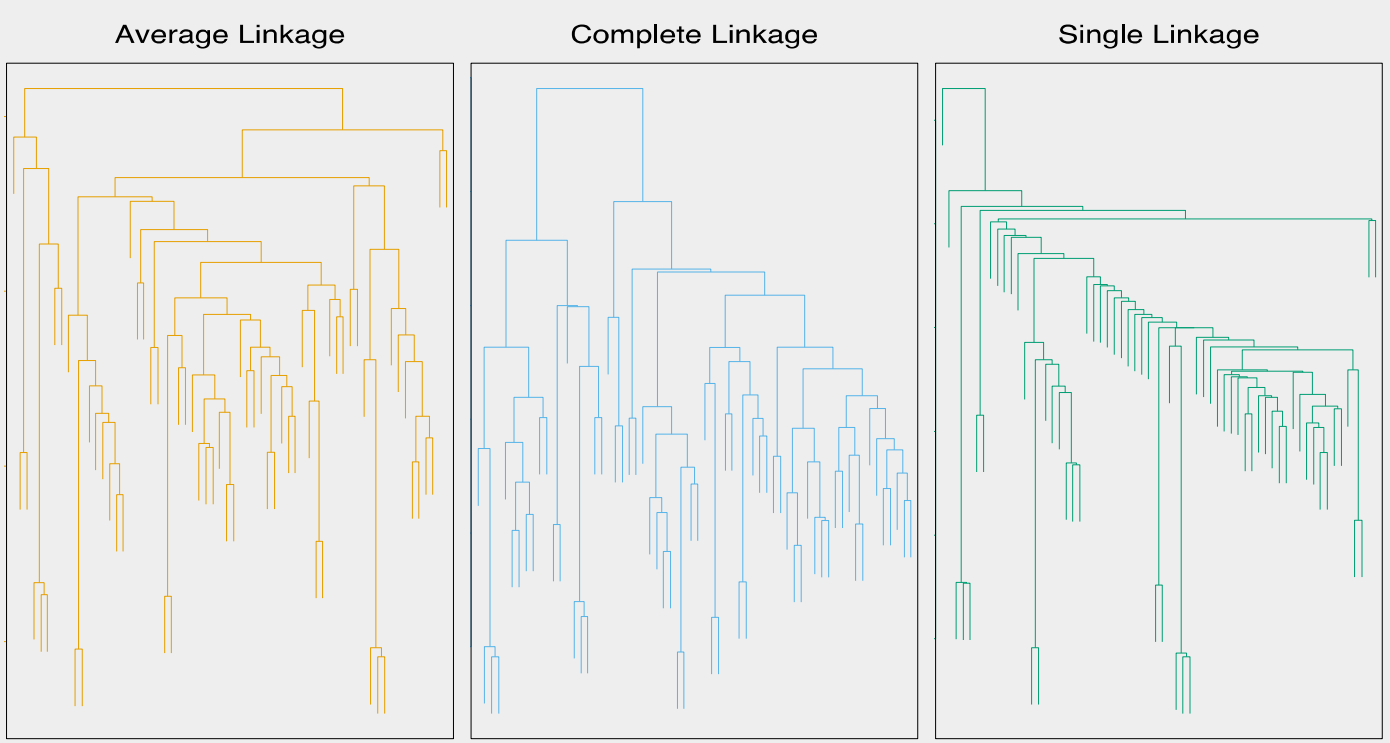
Métodos de *Linkage*: Distancia entre Clusters

- **Complete Linkage:** Distancia máxima o similitud mínima entre observaciones
- **Single Linkage:** Distancia mínima o similitud máxima entre observaciones
- **Average Linkage:** Distancia promedio entre las observaciones
- **Centroid:** Distancias entre centroides de las observaciones.



Importancia del *Linkage*

- Aquí tenemos los tres resultados con los mismos datos.
- La diferencia en los dendrogramas se atribuye al *Linkage* utilizado.
- *Linkage Complete* y *Average* tienden a producir racimos de tamaño uniforme mientras que Single Linkage tiende a producir racimos extendidos para que las hojas individuales se fusionen una a una.



Criterio de Distancia y Similitud

Distancia

$$d(C_i, C_j) = \min_{\substack{x_l \in C_i \\ x_m \in C_j}} \{d(x_l, x_m)\} \quad l = 1, \ldots, n_i \; ; \; m = 1, \ldots, n_j$$

Similitud

$$s(C_i, C_j) = \max_{\substack{x_l \in C_i \\ x_m \in C_j}} \{s(x_l, x_m)\} \quad l = 1, \ldots, n_i \; ; \; m = 1, \ldots, n_j$$

Ejemplo: *Single Linkage*

Matriz Inicial de distancias

	A	B	C	D	E	F	G
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

Nivel K=1

Nivel K=2

	A	B	(C,E)	D	F	G
A	0					
B	2,15	0				
(C,E)	0,7	1,38	0			
D	1,07	1,14	0,29	0		
F	1,16	1,01	0,41	0,22	0	
G	1,56	2,83	1,86	2,04	2,05	0

Nivel K=3

	A	B	(C,E)	(D,F)	G
A	0				
B	2,15	0			
(C,E)	0,7	1,38	0		
(D,F)	1,07	1,01	0,29	0	
G	1,56	2,83	1,86	2,04	0

Nivel K=4

	A	B	((C,E),(D,F))	G
A	0			
B	2,15	0		
((C,E),(D,F))	0,7	1,01	0	
G	1,56	2,83	1,86	0

Nivel K=5

	(A,((C,E),(D,F)))	B	G
(A,((C,E),(D,F)))	0		
B	1,01	0	
G	1,56	2,83	0

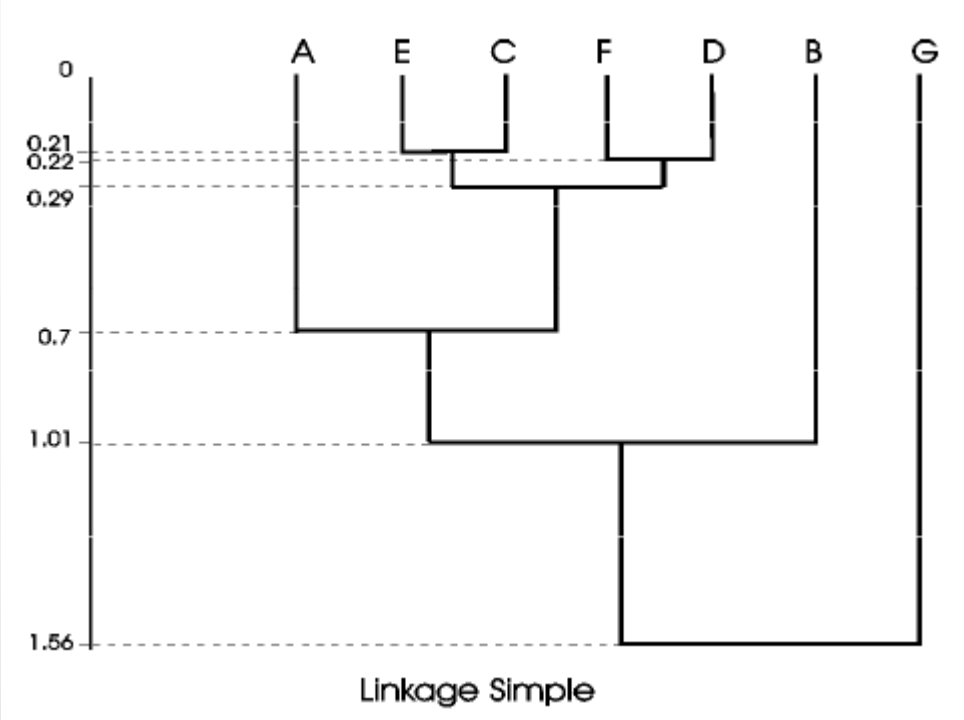
Nivel K=6

	(B,(A,((C,E),(D,F))))	G
(B,(A,((C,E),(D,F))))	0	
G	1,56	0

Ejemplo: *Single Linkage*

Dendrograma Resultante en K=6

En la etapa K-ésima quedan formados n – K clusters



Medidas de distancia/similitud

Variables Numéricas

- Distancia Euclidea
- Distancia de Manhattan
- Distancia Minkowski

Variables Binarias: Coeficiente de Jaccard

Variables Categóricas

Variables Numéricas

Distancia Euclidea

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2},$$

Distancia de Manhattan

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

Distancia Minkowski

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$$

Los requisitos para una función de distancia son:

- Las distancias son siempre **no negativas**
- Sólo la distancia entre un punto y sí mismo es 0.
- La distancia es simétrica; no importa en qué orden se tiene en cuenta los puntos al calcular su distancia.
- Las medidas de distancia obedecen a la **desigualdad triangular**; la distancia de X a Y a Z nunca es menor que la distancia que va desde X a Z directamente.

Coeficiente de Jaccard

A contingency table for binary variables.

		object j		
		1	0	sum
object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Distancia de dos variables binarias

$$d(i, j) = \frac{r+s}{q+r+s}.$$

Similitud

$$sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j).$$

sim(i, j) es conocido como ***coeficiente de Jaccard***

Variables Categóricas

Distancia de dos variables categóricas

donde:

p: es el total de variables

m: es el total de coincidencias

$$d(i, j) = \frac{p - m}{p},$$

A sample data table containing variables of mixed type.

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

Ejemplo para *test-1*

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Evaluación: Coeficiente de Silueta

- El **coeficiente de Silueta** es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de *clustering*.
- El objetivo de Silueta es identificar cuál es el número óptimo de agrupamientos.

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

Donde:

- **a** es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del cluster al que pertenece i
- **b** es la distancia mínima a otro cluster que no es el mismo en el que está la observación i.

  Ese cluster es la segunda mejor opción para i y se lo denomina vecindad de i.

Evaluación: Coeficiente de Silueta

- El valor de s(i) puede ser obtenido combinando los valores de a y b como se muestra a continuación:

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

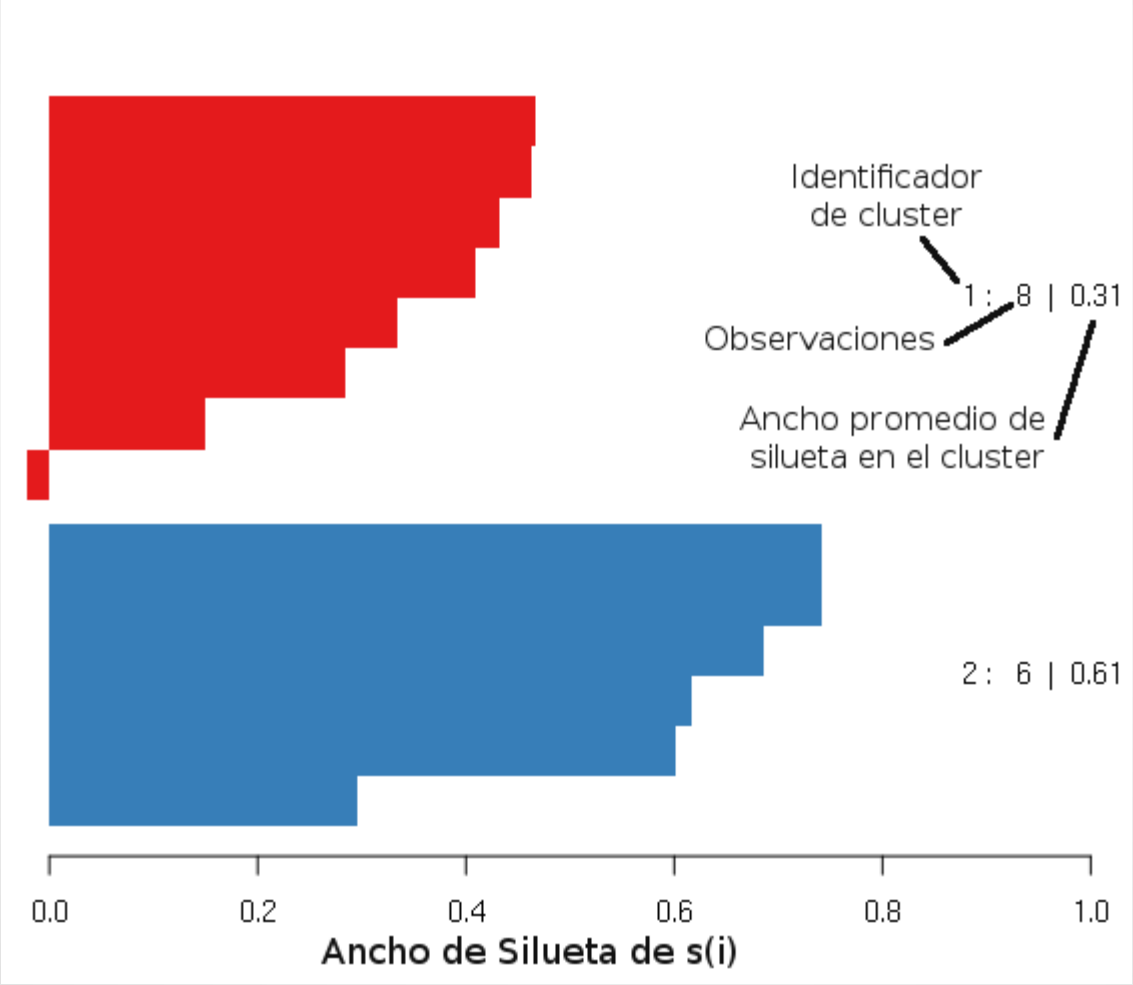
- El coeficiente de Silueta es un valor comprendido entre -1 y 1.

Resumiendo:

- s(i) ≈ 1, la observación i está bien asignada a su cluster
- s(i) ≈ 0, la observación i está entre dos cluster
- s(i) ≈ −1, la observación i está mal asignada a su cluster

Evaluación: Coeficiente de Silueta

Gráfico de ancho de Silueta



REFERENCIAS

- Basado principalmente en: Clustering Chapter 10. IOM 530: Intro. to Statistical Learning. (Traducción Libre)

- Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

- J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org