Análisis Inteligente de Datos

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Profesora: Mónica Cantoni

Ayudantes: Cecilia Oliva - Fabiana Rossi - Pamela Pairo

Clase 06 – Reglas de Asociación - Clustering - Kmeans

Agenda

- **Reglas de asociación**
- **Clustering**

Reglas de asociación

Reglas de asociación

Las reglas de asociación permiten encontrar patrones comunes en los elementos de grandes conjuntos de datos.

Una de las principales aplicaciones de esta técnica es el análisis del carrito de compras. Mediante el cual se pude identificar los productos que se compra de forma conjunta en una tienda.

Uno de los ejemplos más citado sobre los resultados de los análisis del carrito de compra es caso *"cerveza y pañales"*.

La historia es que una gran cadena observó que los clientes que compraban cerveza y pañales al mismo tiempo.

Posiblemente no sea más que una leyenda urbana de la ciencia de datos.

Sin embargo es uno de los ejemplos más reconocidos de los resultados que se pueden obtener con este tipo de análisis.

4

Introducción a las Reglas de asociación

Introducción a las reglas de asociación

Las reglas de asociación generalmente se escriben de la forma:

$$\{Antecedente\} \Longrightarrow \{Consecuente\}$$

Esto indica que existe una relación entre los clientes que compran el antecedente y el consecuente en la misma transacción. La fuerza y el sentido de la relación se mide con diferentes indicadores: el soporte, la confianza y la mejora de la confianza.

5

Reglas de asociación

- Queremos ver cómo se asocian (o no) los artículos en compras
- Datos:
  - compras realizadas en un supermercado
  - Sólo datos de cuáles productos fueron comprados

Datos: versión canastas

- ¿Qué podemos ver en estas canastas de compra?
- ¿Cómo se asocian los artículos?

id	artículos comprados
1	Pan, leche, papas fritas
2	Cerveza, huevos, pañales, pan
3	Cerveza, gaseosas, pañales, leche
4	Cerveza, pan, pañales, leche, papas fritas
5	Gaseosas, pan, pañales, leche
6	Cerveza, pan, pañales, leche, mostaza
7	Gaseosas, pan, pañales, leche

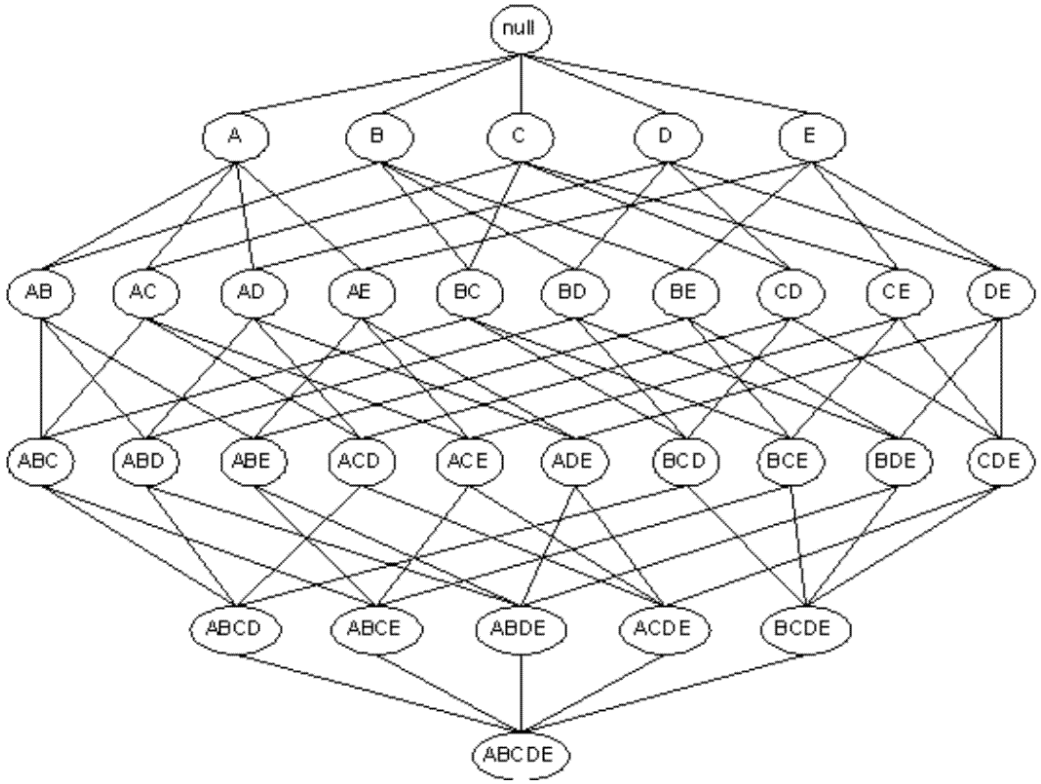
Datos: versión matricial

- ¿Qué podemos ver en estas canastas de compra?
- ¿Cómo se asocian los artículos?

Pan	Leche	Papas fritas	Mostaza	Cerveza	Pañales	Huevos	Gaseosas
1	1	1	0	0	0	0	0
1	0	1	0	1	1	1	0
0	1	0	0	1	1	0	1
1	1	0	0	1	1	0	0
1	1	0	1	0	1	0	1
1	1	0	0	1	1	0	0
1	1	1	0	0	1	0	1

Conjuntos de ítems

- Hay 2^n conjuntos de ítems posibles
- No se puede hacer por fuerza bruta



Soporte

- El soporte es la frecuencia relativa con la que se observa la regla. Es decir, un soporte de 0.15 indica que el antecedente y el consecuente se observan a la vez en el 15% de las transacciones. Este indicador mide la fuerza de la regla. Al ser un porcentaje, los posibles valores del soporte se encuentran entre 0 y 1.

10

Soporte

- soporte (A) = # transacciones donde aparece A / # total de transacciones
  - soporte({huevos}) = 1/7
  - soporte({pan}) = 6/7
  - soporte({pan, leche}) = 5/7
  - soporte({pan, leche, cerveza}) = 2/7
- Se buscan conjuntos de ítems con soporte entre dos umbrales
- Observación: el soporte baja a medida que crece la cantidad de ítems del conjunto

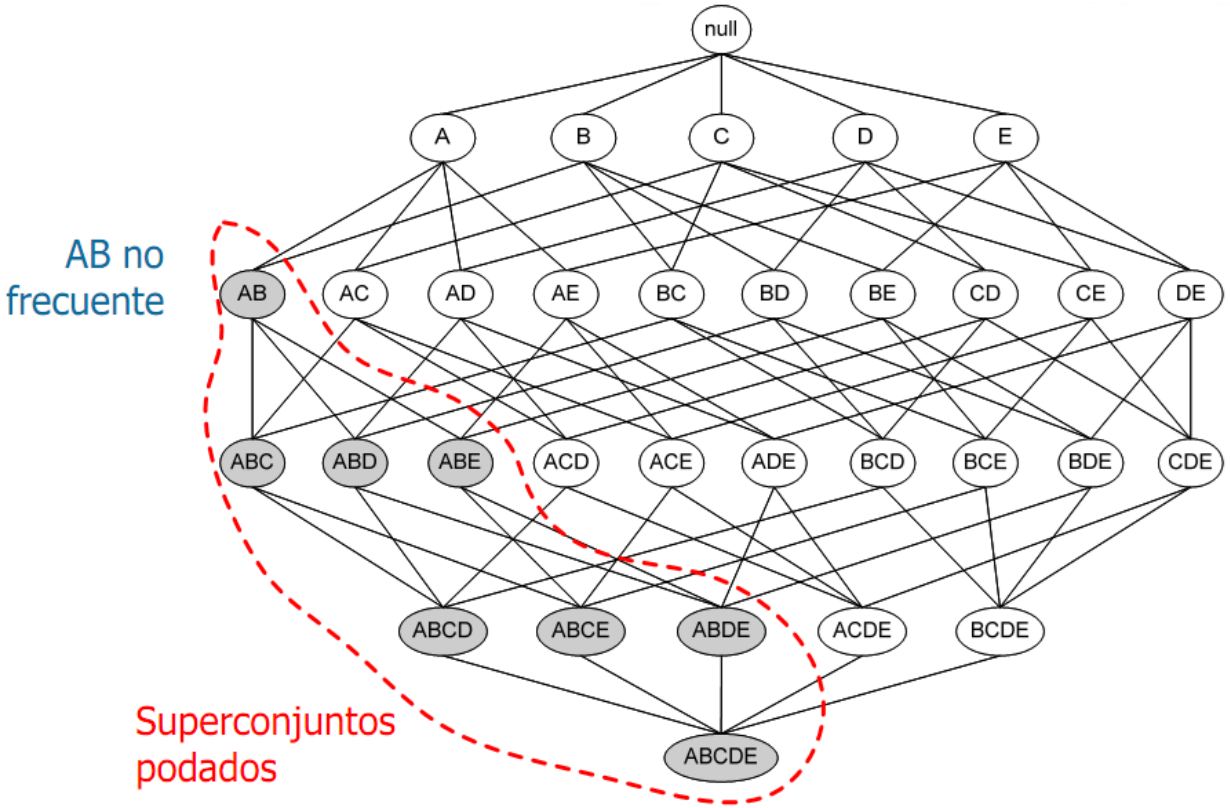
Propiedad *a priori*

- Si un conjunto de ítems no es frecuente, también lo son todos sus subconjuntos

- Puedo eliminar (podar) todos los conjuntos que no son frecuentes y todos sus subconjuntos

12

Poda de conjuntos



Reglas de asociación

- Regla: X → Y
  - X e Y son conjuntos de ítems
- Ejemplos:
  - {pan} → {huevos}
  - {pan, leche} → {huevos}
- Ojo: "→" significa co-ocurrencia, no causalidad

Confianza

La confianza es el porcentaje de las transacciones en las que aparece el antecedente en la que también aparece el consecuente. Lo que mide este indicador es la fiabilidad de la regla.

Matemáticamente se puede obtener utilizando la expresión

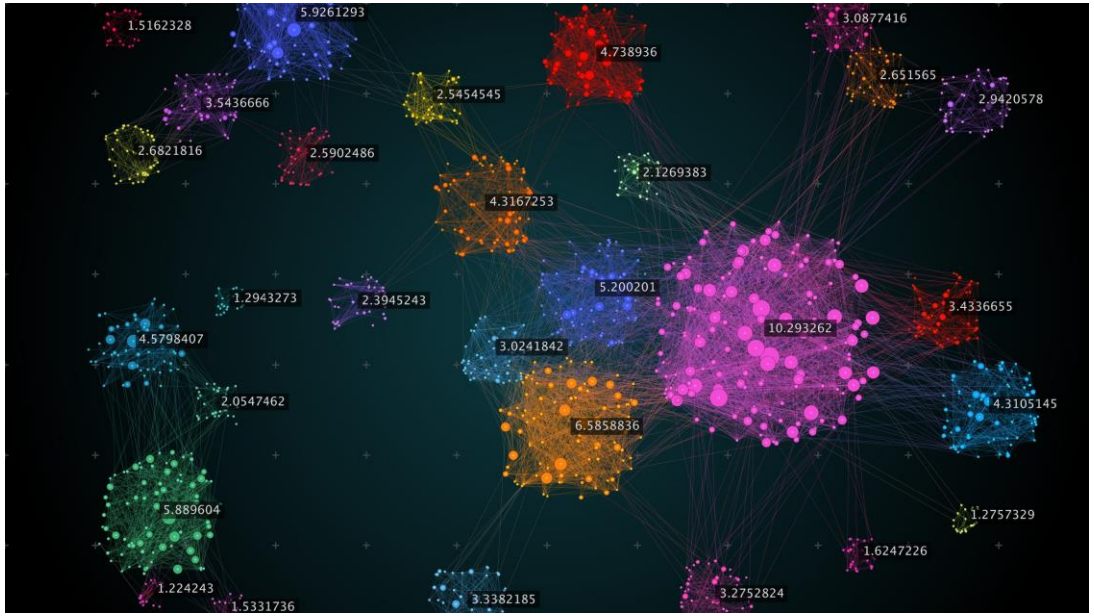
$$\boldsymbol{conf}(\{\boldsymbol{Antecedente}\} \Longrightarrow \{\boldsymbol{Consecuente}\}) = \frac{\boldsymbol{soporte}(\{\boldsymbol{Antecedente}, \boldsymbol{Consecuente}\})}{\boldsymbol{soporte}(\{\boldsymbol{Antecedente}\})}$$

En donde $\boldsymbol{conf}(\{\boldsymbol{Antecedente}\} \Longrightarrow \{\boldsymbol{Consecuente}\})$ es la confianza de los registros en los que aparece a la vez el antecedente y consecuente.

Confianza

- confianza(X → Y) = # transacciones donde aparece X e Y / # transacciones donde aparece X = soporte({X,Y}) / soporte({X})
  - soporte({pan}) = 6/7
    soporte({leche}) = 6/7
  - soporte({pan, leche}) = 5/7
  - confianza({pan} → {leche}) = 5/6
  - confianza({leche} → {pan}) = 5/6
- Tarea: Buscar las reglas con confianza entre dos umbrales

Clustering


1.5162328
5.9261293
4.738936
3.0877416
2.651565
2.9420578
3.5436666
2.5454545
2.6821816
2.5902486
2.1269383
4.3167253
1.2943273
2.3945243
5.200201
3.4336655
10.293262
4.5798407
3.0241842
2.0547462
6.5858836
4.3105145
5.889604
1.2757329
1.6247226
1.224243
1.5331736
3.3382185
3.2752824

Análisis de Conglomerados (Clustering) Introducción

- También llamado análisis de *clúster.*
- Es una técnica diseñada para clasificar distintas observaciones en grupos de tal forma que:

1. Cada grupo (conglomerado o *clúster*) sea homogéneo respecto a las variable utilizadas para caracterizarlo, que cada observación contenida en él sea parecida a todas las que estén incluidas en ese grupo.
2. Que los grupos sean lo más distintos posible unos de otros respecto a las variables consideradas.

- Los grupos son desconocidos a priori y es necesario derivarlos de las observaciones. Por ejemplo en el análisis discriminante las observaciones ya estaban previamente clasificadas en uno o más grupos.

Secuencia lógica del Análisis de Conglomerados

1. Inicialmente el investigador dispone de n observaciones (individuos, empresas, etc.) de las cuales tiene información sobre *k* variables.
2. Se establece un indicador que indique en qué medida cada par de observaciones se parece entre sí. A esta medida se la denomina distancia o similaridad.
3. Hacer grupos con aquellas observaciones que más se parezcan entre sí, de acuerdo con una medida de similaridad calculada anteriormente.
4. Elegir entre los dos tipos de análisis jerárquico o no jerárquico y el método de conglomerado para el tipo de análisis elegido (centroide o vecino más cercano en el caso de conglomerado jerárquico).
5. Finalmente el investigador debe describir los grupos que obtuvo y comparar los unos con los otros. Para ello debe observar que valores promedio toman las *k* variables utilizadas en el análisis de conglomerados en cada uno de los *g* grupos obtenidos $(g \leq n)$.

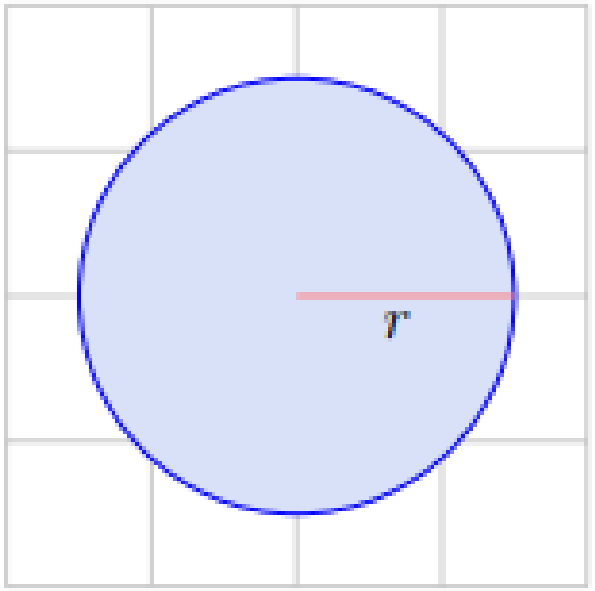
Función de distancia

- Atributos numéricos (altura, facturación, latitud, longitud): Distancia euclidiana
- Atributos ordinales (chico, mediano, grande...): Se pasa a numérico
- Atributos nominales (verde, rojo, azul): *Matching*

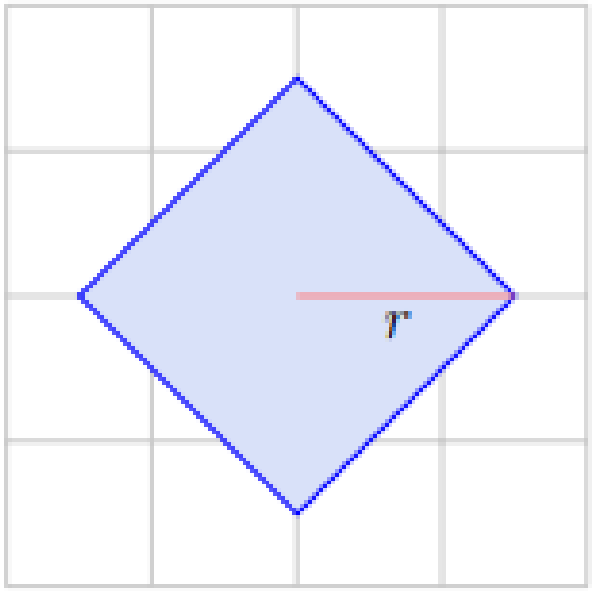
Medidas de Similaridad para variables métricas

- Distancia euclídea
- $D_{i,j} = \sqrt{\sum_{p=1}^{k}(x_{ip} - x_{jp})^2}$
- Distancia euclídea al cuadrado
- $D_{i,j} = \sum_{p=1}^{k}(x_{ip} - x_{jp})^2$
- Distancia de Chebyshev
- $d_{\infty}(x, y) = \max_{1 \leq i \leq k} |x_i - y_i|$
- Distancia de Minkowski
- $D_{i,j} = [\sum_{p=1}^{k} |x_{ip} - x_{jp}|^2]^{1/n}$
- Si n = 2 es la distancia euclídea
- Distancia de Manhattan
- $D_{i,j} = \sum_{p=1}^{k} |x_{ip} - x_{jp}|$

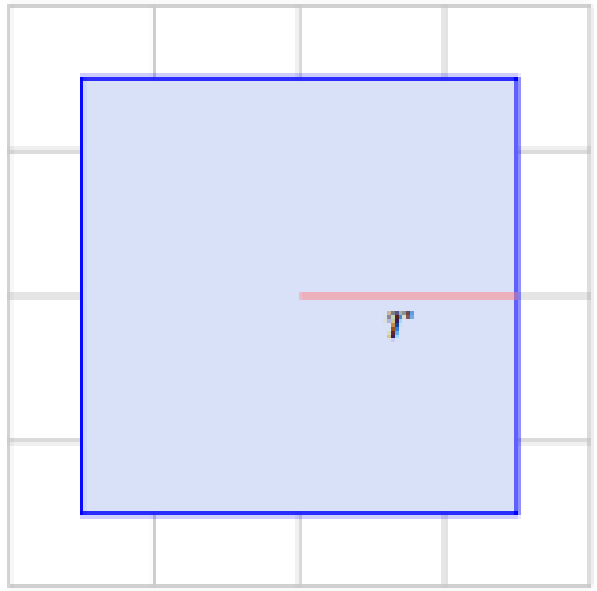
Interpretación geométrica de la distancia



(a) Euclídea



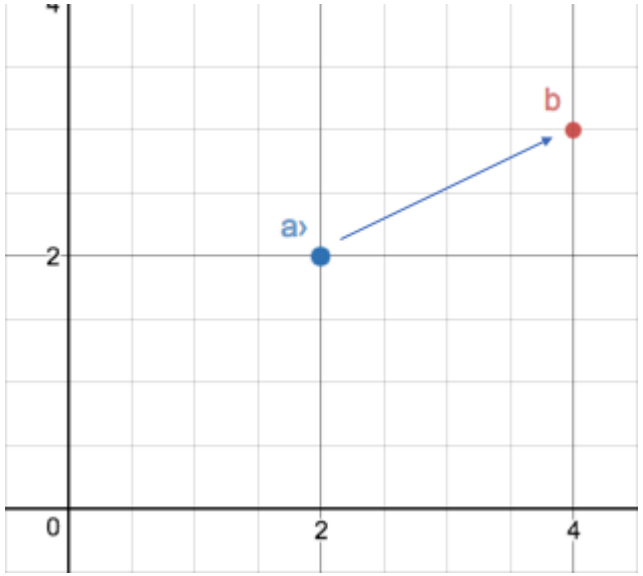
(b) Manhattan



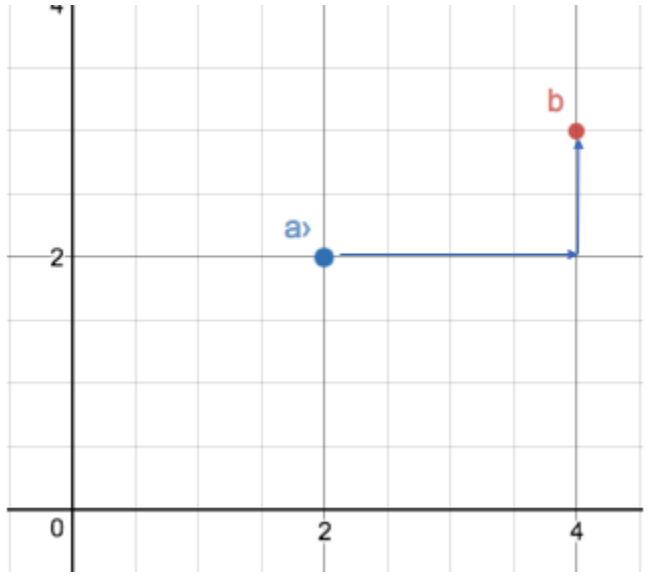
(c) Chebyshev

Distancias

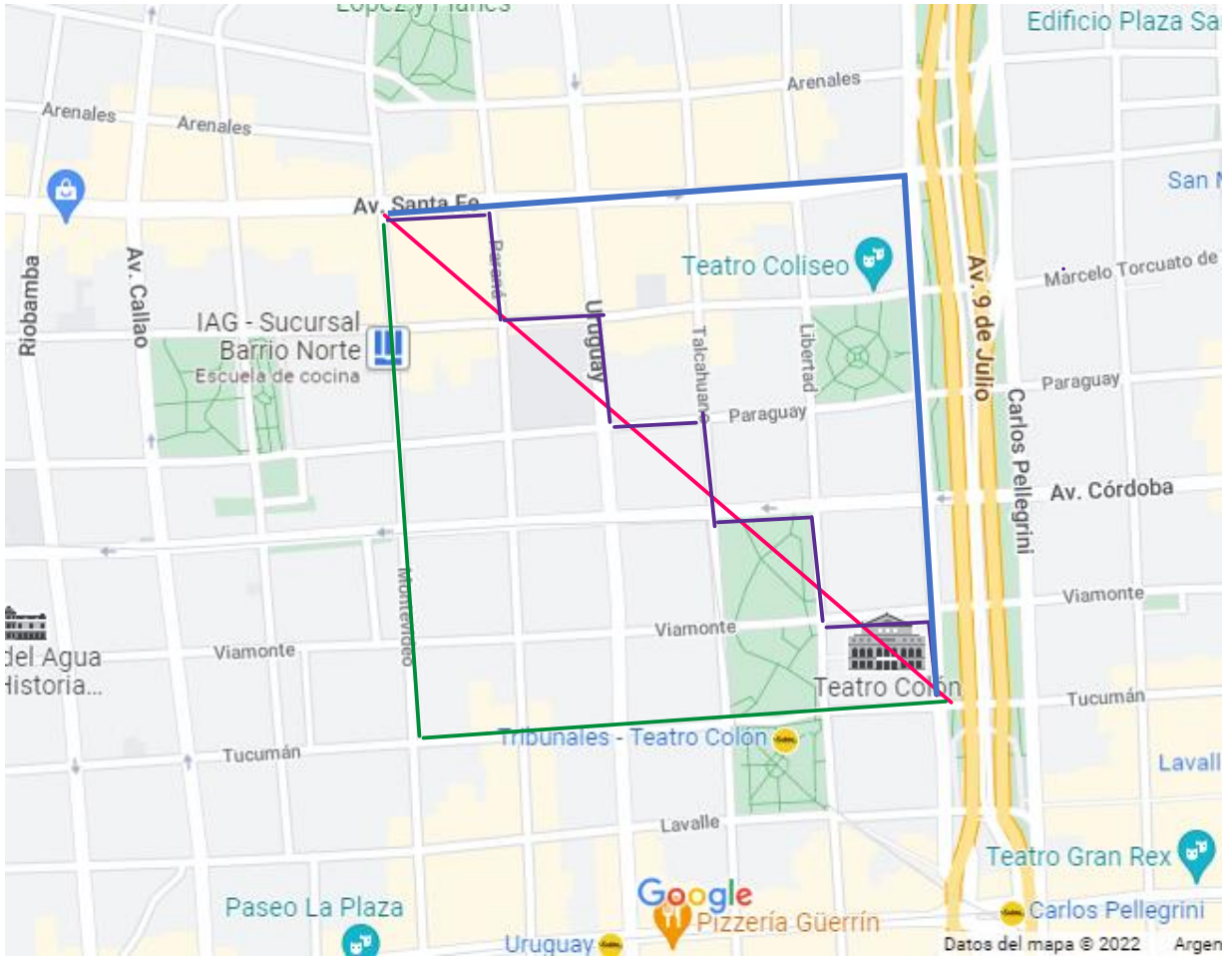
Distancia euclídea



Distancia de Manhattan



Distancia de Manhattan



Ejemplo de distancia

Ejemplo 10.1. En la Tabla 10.2 se consideran las observaciones correspondientes a tres nadadores para los cuales se han registrado los tiempos empleados para correr cada uno de los cuatro tramos en los que se dividió la carrera.

Nadador	x_1	x_2	x_3	x_4
A	10	10	13	12
B	12	12	14	15
C	11	10	14	13

Tabla 10.2: Tiempos para los nadadores

- **Distancias euclídeas**

$$d_2(A,B) = \sqrt{(10-12)^2 + (10-12)^2 + (13-14)^2 + (12-15)^2} = \sqrt{18},$$
$$d_2(A,C) = \sqrt{(10-11)^2 + (10-10)^2 + (13-14)^2 + (12-13)^2} = \sqrt{3},$$
$$d_2(B,C) = \sqrt{(12-11)^2 + (12-10)^2 + (14-14)^2 + (15-13)^2} = 3.$$

- **Distancias de Manhattan**

$$d_1(A,B) = |10-12| + |10-12| + |13-14| + |12-15| = 8,$$
$$d_1(A,C) = |10-11| + |10-10| + |13-14| + |12-13| = 3,$$
$$d_1(B,C) = |12-11| + |12-10| + |14-14| + |15-13| = 5.$$

Medidas de Similaridad

Ejemplo: Inversión en publicidad y ventas de una empresa hipotética

Valores expresados en unidades monetarias

Agrupación intuitiva:

E1 y E2

E3 y E4

E5 y E6

E7 y E8

Empresa	Inversión en publicidad	Ventas
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27

Matriz de distancias euclídeas para los datos del ejemplo

	1	2	3	4	5	6	7	8
1	0							
2	5.66	0						
3	13.42	8.25	0					
4	15.52	11.00	3.61	0				
5	29.00	33.24	37.00	36.25	0			
6	34.37	38.01	40.61	39.29	7.07	0		
7	32.65	34.79	35.13	33.00	15.00	11.18	0	
8	38.01	40.16	40.31	38.05	17.72	12.00	5.39	0

Medidas de Similaridad

Paso 2

Se fusionan E3 y E4

Se calcula la matriz de distancias

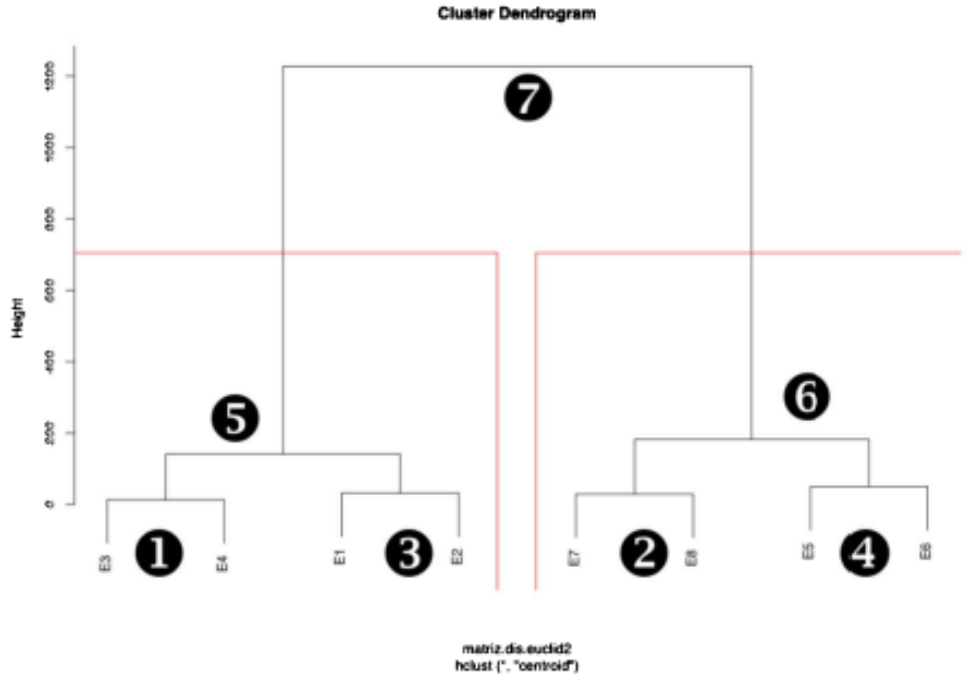
Se observa la menor distancia

Se fusionan los elementos….

Empresa	Inversión en publicidad	Ventas
E1	16	10
E2	12	14
E3-4	11	23,5
E5	45	10
E6	50	15
E7	45	25
E8	50	27

Dendograma

Figura 3.4.: Dendograma



Métodos

- Vecino más cercano
- La distancia entre los grupos se obtiene calculando la distancia entre los centroides
- Vecino más lejano
- La distancia entre los grupos se mide por la distancia entre sus miembros más alejados
- Vinculación promedio
- La distancia promedio entre todos los pares de observaciones que pueden formarse tomando un miembro de un grupo y otro miembro de otro grupo

Agrupamiento no jerárquico o partitivo K-means

Un dato de entrada es el número de k conglomerados deseados.

Pasos

1) se divide a los objetos en k subconjuntos no vacíos.

¿Cómo se eligen los subconjuntos k?

Se eligen aleatoriamente los centros.

Se toman los centros de los grupos como k puntos más alejados entre sí.

Se utiliza información previa disponible para elegir los k centros.

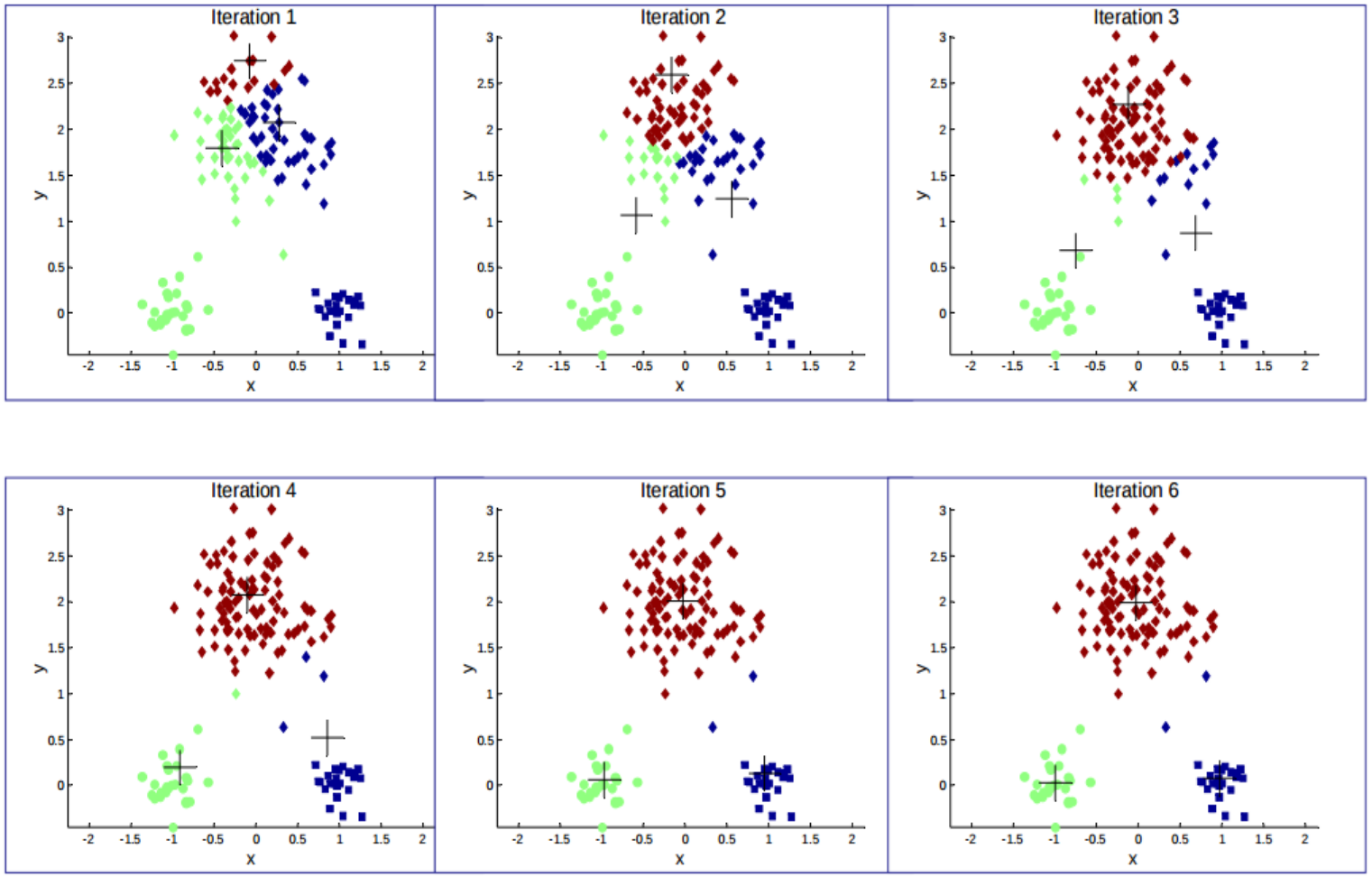
2) Se asigna cada elemento a una de los k clúster cuando la distancia al centroide de ese lugar es mínima.

La asignación es secuencial, cada vez que se reasigna un elemento, se calcula el centroide del nuevo clúster.

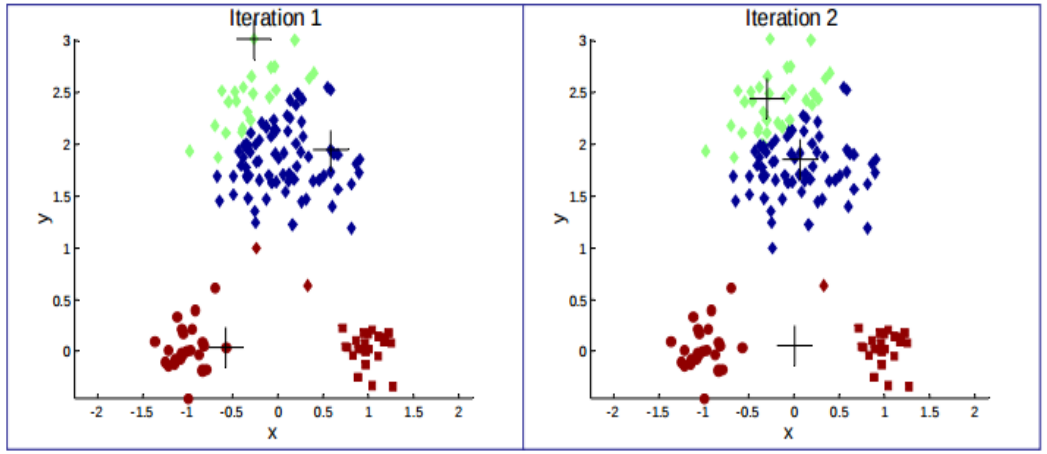
3) Se asigna cada objeto al centroide del clúster más cercano.

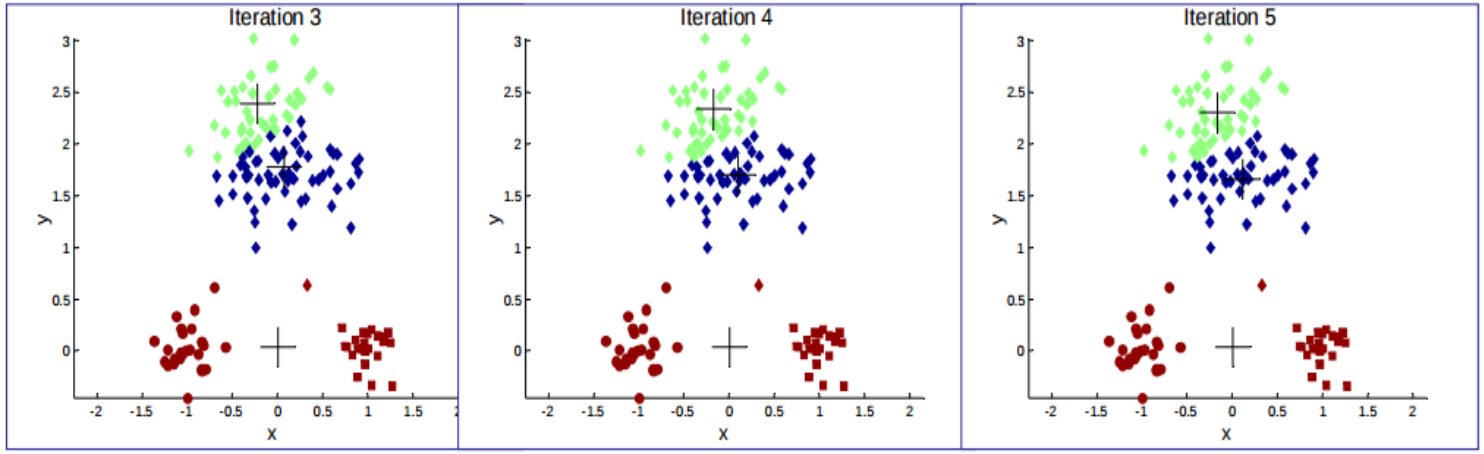
4) Se vuelve al paso 2) y se repite hasta que no hay necesidad de realizar nueva asignación.

Algoritmo k-means



Cuidado con los centroides iniciales





K-Means

- **Ventajas**
- Es relativamente eficiente
- Es computacionalmente rápido
- Se trabaja bien con datos faltantes (missing values)

- **Desventajas**
- Sólo se aplica si la media está definida
- Es sensible a la presencia de outliers.
- Es necesario especificar el número de conglomerados a priori
- No se satisface el criterio de optimización en forma global, en general converge al óptimo local

¿Cómo elegir la medida de proximidad de distancia?

- Depende de la naturaleza de los datos
- Puede resultar conveniente discretizar alguna variable continua considerando ciertos intervalos.
- Formas de hacer la subdivisión: ejemplo en estudios médicos podría hacerse la subdivisión en función del riesgo, en un estudio de mercado podría estar en función de los niveles de consumo.
- La selección de la distancia y de la técnica pueden cambiar la composición final de los clústeres.

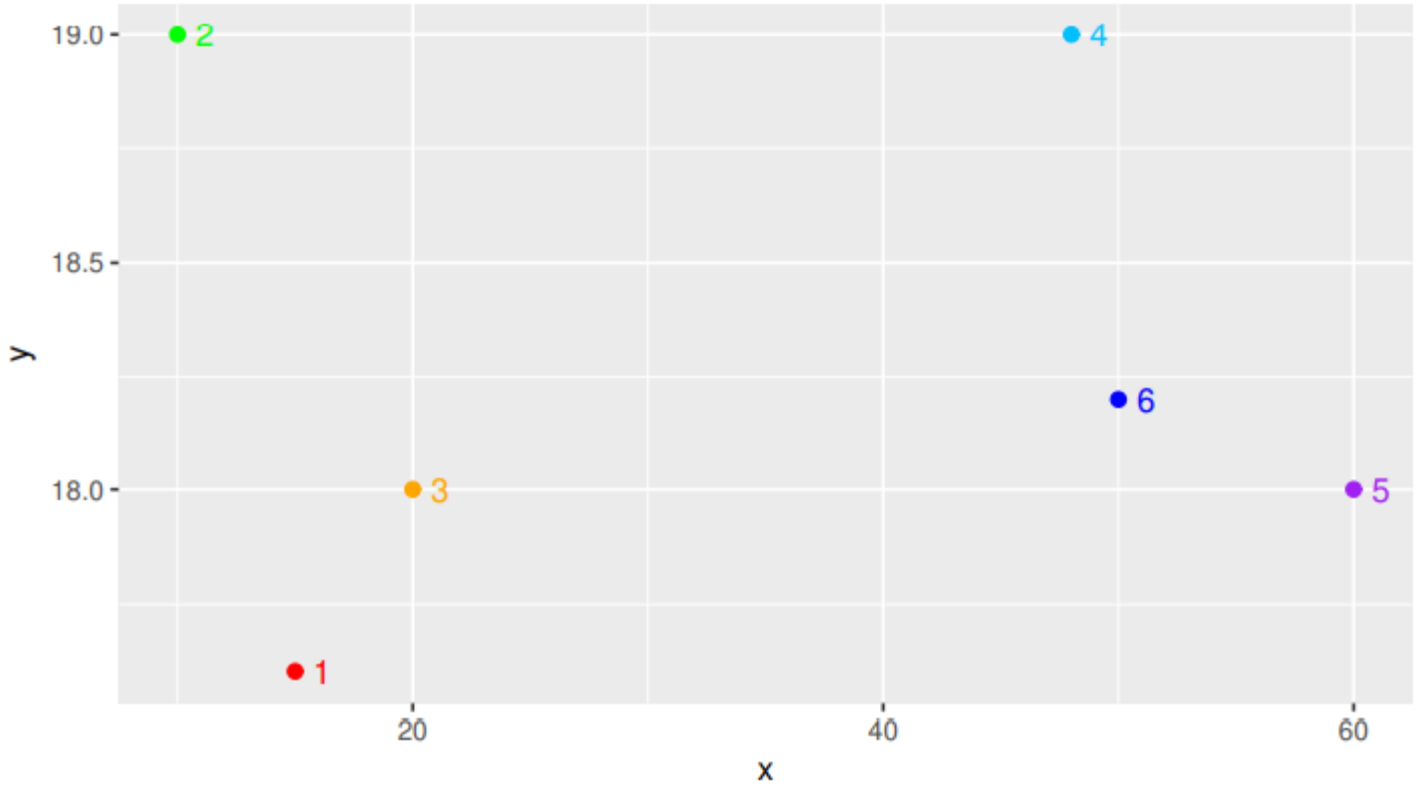
¿Cómo tratar los valores perdidos o missing data?

- Utilizar los registros completos. Pero puede reducir la cantidad de información.
- No debe utilizarse la técnica de completar datos faltantes con la media o la mediana en el análisis de clúster.

Ejemplo:

x	y
15	17.6
10	19.0
20	18.0
48	19.0
60	18.0
50	18.2

Elegimos como centroides a los puntos más alejados
C1 = (10; 19) y C2 = (60;18)



Ejemplo:

Punto	Distancia a c_1	Distancia a c_2	*Cluster* asignado
1	5.19	45.00	1
2	0.00	50.01	1
3	10.05	40.00	1
4	38.00	12.04	2
5	50.01	0.00	2
6	40.01	10.00	2

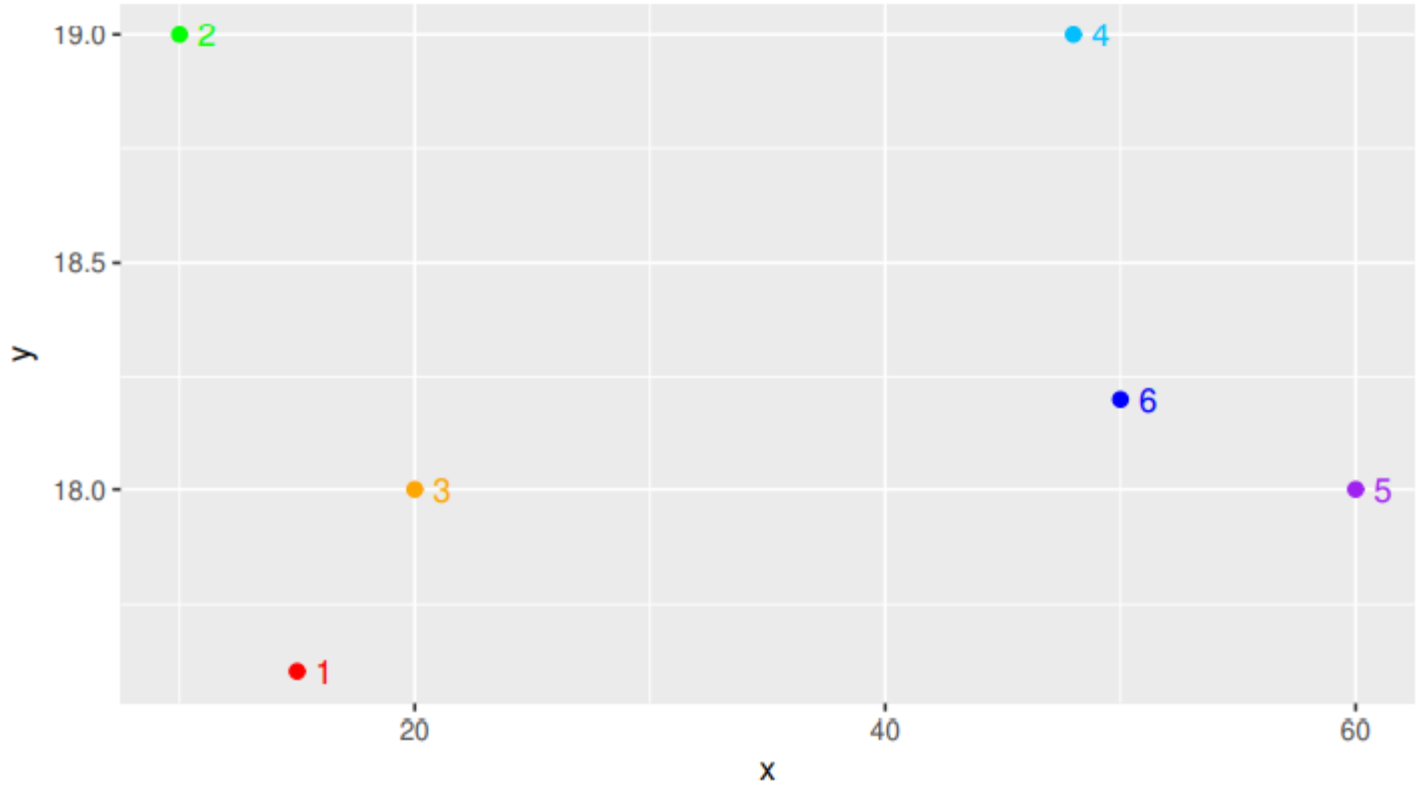
Se calculan los centroides de los primeros tres puntos y de los últimos tres c3 = (15, 18.2) y c4 = (52.67, 18.4).

Punto	Distancia a c_3	Distancia a c_4	*Cluster* asignado
1	0.60	37.68	1
2	5.06	42.67	1
3	5.00	32.67	1
4	33.01	4.71	2
5	45.00	7.34	2
6	35.00	2.68	2

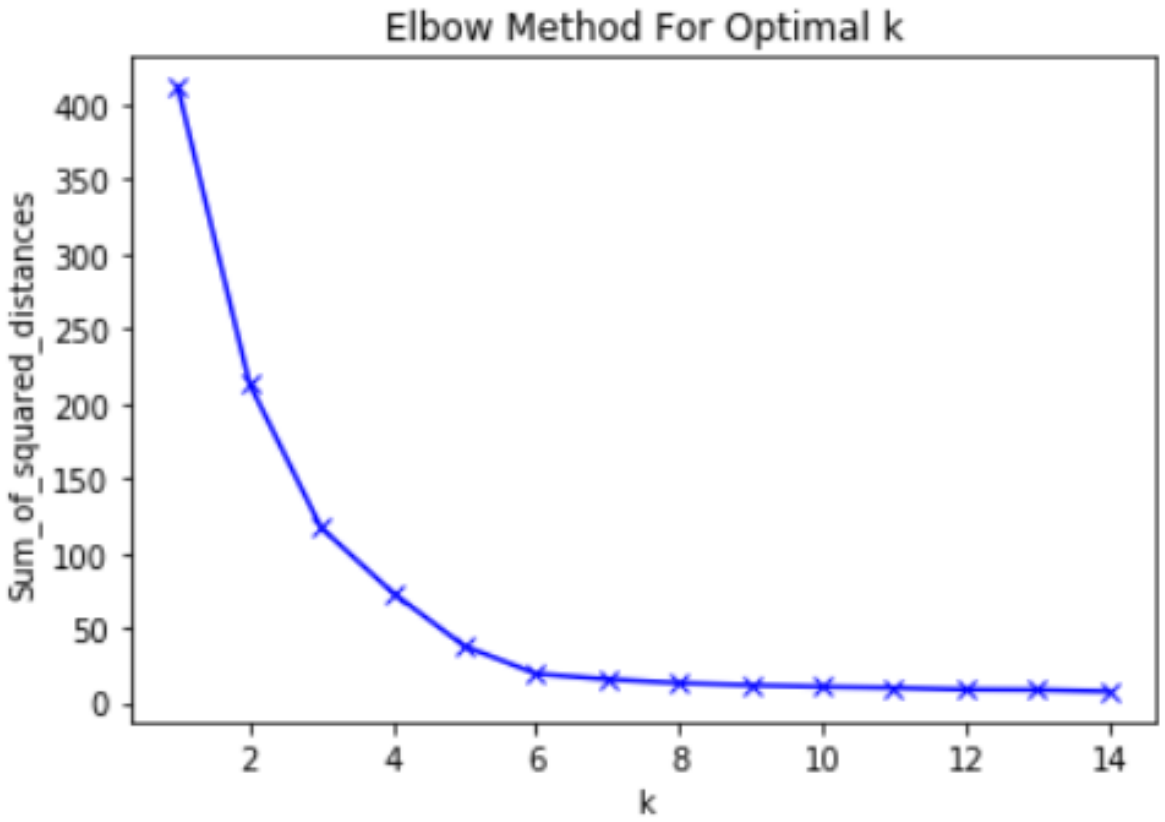
Se calculan las distancias pero se observa que no cambió la clasificación.

Ejemplo:

x	y
15	17.6
10	19.0
20	18.0
48	19.0
60	18.0
50	18.2



Método del codo



Usos del clustering

- Segmentación de personas, cosas, lugares
- Agrupamiento de documentos similares
- Clases iniciales cuando se tiene que hacer una clasificación a mano
- Ver valores atípicos

UPGMA (Unweighted Pair Group Method using Arithmetic averages)

- **UPGMA (método de grupos de pares no ponderados con media aritmética)**
- Es un método de agrupamiento jerárquico aglomerativo simple.
- **Algoritmo UPGMA:**
- Este método asume que las especies son grupos por si mismas, luego relaciona los dos grupos más cercanos basado en la matriz de distancias, recalcula la matriz de distancia y repite el proceso hasta que todas las especies estén conectadas a un único grupo.
- El método UPGMA realiza todos sus cálculos con la matriz calculada hallando la distancia genética entre los elementos.

UPGMA (Unweighted Pair Group Method using Arithmetic averages)

- **Descripción del método**
- Método para análisis filogenético definido por Peter H. A. Sneath y Robert R. Sokal 1973., principales representantes de la escuela fenética. Es un algoritmo heurístico que usualmente encuentra una solución muy acertada.
- Consiste en la búsqueda de la distancia más pequeña en la matriz de distancias genéticas y agrupar las unidades que la conforman como una sola unidad taxonómica independiente.
- Se calculan los promedios de la nueva unidad contra las restantes creando una nueva matriz y se repite el proceso hasta que todas las unidades queden unidas a un único elemento (ancestro hipotético).

UPGMA (Unweighted Pair Group Method using Arithmetic averages)

- Ventajas
- Es un método muy sencillo.
- Al realizar los cálculos basados en la matriz de distancias y no directamente sobre las secuencias es mucho más rápido computacionalmente.

- Desventajas
- Las secuencias no son consideradas como tales, sino que se trabaja con la matriz de distancias lo que puede causar pérdida de información.

UPGMA (Unweighted Pair Group Method using Arithmetic averages)

- Algoritmo
- El algoritmo UPGMA construye un dendrograma que refleja la estructura presente en una matriz de similitud por pares.
- En cada paso, los dos clústeres más cercanos se combinan en un clúster de nivel superior.
- La distancia entre los clústeres A y B cada uno de tamaño (cardinalidad) |A|y |B| se toma como el promedio de todas las distancias $d(x, y)$ entre los pares de objetos x en A y y en B, es decir, la distancia media entre elementos de cada grupo:
- $\frac{1}{|A|.|B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$

UPGMA (Unweighted Pair Group Method using Arithmetic averages)

En otras palabras, en cada paso de agrupación, la distancia actualizada entre los grupos unidos $A \cup B$ y un nuevo grupo X viene dado por el promedio proporcional de la $d_{A,X}$ y $d_{B,X}$ distancias:

$$d_{(A \cup B),X} = \frac{|A|.d_{A,X}.|B|d_{B,X}}{|A| + |B|}$$

El algoritmo UPGMA produce dendrogramas enraizados y requiere una suposición de tasa constante, es decir, asume un árbol ultramétrico en el que las distancias desde la raíz hasta la punta de cada rama son iguales. Cuando las puntas son datos moleculares (*es decir* , ADN ,ARN y proteína) muestreados al mismo tiempo, la suposición de ultrametricidad se vuelve equivalente a asumir un reloj molecular.

Ejemplo:
Primer agrupamiento

Este ejemplo se basa en una matriz de distancia genética JC69 calculada a partir de la alineación de secuencias de ARN ribosomal 5S de cinco bacterias: Bacillus subtilis (a), Bacillus stearothermophilus (b), Lactobaillus viridescens (c), Acholesplasma modicum (d) y Micrococcus luteus (e).

Supongamos que tenemos cinco elementos (a, b, c, d, e) y la siguiente matriz D_1 de distancias de pares entre ellos.

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

$D_1(a, b) = 17$ es el valor más pequeño de D_1 por lo que unimos los elementos a y b.

Referencias: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC341310/pdf/nar00294-0004.pdf

Ejemplo:
Primer agrupamiento

Estimación de la longitud de la primera rama

U es el nodo al que a y b están conectados.

$\delta(a,u) = \delta(b,u) = D_1(a,b)/2$

Los elementos a y b son equidistantes de u

Las ramas a y b que se unen a u tienen longitudes:

$$\delta(a,u) = \delta(b,u) = 17/2 = 8{,}5$$

Primera actualización de la matriz de distancias

$$D_2\big((a,b),c\big) = \frac{(D_1(a,c)\ x\ 1 + D_1(b,c)\ x\ 1}{1+1} = (21+30)/2 = 25{,}5$$

- $D_2\big((a,b),d\big) = \frac{(D_1(a,d)\ x\ 1 + D_1(b,d)\ x\ 1}{2} = (31+34)/2 = 35{,}2$
- $D_2\big((a,b),e\big) = \frac{(D_1(a,e)\ x\ 1 + D_1(b,e)\ x\ 1}{2} = (23+21)/2 = 22$

Segundo agrupamiento

	(a,b)	c	d	e
(a,b)	0	**25,5**	**32,5**	**22**
c	**25,5**	0	28	39
d	**32,5**	28	0	43
e	**22**	39	43	0

$D_2((a,b),e) = 22$

22 es el elemento más pequeño por lo que unieron el clúster (a,b) y el elemento e

Estimación de la longitud de la segunda rama

v es el nodo al que ***a*** y ***b*** están conectados a ***e***

$$\delta(a,v) = \delta(b,v) = \delta(e,v) = 22/2 = 11$$

Los elementos a y b son equidistantes de v

Las ramas a y b que se unen a v tienen longitudes:

Longitud de la rama faltante: $\delta(u,v) = \delta(e,v) - \delta(a,u) = \delta(e,v) - \delta(b,u) = 11 - 8{,}5 = 2{,}5$

Tercer agrupamiento

- **Actualización de la segunda matriz de distancia**

$$D_3(((a,b),e),c) = \frac{(D_2((a,b),c)\ x\ 2 + D_2(e,c)\ x\ 1)}{2+1} = \frac{25{,}5x2 + 39x1}{3} = 30$$

$$D_3(((a,b),e),d) = \frac{(D_2((a,b),d)\ x2 + D_2(e,d)\ x\ 1)}{2+1} = (32{,}5x2 + 43x1)/3 = 36$$

	((a,b),e)	**c**	**d**
((a,b),e)	0	**30**	**36**
c	**30**	0	28
d	**36**	28	0

$D_3(c,d) = 28$ es el valor más pequeño de D3 por lo que se unen los elementos c y d

Estimación de la longitud de la tercera rama $\delta(c,w) = \delta(d,w) = 28/2 = 14$

Paso final

- Actualización de matriz de distancia

$$D_4\big((c,d),(a,b),e\big) = \frac{(D_3(c,((a,b),e)\ x\ 1 + D_3(d,((a,b),e)\ x\ 1)}{1+1} = \frac{30x1 + 36x1}{2} = 33$$

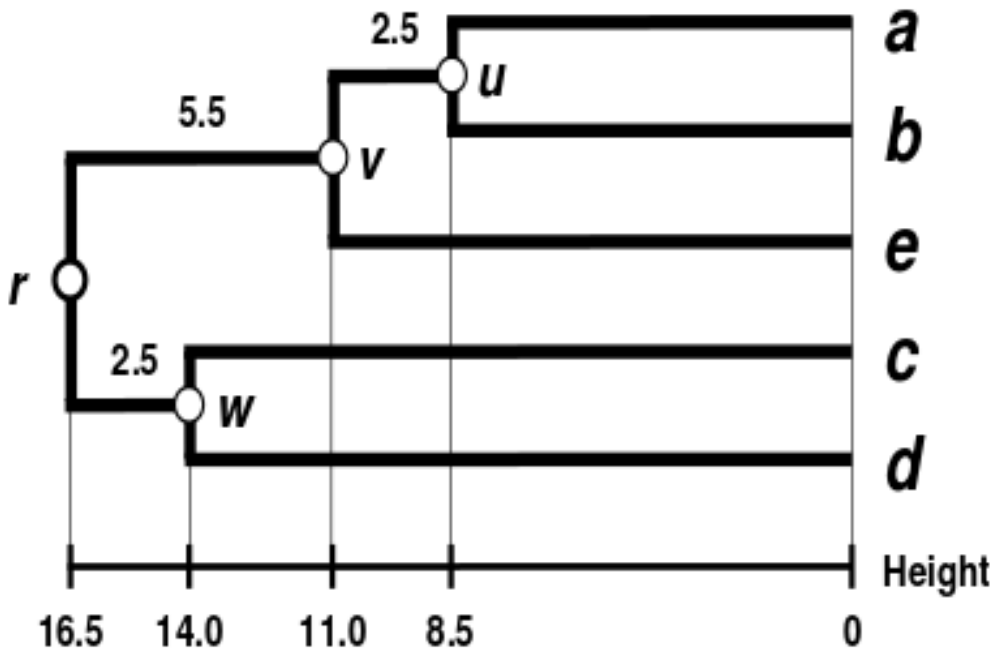
	((a,b),e)	**(c,d)**
((a,b),e)	0	**33**
(c,d)	**33**	0

Estimación de la longitud de la cuarta rama $\delta(((a,b),e),r) = \delta((c,d),r) = 33/2 = 16{,}5$

$$\delta(v,r) = \delta\big(((a,b),e),r\big) - \delta(e,v) = 16{,}5 - 11 = 5{,}5$$

$$\delta(w,r) = \delta\big((c,d),r\big) - \delta(c,w) = 16{,}5 - 14 = 2{,}5$$

Dendograma UPGMA



$$\delta(a,r) = \delta(b,r) = \delta(c,r) = \delta(d,r) = \delta(e,r) = 16{,}5$$

r es el nodo más profundo

Usos

- En ecología, es uno de los métodos más populares para la clasificación de unidades de muestreo (como parcelas de vegetación) sobre la base de sus similitudes por pares en variables descriptoras relevantes (como la composición de especies). Por ejemplo, se ha utilizado para comprender la interacción trófica entre las bacterias marinas y los protistas.
- En bioinformática, UPGMA se utiliza para la creación de árboles fenéticos (fenogramas). UPGMA se diseñó inicialmente para su uso en estudios de electroforesis de proteínas, pero actualmente se usa con más frecuencia para producir árboles guía para algoritmos más sofisticados. Este algoritmo se utiliza, por ejemplo, en procedimientos de alineación de secuencias, ya que propone un orden en el que se alinearán las secuencias. De hecho, el árbol guía tiene como objetivo agrupar las secuencias más similares, independientemente de su tasa evolutiva o afinidades filogenéticas, y ese es exactamente el objetivo de UPGMA.
- En filogenética, UPGMA asume una tasa de evolución constante (hipótesis del reloj molecular) y que todas las secuencias se muestrearon al mismo tiempo, y no es un método bien considerado para inferir relaciones a menos que esta suposición haya sido probada y justificada para el conjunto de datos que se está analizando. usó. Tenga en cuenta que incluso bajo un 'reloj estricto', las secuencias muestreadas en diferentes momentos no deberían conducir a un árbol ultramétrico.

¿Preguntas?