8. ANÁLISIS DE CONGLOMERADOS (AC) CLUSTER ANALYSIS

USOS DEL AC

- **Objetivo:** Dada una muestra de ***n objetos/ observaciones***, cada uno de los cuales tiene datos/registros numéricos sobre ***p variables***, proyectar un esquema para agrupar los objetos/observaciones en clases, de modo que objetos/observaciones "similares" se ubiquen en la misma clase. El método debe ser completamente numérico y el número de clases no es conocido.
- **Comparación AFD – AC:** El problema es más complejo, puesto que en el AFD el número de grupos es conocido.

USOS DEL AC

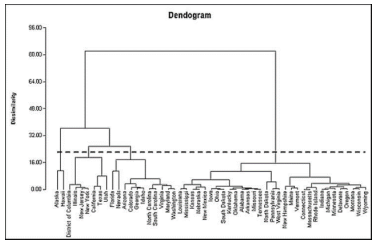
Razones para utilizar el AC

- Identificar objetivamente los "verdaderos" grupos.
- Reducir los datos a considerar (por ej., monitoreo de un representante de cada grupo).
- Sugerir relaciones que deben ser investigadas (en aquellos casos en que el AC genera agrupaciones inesperadas).

TIPOS DE AC

Existen muchos algoritmos para realizar un Análisis de Conglomerados o Clusters.

- **Técnicas Jerárquicas** que producen un Dendrograma



Estos métodos comienzan con el **cálculo de las distancias** de cada objeto/observación a todos los demás objetos/observaciones. Los grupos se forman entonces mediante un proceso de **aglomeración o división**.

TIPOS DE AC

Con ***aglomeración***, todos los objetos comienzan estando solos en grupos de uno. Los **grupos cercanos se fusionan** gradualmente hasta que finalmente todos los objetos están en un solo grupo. Con la ***división***, todos los objetos comienzan en un solo grupo. Éste se divide entonces en dos grupos, a su vez los dos grupos se dividen, y así sucesivamente hasta que todos los objetos estén en su propio grupo.

TIPOS DE AC

- **Técnicas que involucran la partición**, permitiendo a los objetos/observaciones moverse (entrar y salir) de los grupos en diferentes fases del análisis.

El enfoque básico consiste en **elegir unos centros de grupos más o menos arbitrarios**, con lo que **los objetos/observaciones se asignan al centro más cercano.** A continuación **se calculan nuevos centros** que representan los promedios de los objetos en los grupos. Un objeto se traslada a un nuevo grupo si está más cerca del centro de ese grupo que al centro de su grupo actual. Los grupos que están cerca se fusionan y los grupos que se extienden o que se esparcen se dividen, siguiendo algunas reglas definidas. El proceso continúa iterativamente hasta que se logra estabilidad con un número predeterminado de grupos. Usualmente se plantea un rango de valores para el número final de grupos.

TIPOS DE AC

- **Técnicas Jerárquicas** que producen un Dendrograma
  - ✓ **Vecino más cercano**
  - ✓ **Vecino más distante**
  - ✓ **Media del grupo**

- **Técnicas que involucran la partición**, permitiendo a los objetos/observaciones moverse (entrar y salir) de los grupos en diferentes fases del análisis.

TÉCNICAS JERÁRQUICAS

Vecino más cercano

Supongamos que la matriz de distancias entre 5 observaciones sea:

	1	**2**	**3**	**4**	**5**
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

Los grupos se fusionan a un determinado nivel de distancia, si una de las observaciones en un grupo está a esa distancia, o más próximo a por lo menos una observación en el segundo grupo.

TÉCNICAS JERÁRQUICAS

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

Distancia	Grupos
0	1, 2, 3, 4, 5
2	(1,2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)

- **Distancia 0:** Las 5 observaciones están cada una en un grupo.
- **Distancia 2:** La matriz de distancias muestra que la menor distancia entre dos observaciones es 2, entre la observación (1) y (2). De aquí, existen cuatro grupos a un nivel de distancia 2: (1, 2), (3), (4) y (5).

TÉCNICAS JERÁRQUICAS

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

Distancia	Grupos
0	1, 2, 3, 4, 5
2	(1,2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)

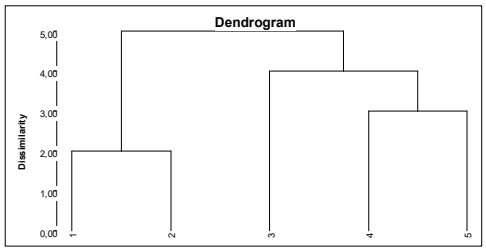
- **Distancia 3:** La distancia que sigue, de menor a mayor, es 3, entre las observaciones 4 y 5. De aquí, existen tres grupos a una distancia 3 (1, 2), (3) y (4, 5).
- **Distancia 4:** La distancia que sigue es 4, entre las observaciones 3 y 4. De aquí, existen dos grupos a este nivel de distancia (1, 2) y (3, 4, 5).

TÉCNICAS JERÁRQUICAS

- **Distancia 5:** Finalmente, la próxima distancia más pequeña es 5, entre las observaciones 2 y 3 y entre las observaciones 3 y 5. De aquí, para este nivel de distancia, los dos grupos se unen en un solo grupo (1, 2, 3, 4, 5) y el análisis está completo.

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

Distancia	Grupos
0	1, 2, 3, 4, 5
2	(1,2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)



TÉCNICAS JERÁRQUICAS

Vecinos más distantes

Con la unión de **"vecinos más distantes"** dos grupos se unen solamente si *los miembros más distantes de los dos grupos* están lo suficientemente próximos.

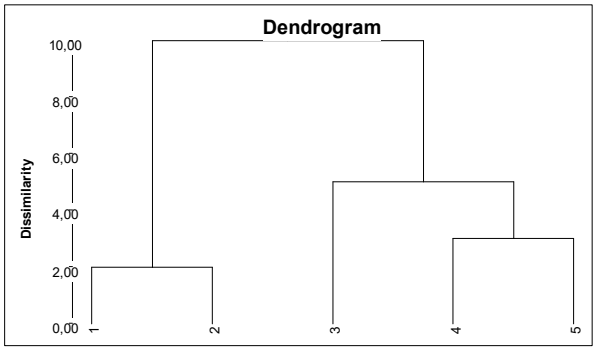
	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

Distancia	Grupos
0	1, 2 ,3 , 4, 5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
5	(1, 2), (3, 4, 5)
10	(1, 2, 3, 4, 5)

El objeto 3 no se une con los objetos 4 y 5 hasta el nivel de distancia 5 porque ésta es la distancia al objeto 3 de los objetos más alejados 4 y 5.

TÉCNICAS JERÁRQUICAS

El dendrograma resultante es:



TÉCNICAS JERÁRQUICAS

Media de grupo

Con la unión **"media de grupo"** dos grupos se unen si la distancia media entre ellos es suficientemente pequeña.

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

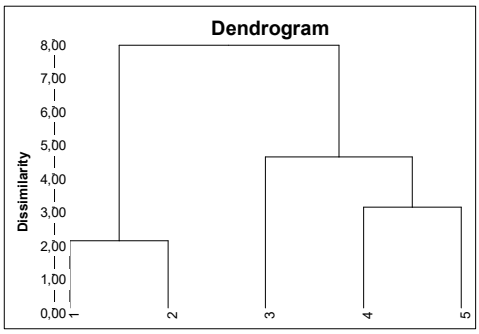
Distancia	Grupos
0	1, 2, 3, 4, 5
2	(1, 2), 3, 4, 5
3	(1, 2), 3 (4, 5)
4,5	(1, 2), (3, 4, 5)
7,8	(1, 2, 3, 4, 5)

Por ejemplo, los grupos (1, 2) y (3, 4, 5) se unen en el nivel de distancia 7,8 puesto que esa es la distancia media de los objetos 1 y 2 a los objetos 3, 4 y 5.

1-3	6
1-4	10
1-5	9
2-3	5
2-4	9
2-5	8
	7,8

TÉCNICAS JERÁRQUICAS

- El dendrograma resultante es:



TÉCNICAS QUE INVOLUCRAN LA PARTICIÓN

Los **métodos jerárquicos divisivos** se han utilizado menos que los métodos aglomerativos.

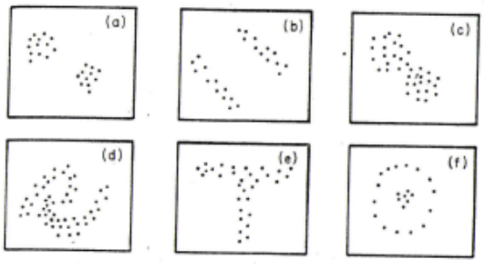
Los objetos se agrupan **inicialmente** en **un grupo**, que **luego se divide en dos grupos** separando el objeto que está más alejado de la media de los otros objetos. Luego, los individuos del grupo principal se mueven para un nuevo grupo si ellos están más próximos al nuevo grupo que lo que lo están al grupo principal. A medida que la distancia que se permite entre los individuos ubicados en un mismo grupo se reduce tienen lugar ulteriores subdivisiones. Finalmente, todos los objetos están en grupos de tamaño uno.

PROBLEMAS DEL AC

- ✓ Existen muchos algoritmos para realizar un análisis de conglomerados o clusters. Sin embargo, no existe ningún método aceptado como "el mejor". Lamentablemente, diferentes algoritmos no necesariamente producen los mismos resultados sobre un conjunto dado de datos.
- ✓ Una prueba justa de cualquier algoritmo consiste en tomar un juego de datos con una estructura de grupos conocida y verificar si el algoritmo puede reproducirla. Sin embargo, parece ser que esta prueba sólo es eficaz en aquellos casos en que los grupos son muy distintos.

PROBLEMAS DEL AC

- ✓ En algunos casos surgen dificultades debido a la forma de los conglomerados. **Ej.:** X_1 y X_2 son dos variables y las observaciones se grafican de acuerdo a sus valores.



- ☐ **(a) y (b):** Cualquier algoritmo debería reproducir los grupos