Análisis Cluster o de Conglomerados

Dra. Marta Quaglino

2020

Generalidades

- Los métodos de análisis Cluster engloban a un conjunto muy amplio de técnicas desarrolladas en distintos tiempos y ámbitos, que han respondido a la necesidad de conseguir el agrupamiento de unidades u objetos en clases mas o menos homogéneas
- Algunos ejemplos han sido la clasificación de especies (animales, plantas, etc.), la definición de áreas geográficas similares en cuanto a desarrollo económico,
- El hecho de provenir de distintos ambientes, produce una gran variabilidad entre las propuestas y la falta de un cuerpo orgánico, teórico, que guie a los procedimientos. Estos son en general soluciones ad-hoc, basadas en criterios particulares y que se adoptan, porque han funcionado
- Los métodos, ha sido reunidos y "clasificados" a posteriori, buscando los elementos comunes a los cuales responden
- El Análisis Cluster responde a la denominada clasificación no supervisada, porque la regla se deriva y se aplica SIN tener a priori un grupo donde ya estén las etiquetas identificadoras de grupos (como lo hace el Análisis Discriminante)

Objetivos

- Las técnicas de clasificación, pretenden lograr agrupamientos de **variables o de objetos**, en función de mediciones múltiples sobre un conjunto más o menos extenso de observaciones
- Los grupos formados, deberían ser homogéneos, con fuertes parecidos internos, y a la vez diferentes entre ellos
- La reunión de objetos o agrupamientos deben respetar los parecidos entre objetos, identificados por el conjunto de mediciones identificadas como importantes para la clasificación.
- El objetivo es lograr grupos homogéneos en cuanto a las variables medidas (si cambian los indicadores elegidos, se espera que el agrupamiento sea diferente)

¿Cuando se puede aplicar AC?

- No tiene restricciones
- Las variables pueden ser cuanti o cualitativas, pero de ello dependerá la definición de los parecidos o diferencias entre objetos, mediciones clave en muchos algoritmos, para identificar los grupos
- Pueden utilizarse las mismas estrategias para agrupar variables o individuos.
- Los algoritmos de agrupamiento nunca reconocen distintos roles en las variables que participan.
- En principio las unidades a agrupar no están clasificadas, sino que forman un todo (a diferencia de las aplicaciones de AD)

Como se logra el objetivo?

- En general debe definirse algún criterio de cercanía o parecido entre los objetos. Esto constituye un desafío en si mismo, porque pueden aplicarse muchos enfoques para definirlo
- Después de adoptar alguna definición, debe identificarse algún algoritmo (automatizado) que vaya realizando comparaciones sucesivas y reuniendo a los objetos mas parecidos
- El algoritmo de agrupamiento puede o no seguir una secuencia “ordenada”
- En el proceso de obtener los grupos, se constata que efectivamente resulten diferentes y si el grado de homogeneidad dentro es aceptable
- Finalmente, se describen cuales son las características propias de cada grupo.
- El numero de grupos, puede ser requerido a priori del agrupamiento. En algunos métodos se decide al final

Clasificación de los algoritmos

- Los propios algoritmos de clasificación, pueden ser agrupados, según la forma en que consiguen la identificación de los grupos y la cantidad de variables que vayan considerando en cada paso
- ***Divisivo, Aglomerativo o Particionante***. Un algoritmo es Divisivo si comienza considerando a todo el conjunto como un grupo y se va dividiendo según las diferencias mas notables. Es Aglomerativo si el proceso es a la inversa, en principio cada objeto es un grupo y se van juntando los mas parecidos. Los Particionantes, separan en grupos arbitrariamente, y luego se van intercambiando las unidades de forma de lograr homogeneidad interna
- ***Jerárquico o No jerárquico:*** En los jerárquicos, el agrupamiento es secuencial, respetando el orden de los parecidos. En los no jerárquicos no hay secuencialidad entre paso y paso.
- ***Politético o monotético***. Los monotéticos consideran de a una variable por vez para buscar los parecidos y agrupar (métodos más antiguos, usados en botánica o zoología). Los Politeticos consideran a todas las variables a la vez (más vinculados a métodos multivariados)

Otras características distintivas

- Con los métodos no jerárquicos, debe decidirse a priori el número total de grupos a formar y alguna medida de homogeneidad dentro y/o heterogeneidad entre grupos, a ser optimizada en los sucesivos pasos
- En los métodos jerárquicos, se hace una evaluación final de la historia de la formación sucesiva de grupos y se decide el numero de grupos a posteriori
- Para los métodos jerárquicos, es vital definir antes del algoritmo una medida del parecido (o de las diferencias) entre unidades, en cambio los no jerárquicos requieren de una medida de homogeneidad para saber cuando hacer el stop en el procedimiento
- Los algoritmos jerárquicos son menos eficientes frente a grandes matrices de datos y requieren mas memoria que los jerárquicos. Esto impone ciertas restricciones al momento de elegir la estrategia

Algoritmos jerárquicos

Algoritmos jerárquicos

- Fueron las primeras propuestas de clasificación en aparecer
- El punto de partida siempre es la construcción de una matriz de parecidos o de diferencias (***S*** o ***D***), a partir de la matriz de información primaria de Individuos x Variables (***X***)
- Esa matriz (***S*** o ***D***) es cuadrada, simétrica y de dimensión "*nxn*" o "*pxp*", según se busque agrupar individuos o variables
- Hay distintas propuestas de medidas de distancia o de similaridades, que dependen del tipo de variables disponibles. No hay acuerdo sobre cuál es más conveniente.

Medidas de similaridad o distancia

- Se pueden definir parecidos entre variables o entre individuos
- En g… le 0 es disc…
- Los … inversa reco…
- Existen múltiples propuestas basadas en a, b, y c , o en a, b, c y d

Individuo i / Individuo j	presente	ausente
presente	a	b
ausente	c	d

Medidas de similaridad o distancia

- Con variables dicotómicas se usa: SIMILARIDAD
- Se plantean en función del número de características comunes (a), no comunes (d), y alternadas, una si y otra no (b y c)

Individuo i / Individuo j	presente	ausente
presente	a	b
ausente	c	d

- Se pueden definir parecidos entre variables o entre individuos
- En general las medidas varían entre 0 y 1, donde 0 es discrepancia y 1 parecido completo.
- Los que comparan variables varían entre -1 y 1, reconociendo la posibilidad de una asociación inversa
- Existen múltiples propuestas basadas en a, b, y c , o en a, b, c y d

Definiciones de Similaridades (variables 0-1)

- Recordar que a=nº de características presentes comunes, d=idem ausentes, b y c presente/ausente y p=total variables
- Sokal y Michener: (a+d)/p (simétrico respecto a “a” y “d”, en coincidencias en si o en no).
- Roger y Tanimoto: p-(b+c)/p+(b+c) (simétrico)
- Ochiai: a/sqrt[(a+b) (a+c)] (no simétrico)
- Sokal y Sneath: a/a+2(b+c) (no simétrico)
- Jaccard: a/(a+b+c) (no simétrico)
- Russel y Rao: a/p (no simétrico)
- Dice: 2a/(a+b)(a+c) (no simétrico)
- Hamman: 1- (2(b+c)/p) (simétrico)
- Yule:(ad-bc)/(ad+bc) (estadístico)
- Pearson:(ad-bc)/sqrt[(a+c)(b+d)(a+d)(c+d)] (estad.)

Propuestas para variables cuantitativas

- Variables cuantitativas o de rango: pueden definirse SIMILARIDADES o DISTANCIAS.
- Para SIMILARIDADES se usan correlaciones entre individuos o variables, según el caso (Pearson, Spearman, Kendall). La definición es natural para parecidos entre variables, pero se puede usar como parecido entre individuos, cambiando el sentido de la matriz ***X***
- Para definir DISTANCIAS se usan conceptos geométricos como distancias Euclídeas. Hay también medidas menos usuales como Manhatan o Minkovsky, que generalizan el concepto y son mas robustas frente a outliers
- En ocasiones pude convenir realizar una estandarización previa de las variables para equilibrar su influencia. Se eligen estandarizaciones que no oculten la capacidad de separación de grupos, por ejemplo: x_{ij}/rango, x_{ij}/max(x_{ij}), x_{ij}/suma(x^2_{ij})

Definiciones de distancia

- **Minkowski:**
- $d_{ij} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right]^{1/r}$, con r ≥1
- **Euclidea:**
- $d_{ij} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2\right]^{1/2}$ (Minkowski con n=2)
- **Manhattan (city block)**
- $d_{ij} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|\right]$ (Minkowski con n=1)
  menos influenciada por valores extremos
- **Distancia de Mahalanobis o estadística**
- ${d_{ij}}^2 = (X_i - X_j)' \sum^{-1} (X_i - X_j)$, con $\sum$ matriz de var y cov de ***X***
- **Distancia de Tchebicheff o del supremo**
- $d(i,k) = \max_{i \le j \le p} |x_{ij} - x_{kj}|$ (Minkowski con n →∞)

Propiedades de las distancias

- **Propiedades**:
- Positividad, $d_{ij} \geq 0$, para todo par de objetos (i,j)
- Simetría, $d_{ij} = d_{ji}$
- Valor mínimo nulo, $d_{ij} = 0$ si y solo si $i=j$
- Desigualdad triangular
  $d(i,k) \leq d(i,m) + d(m,k) \quad , \quad \forall (i,k,m)$

Propuesta para mezcla de variables

- El coeficiente general de Gower, define la similitud de dos indivíduos como promedio de las similitudes por variable: $s_{ii'j}$ según "j" sea cuanti o cualitativa, j=1...p
- Si alguna variable "j" es binaria, los se asigna valor uno a $s_{ii'j}$ si x_{ij} y $x_{i'j}$ son iguales, y cero si son diferentes.
- Si la variable es cuantitativa $s_{ii'j}$ es = $1-|x_{ij}- x_{i'j}| / r_j$
- Luego $s_{ii'} = \frac{\sum_{j=1}^{p_1}\left(1-\frac{|x_{ij}-x_{i'j}|}{r_j}\right)+a+d+\alpha}{p_1+p_2+p_3}$, donde
- p_1 es el número de variables continuas
- p_2 número de variables binarias
- p_3 número de variables cualitativas
- r_k rango de la k-ésima variable continua
- a número de coincidencias "en 1" de las variables binarias
- d número de coincidencias "en 0" de las variables binarias
- α es el numero de coincidencias en variables cualitativas

Otra forma de generar distancias que refleje parecidos

- Frente a variables de tipo continuo, existen mas alternativas naturales para definir distancias, por lo tanto una propuesta es aplicar una estrategia de resumen de información multivariada, como paso previo a la búsqueda de agrupamientos
- En un primer paso, se aplica una técnica factorial, generando nuevas coordenadas (ACM o ACP) que tienen la propiedad de preservar el parecido entre individuos o variables
- Se retienen menos factores, pero más de los que se retendrían para explicar las asociaciones o los parecidos. De este modo se elimina “ruido”, reteniendo los parecidos mas evidentes.
- Luego se procede como con variables continuas, reemplazando la matriz de datos originales, por la matriz de individuos (o variables) por coordenadas factoriales, con un numero mas reducido de factores

Distancias o Similaridades?

- La elección de calculo de distancias o de similaridades, depende generalmente del problema y del área a la que corresponde
- Sin embargo en cualquier situación es posible pasar de una a otras con alguna función especialmente definida
- Si se define $s_{ij} \rightarrow d^2_{ij}=1-s^2_{ij}$
- Si se define $d_{ij} \rightarrow s_{ij}=1-d^2_{ij}$ (normalizando previamente las distancias a valores en (0,1))
- O bien $d_{ij} \rightarrow s_{ij} =-1/2\ d^2_{ij}$
- Otras relaciones posibles son:
- $s(i,k) = \frac{1}{1+d(i,k)}$ o $s(i,k) = c - d(i,k)$

Algoritmos jerárquicos aglomerativos:

- Los algoritmos jerárquicos mas utilizados son aglomerativos, politeticos
- Se adopta una medida adecuad de distancias o similaridades y se construye la matriz ***S*** o ***D***
- Paso 1: Cada unidad (variable o individuo) es un grupo
- Paso 2: Se agrupan las unidades con s_{ij} mayor, o d_{ij} menor
- Paso 3: Se re-calculan los parecidos entre grupos y unidades o entre grupos (no hay consenso acerca de cómo proceder en este paso), conformando una nueva matriz de similaridades o de distancias de menor dimensión
- Se agrupan nuevamente las unidades o grupos más parecidos (Paso 2)
- Se continúa con 1 a 3 hasta tener un solo grupo

Parecidos entre grupos (paso 3)

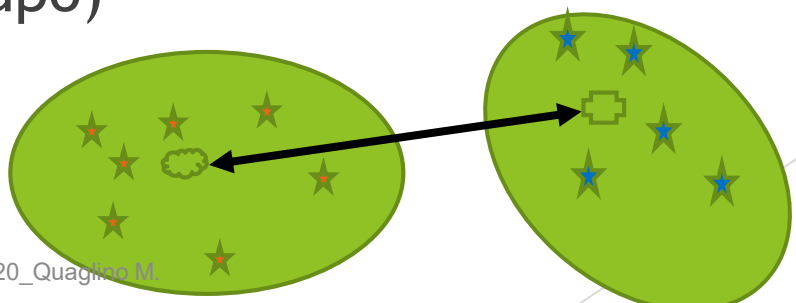
- Entre las propuestas para definir distancias entre grupos en las etapas de agrupamiento están:
- Encadenamiento simple:

$$d_{c,(a,b)}=\min (d_{c,b}, d_{c,a})$$

- Encadenamiento completo:

$$d_{c,(a,b)}=\max (d_{c,b}, d_{c,a})$$

- Estos dos metodos son “algo” extremos
- Estrategia de promedios simples o ponderados $d_{c,(a,b)}= (d_{c,b} + d_{c,a})/2$
- O la mediana del conjunto de distancias
- Método del centroide (distancia entre los centros de cada grupo)



Efecto de los criterios de distancia entre grupos

Tres estructuras de datos bivariados a clasificar

Fuente: Escudero, L. *Reconocimiento de patrones*

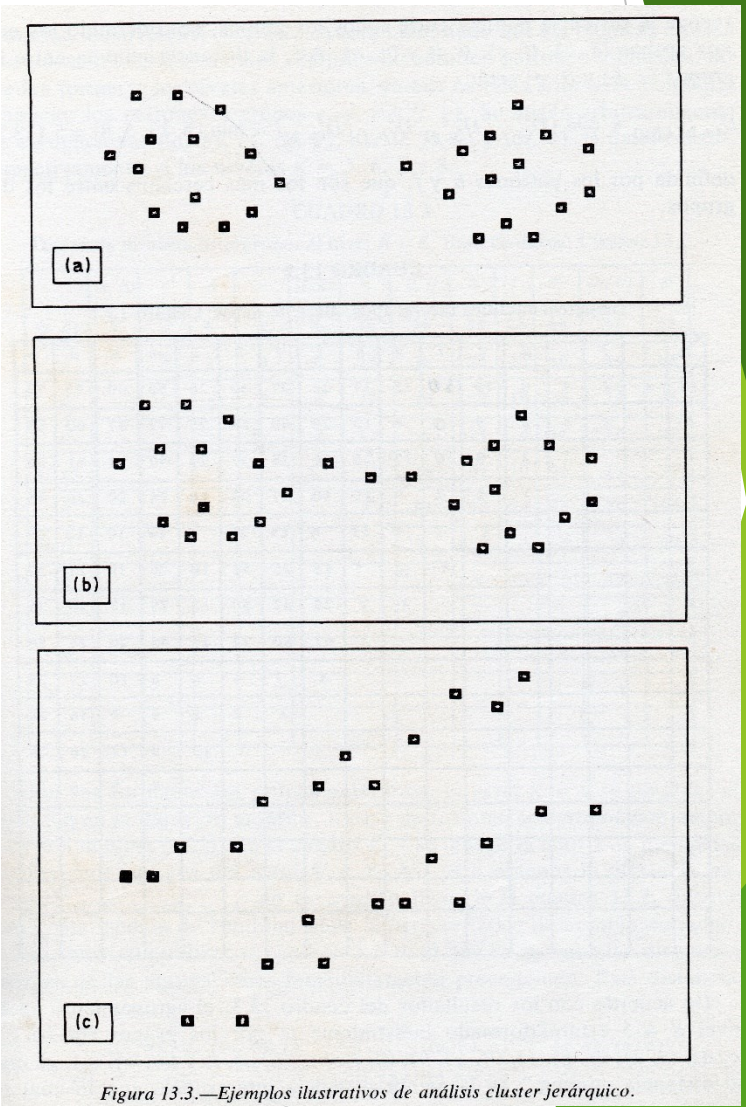


Figura 13.3.—Ejemplos ilustrativos de análisis cluster jerárquico.

Agrupamiento logrado con encadenamiento completo

El encadenamiento completo o metodo de la distancia maxima, tiende a formar grupos alargados.

Logra reconocer los cluster bien separados y no separa bien en presencia de puntos intermedios

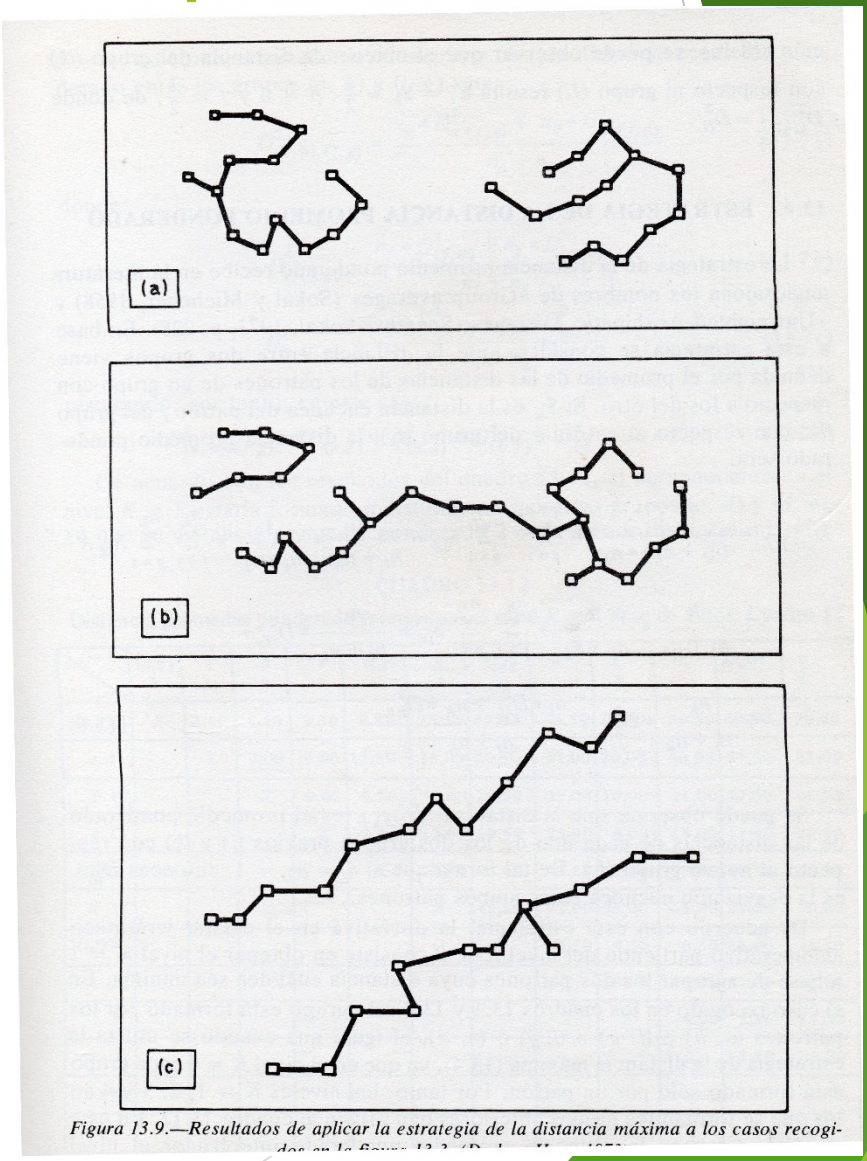


Figura 13.9.—Resultados de aplicar la estrategia de la distancia máxima a los casos recogi-

Agrupamiento logrado con encadenamiento simple

El encadenamiento simple o metodo de la distancia minima, tiende a formar grupos esfericos.

No logra reconocer los grupos alargados, y clasifica en dos grupos mas compactos

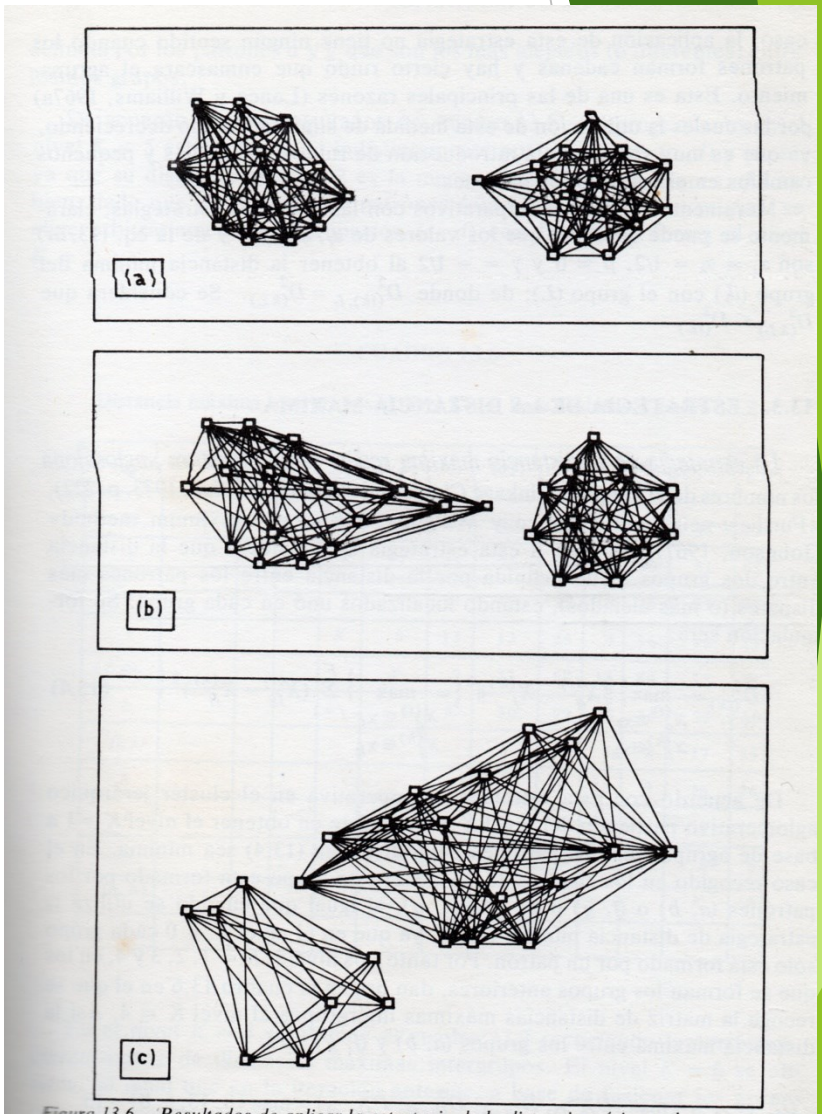


Figura 13.6.—Resultados de aplicar la estrategia de la distancia mínima a los casos recogidos en la figura 13.3 (Duda y Hart, 1973).

Metodo de Ward

- Es un algoritmo jerarquico muy utilizado, que no considera distancias para decidir cuales son los grupos a formar en cada caso, sino el incremento de variabilidades intra
- En cada etapa, se unen los dos clusters para los cuales se tenga el minimo aumento en la suma de los cuadrados de las diferencias de cada individuo respecto al centroide, dentro de cada cluster
- La suma de variabilidades intra es monótona no decreciente al ir aglomerando individuos
- Hay formulas de recurrencia que minimizan la cantidad de operaciones y aceleran el tiempo computacional necesario. Tambien es menor la memoria que requiere.

Caso particular: Método de Ward

- Este algoritmo funciona en forma aglomerativa, SIN calcular la matriz de similaridades. Define un criterio de optimización: “minimizar las sumas de cuadrados dentro”: $W=\Sigma_k\Sigma_i(X_{ik}-X_{mediak})'_i(X_{ik}-X_{mediak})$
- Parte de n grupos individuales con W=0
- Agrupa aquellos que logren el menor cambio en W (une los de menor distancia euclídea)
- Se continúa uniendo los elementos o conglomerados, que incrementen lo menos posible a W. Esto equivale a unir a y b tal que $\min (n_a n_{b)}/(n_a+n_b)(X_{media\ a}-X_{media\ b})'(X_{media\ a}-X_{media\ b})$
- Propiedades:
  - Tiende a combinar conglomerados con pocas observaciones
  - Produce conglomerados de tamaños similares
  - Requiere menos memoria para el procesamiento

Árbol de decisión o dendograma

- En los procedimientos jerárquicos, hay un agrupamiento sucesivo. Cuando algunas unidades se unen, ya no se separan.
- Se conserva la secuencia de los agrupamientos y se presenta el historial de los pasos sucesivos en un diagrama tipo árbol, donde el largo de las ramas va indicando el nivel al que se juntaron las unidades
- Este diagrama se utiliza para decidir el número de grupos o clases que se identifican. Es muy informativo para descubrir la estructura de parecidos entre unidades
- Cuando hay muchas unidades a clasificar se vuelve confuso o imposible de observar!

Un pequeño ejemplo

- Consideramos 7 individuos a clasificar, cuyas interdistancias se dan en una table D

	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

Distancia minima en la matriz → Se forma el primer grupo (C,E)

La matriz de distancias reducida

- La proxima matriz de distancias ya considera el grupo (C,E)

	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

Ahora la mas chica es 0,22

Se unen (F,D)

Estas distancias NO son las originales
d(C,E);D)=min[d(C,D);d(E,D)]=min(0,43; 0,29)=0,29
d(C,E);A)=min[d(C,A);d(E,A)]=min(0,7; 0,85)=0,7
Etc.

Proximos pasos

	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

Nuevo agrupamiento: (C, D, E, F)

	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

(A, C, D, E, F)

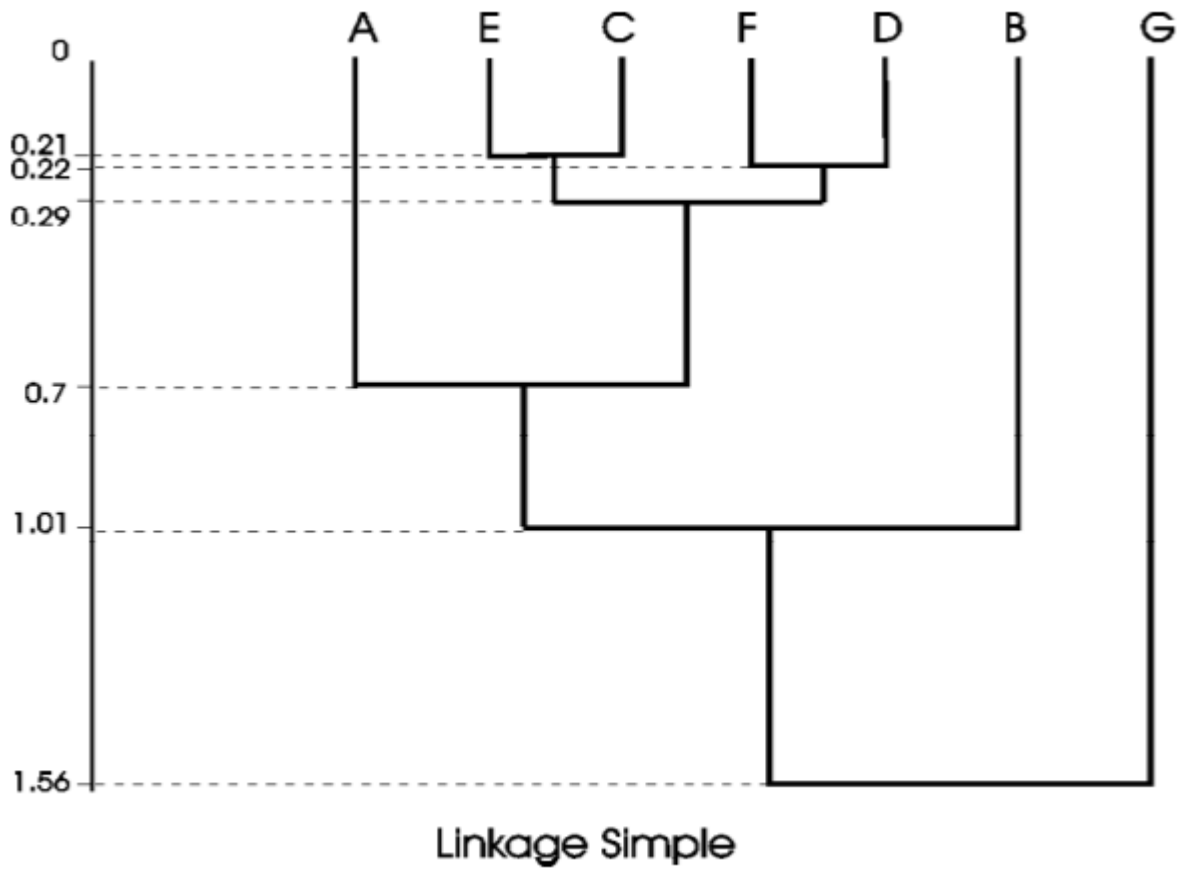
	(A,((C,E),(D,F)))	B	G
(A,((C,E),(D,F)))	0		
B	1,01	0	
G	1,56	2,83	0

(A, B, C, D, E, F)

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

Y finalmente se agrega G

Historial del agrupamiento secuencial

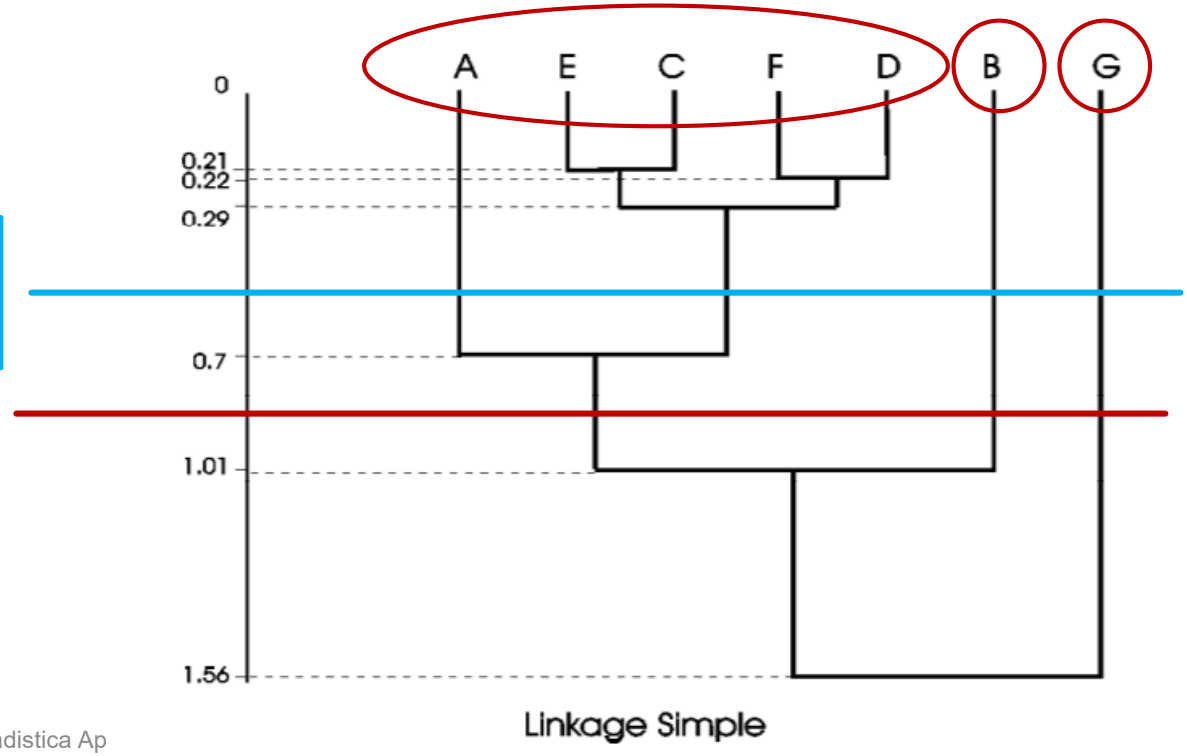


Si se calculan las interdistancias con el criterio del maximo….cambiaran los grupos?

Identificacion del numero de clusters

- Una vez que se construye el dendograma, para elegir un cierto numero de clusters (el algoritmo siempre parte de los clusters con una sola unidad y llega a un solo grupo), se traza una paralela al eje de abcisas por alguna altura en particular
- Las ramas que se cortan, identifican los grupos formados

{A} {ECDF} {B} {G}



Calidad del agrupamiento

- Los algoritmos jerárquicos pueden deformar el concepto original de parecidos entre unidades por el (necesario) cálculo de inter-distancias entre grupos del Paso 3
- Si las unidades tienen un parecido claramente ordenable, el agrupamiento será correcto
- Cómo evaluarlo sin “ver” los grupos previo a aplicar el algoritmo?
- Puede calcularse con una medida natural de asociacion, un coeficiente de correlación de Pearson
- Las medidas que se comparan son las distancias iniciales (elementos de un triangulo de la matriz de distancias o similaridades inicial), con la longitud de la rama que une a las dos unidades en el dendograma (n(n-1)/2 pares)
- La rama del dendograma es la distancia a la quedan agrupados según el algoritmo (es un elemento de la matriz de distancias del paso intermedio, en el cual se agruparon las unidades).

Interpretación del coeficiente

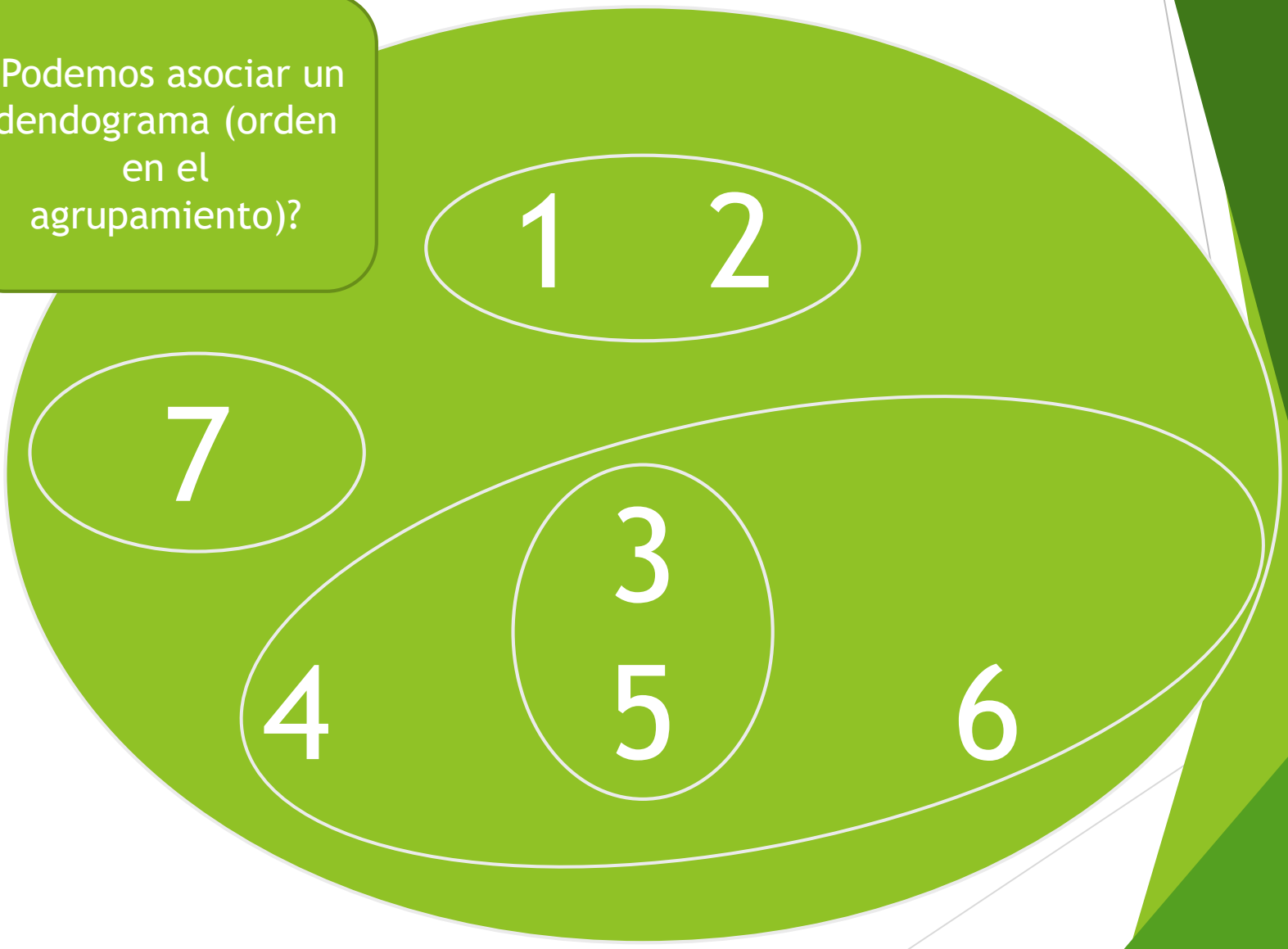
- Un coeficiente cercano a 1 indica que la estructura original de los datos es jerárquica y está siendo reflejada correctamente por el agrupamiento
- Si es cercano a 0, hay una gran distorsión y el agrupamiento no refleja los parecidos originales. Debe cambiarse el algoritmo o la definición usada en el paso 3
- Hay otras propuestas de indicadores de la concordancia entre estas medidas, en particular algunas se definen con un sentido inverso, el 0 indica mayor acuerdo
- Calcular, a modo de ejercicio, el coeficiente cofenetico en el ejemplo de los 7 objetos

Jerarquías:

- La distorsion en la formación de clusters producida por un algoritmo jerarquico, se da porque los parecidos entre los elementos no se ajustan a jerarquías. Hay conjuntos, que responden a una estructura "jerarquica", que se corresponden con una configuración geométrica muy particular
- Una jerarquía es una clase especial de subconjuntos de un conjunto no vacío, X={Xi, i=1…n}
- La clase $H=\{h_k / X \supset h_k\}$ es jerarquía sii:
- P1) $X \varepsilon H$
- P2) $Xi \varepsilon H$, para todo i=1…n
- P3) Dados h_k y $h_{k'}$, su intersección es vacía o están uno contenido en el otro (2 clases del mismo nivel son disjuntas)
- P4) Si $h_k \neq \{Xi\}$ entonces hay dos elementos $h_{k'}$, y $h_{k'}$ / $h_{k'} \cup h_{k'} = h_k$

Ejemplo de H ¿es Jerarquía?

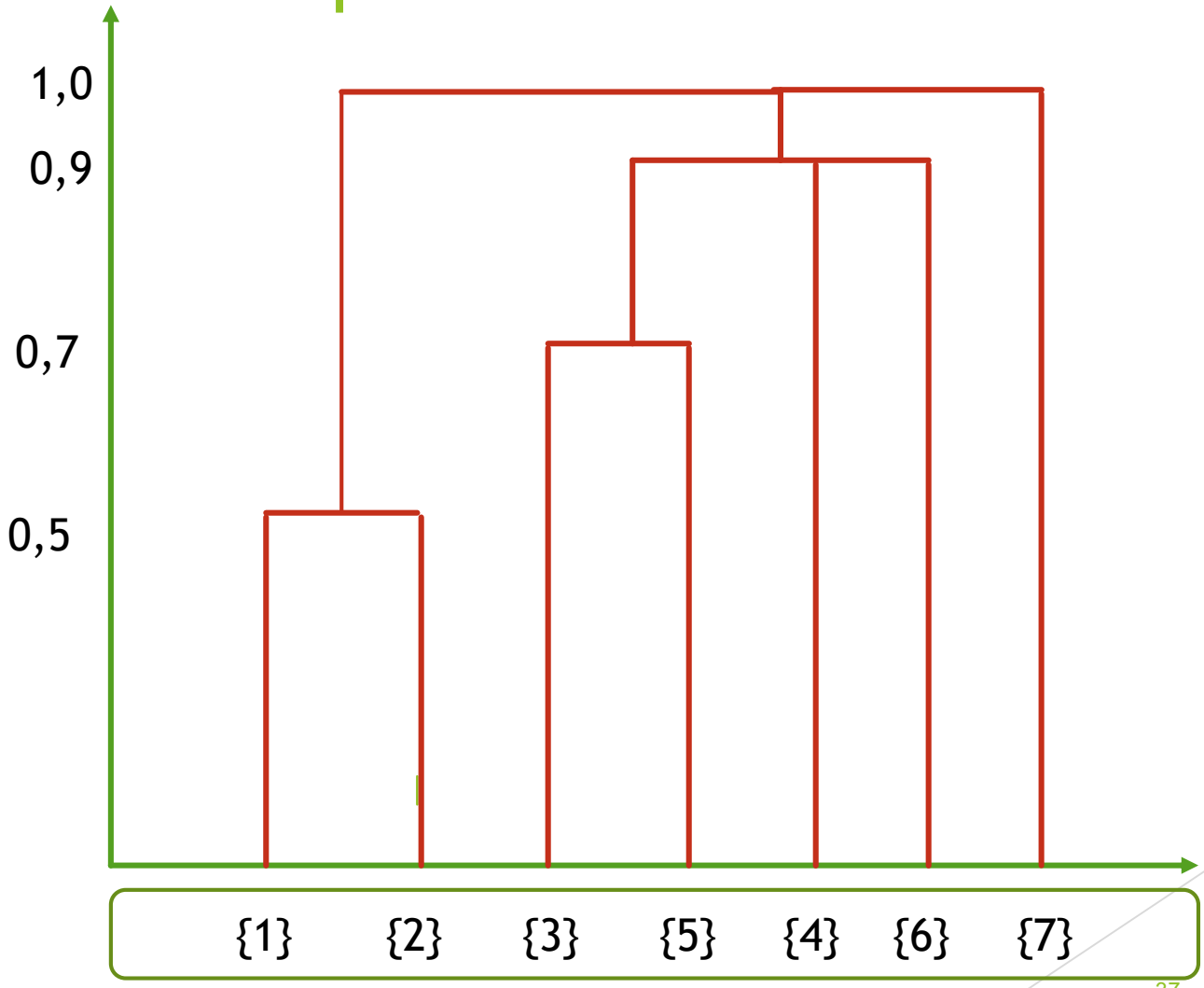
¿Podemos asociar un dendograma (orden en el agrupamiento)?



Jerarquía indexada:

- Sea X={1, 2, 3, 4, 5, 6} conjunto
- H = {X, {1,2}, {7}, {3,5}, {3,4,5,6}, {1}, {2}, {3}, {4}, {5}, {6}, {7} }, conjunto de subconjuntos de X
- Una Jerarquía H está indexada si sus elementos tienen asociados valores $d(h_k)$, respetando el orden de los parecidos.
- Los conjuntos elementales se corresponden con d=0 y si un elemento contiene a otro, el conjunto mas pequeño, el que esta contenido en otro, debe tener "d" menor
- Con este índice se puede armar la matriz de distancias entre los elementos de X, haciendo corresponder a d(Xi,Xj) el $d(h_k)$ con h_k el menor subconjunto que los contenga
- En el ejemplo podriamos asignar el índice d como:
- d({1})=d({2})=d({3})=d({4})=d({5})=d({6})=0
- d({1,2})=0,5 y d({3,5})=0,7
- d({3,4,5,6})=0,9
- d(X)=1…..y como quedaría el dendograma?

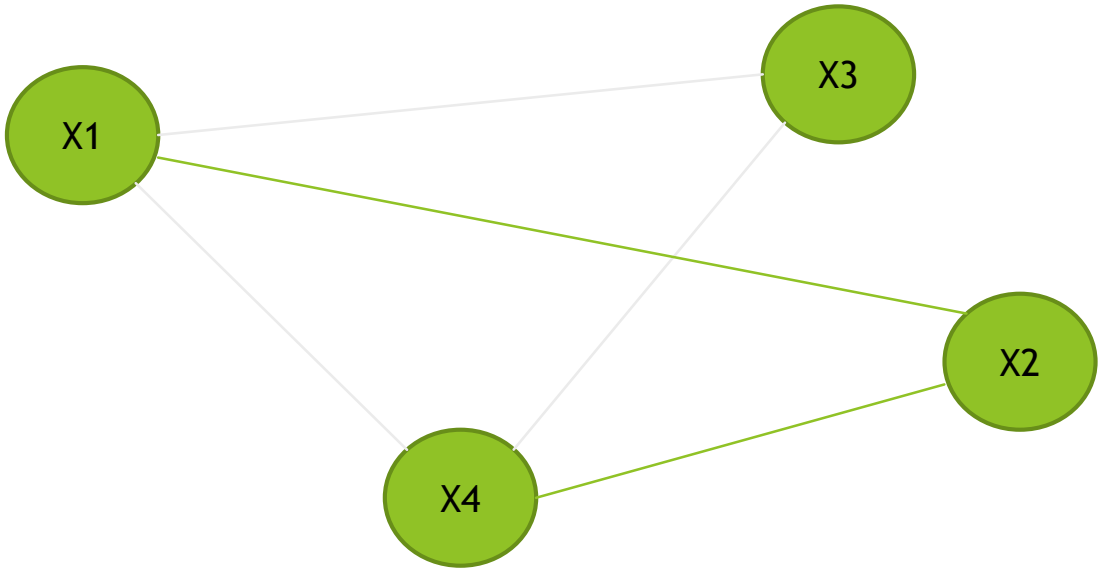
Dendograma asociado a la indexación de la Jerarquia



Distancia ultramétrica

- La distancia que se corresponde exactamente con una jerarquía indexada se denomina distancia ultramétrica y mantiene las propiedades de la euclidea y tiene una propiedad más fuerte que la desigualdad triangular
- Positividad: $d(i,j) \geq 0$ (el = sólo se cumple si i=j)
- Simetria: $d(i,j) = d(j,i)$
- Desigualdad ultrametrica: $d(i,j) \leq máx(d(i,i'),d(j,i'))$
- Desigualdad triangular: $d(i,j) \leq d(i,i') + d(j,i')$
- La desigualdad ultrametrica equivale, en un espacio euclideo, a que los puntos estén relacionados entre si formando, de a cualquier conjunto de tres, triángulos isósceles.
- Esta condición es muy difícil que se de en un conjunto de datos reales, con lo cual es de esperar que cualquier algoritmo jerarquico no respete en forma absoluta a los parecidos iniciales

Interpretación geométrica de distancia ultramétrica



Cada distancia entre dos puntos cualesquiera debe ser menor o igual que la mayor de las distancias entre esos puntos y un tercero

¿CÓMO DEBERIAN SER LOS TRIANGULOS?

Algoritmos no-jerárquicos

Algoritmos no jerárquicos

- Son preferibles en situaciones donde haya muchos objetos para agrupar (los jerárquicos serían muy lentos o imposibles de aplicar) o bien cuando no hay indicios que los objetos de asemejen jerárquicamente
- A diferencia de los jerárquicos se necesita establecer el numero de clases a priori
- Parten de la matriz original de datos o de una matriz de coordenadas factoriales, SIN pasar a la matriz de distancias o similaridades

Criterios no jerárquicos

- La idea general de todos los algoritmos es realizar una partición en C clases de modo que se optimice alguna medida como la minimización de inercias intra-clases
- Una alternativa (no-viable) sería armar todas las variantes de particiones posibles, que logren las C clases, calcular las inercias y decidir por la mejor
- La propuesta es armar C clases iniciales e ir reasignando individuos a clases diferentes, de modo de mejorar las inercias intra-clases (u otra medida), en forma progresiva
- Se continúa hasta que no se produzcan mejoras

Medidas de optimalidad

- SE puede optar por utilizar distintos criterios de homogeneidad para mover los elementos entre grupos hasta obtener los clusters definitivos
- Minimización de las distancias a los centros de grupo
- Minimización de la traza de la matriz de variancias y covariancias “dentro” (S_W)
- Minimización del determinante de S_W
- Maximizar la traza de $S^{-1}_W B$, etc.
- En cualquier caso, estas medidas deben ser re-calculadas en cada iteración del algoritmo y comparadas con las obtenidas en el paso previo.
- Los requerimientos de memoria son mucho menores que con los cluster jerarquicos

Posibilidad de repetición del algoritmo

- La configuración final de las clases, depende de la partición elegida en el primer paso
- Esa partición puede hacerse al azar, o alrededor de centros de clase pre-definidos (semillas)
- Para minimizar el efecto de esta elección, puede aplicarse el algoritmo repetidas veces, con distintas particiones iniciales
- Finalmente se comparan los resultados y se identifican las “clases fuertes”

Calidad de los grupos formados

- Se aconseja evaluar a posteriori, la homogeneidad de las clases lograda a través del algoritmo, de modo de chequear también la pertinencia del número de clases elegido
- Las medidas utilizadas son magnitudes positivas pero no acotadas, y no tienen un “valor óptimo”
- Frecuentemente para decidir el número de grupos a formar, se obtienen varios resultados con diferente numero de clusters (k), se evalúan varias medidas de homogeneidad (o heterogeneidad entre grupos) y se toma una elección de compromiso- generalmente no se logra optimizar a todas.
- Medidas orientadoras:
- “Root Mean Square Std.Deviation” de las clases (RMSSTD). Debería ser pequeña.

Otras medidas

- Semi-Partial R-Square (SPR). Debería ser pequeño.
- $R^2=SS_b/SS_t$, o proporción de variancia explicada por la pertenencia de las observaciones a los conglomerados. Debe ser grande
- Distancias entre centros de clusters. Deben ser grandes
- Estadísticas F que comparan las variabilidades obtenidas con K y (K+1) grupos haciendo variar K. Debe ser alto
- El criterio de agrupación cúbico (CCC) que establece una medida comparativa de desviación de los conglomerados respecto de la distribución esperada si las observaciones se hubieran obtenido de una distribución uniforme. Debería ser alto

Tratamiento posterior a la formación de cluster

- Una vez definidos los clusters, es usual que se presente el objetivo de describirlos, para identificar sus características particulares
- El análisis primario es la descripción de cada grupo según algunas medidas descriptivas como media, mediana, variabilidad, etc.
- Tambien es frecuente hacer una caracterización multivariada de los grupos a través de un Análisis Discriminante, lo cual también incluye test conjuntos de diferencias entre los vectores de promedios.
- Estos procedimientos también pueden servir para ajustar el número de clases inicial, en caso de observar grupos que no tengan grandes diferencias, o grupos que hayan quedado conformados con pocos individuos
- No hay “reglas” sino criterios que pueden ayudar a definirlos