UTEC Posgrado

UTEC Posgrado

UTEC Posgrado

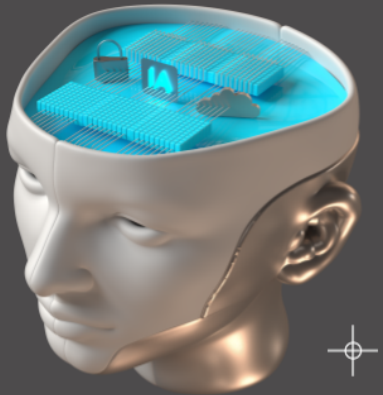
Una escuela para
Early Adopters

MACHINE LEARNING

AGRUPAMIENTO JERÁRQUICO

Motivación

Contexto y necesidad **del agrupamiento jerárquico**



Repaso: Análisis de conglomerados

Objetivo: Agrupar N objetos en subconjuntos (conglomerados) tales que:

- Objetos **dentro** de un conglomerado: muy **similares**

Repaso: Análisis de conglomerados

Objetivo: Agrupar N objetos en subconjuntos (conglomerados) tales que:

- Objetos **dentro** de un conglomerado: muy **similares**
- Objetos en conglomerados **distintos**: muy **diferentes**

Repaso: Análisis de conglomerados

Objetivo: Agrupar N objetos en subconjuntos (conglomerados) tales que:

- Objetos **dentro** de un conglomerado: muy **similares**
- Objetos en conglomerados **distintos**: muy **diferentes**

Central: La noción de **similitud (o disimilitud)** entre objetos, que solo puede provenir de consideraciones del dominio.

Limitaciones de K-means

- Requiere fijar K **de antemano**

Limitaciones de K-means

- Requiere fijar K **de antemano**
- Partición **plana**: no hay estructura anidada

Limitaciones de K-means

- Requiere fijar K **de antemano**
- Partición **plana**: no hay estructura anidada
- Al cambiar K, los clusters **no quedan anidados**

Limitaciones de K-means

- Requiere fijar K **de antemano**
- Partición **plana**: no hay estructura anidada
- Al cambiar K, los clusters **no quedan anidados**
- No proporciona **ordenamiento** dentro de cada cluster

Limitaciones de K-means

- Requiere fijar K **de antemano**
- Partición **plana**: no hay estructura anidada
- Al cambiar K, los clusters **no quedan anidados**
- No proporciona **ordenamiento** dentro de cada cluster

¿Existe una alternativa que NO requiera fijar K y que produzca una jerarquía natural?

$\Longrightarrow$ **Agrupamiento Jerárquico**

¿Qué es el agrupamiento jerárquico?

Produce **representaciones jerárquicas** donde los conglomerados en cada nivel se crean fusionando (o dividiendo) los del nivel adyacente.

- **Nivel más bajo**: cada observación es un conglomerado unitario

¿Qué es el agrupamiento jerárquico?

Produce **representaciones jerárquicas** donde los conglomerados en cada nivel se crean fusionando (o dividiendo) los del nivel adyacente.

- **Nivel más bajo**: cada observación es un conglomerado unitario
- **Nivel más alto**: un único conglomerado con todos los datos

¿Qué es el agrupamiento jerárquico?

Produce **representaciones jerárquicas** donde los conglomerados en cada nivel se crean fusionando (o dividiendo) los del nivel adyacente.

- **Nivel más bajo**: cada observación es un conglomerado unitario
- **Nivel más alto**: un único conglomerado con todos los datos
- **Resultado**: un **dendrograma** (árbol binario)

¿Qué es el agrupamiento jerárquico?

Produce **representaciones jerárquicas** donde los conglomerados en cada nivel se crean fusionando (o dividiendo) los del nivel adyacente.

- **Nivel más bajo**: cada observación es un conglomerado unitario
- **Nivel más alto**: un único conglomerado con todos los datos
- **Resultado**: un **dendrograma** (árbol binario)

▷ Hay $N - 1$ niveles en la jerarquía para N observaciones.

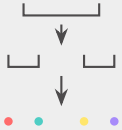
Dos paradigmas

Aglomerativo (bottom-up)

- Comienza con N clusters unitarios
- En cada paso: fusiona los 2 más cercanos
- $N - 1$ fusiones hasta un solo grupo

Divisivo (top-down)

- Comienza con 1 cluster global
- En cada paso: divide un cluster en 2
- $N - 1$ divisiones hasta singletons



Aplicaciones del agrupamiento jerárquico

Área	Aplicación
Genómica	Agrupamiento de genes por patrones de expresión
Taxonomía	Clasificación filogenética de especies
Procesamiento de texto	Organización jerárquica de documentos
Ciencias sociales	Agrupamiento de países por similitud política

Matrices de proximidad

Los datos se representan mediante una **matriz de disimilitudes** D de $N \times N$:

- $d_{ii'} \geq 0$: disimilitud entre objetos i e i'

Matrices de proximidad

Los datos se representan mediante una **matriz de disimilitudes** D de $N \times N$:

- $d_{ii'} \geq 0$: disimilitud entre objetos i e i'
- $d_{ii} = 0$: diagonal nula

Matrices de proximidad

Los datos se representan mediante una **matriz de disimilitudes** D de $N \times N$:

- $d_{ii'} \geq 0$: disimilitud entre objetos i e i'
- $d_{ii} = 0$: diagonal nula
- Simétrica: $d_{ii'} = d_{i'i}$

Matrices de proximidad

Los datos se representan mediante una **matriz de disimilitudes** D de $N \times N$:

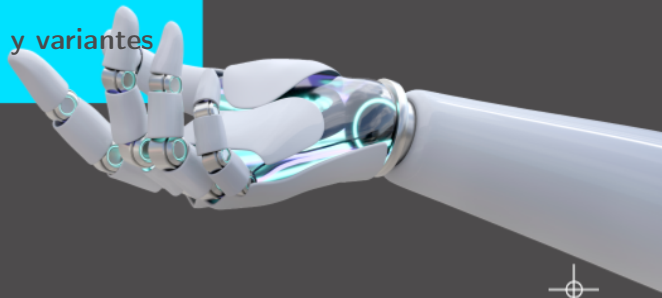
- $d_{ii'} \geq 0$: disimilitud entre objetos i e i'
- $d_{ii} = 0$: diagonal nula
- Simétrica: $d_{ii'} = d_{i'i}$

Distancia más común: euclidiana cuadrática

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

Desarrollo

Agrupamiento jerárquico **Algoritmos y variantes**



Agrupamiento aglomerativo: Idea

1. Iniciar con N conglomerados unitarios (cada observación es un cluster)

Agrupamiento aglomerativo: Idea

1. Iniciar con N conglomerados unitarios (cada observación es un cluster)
2. En cada uno de los $N - 1$ pasos:
   - Encontrar los dos clusters más cercanos
   - Fusionarlos en un solo cluster

Agrupamiento aglomerativo: Idea

1. Iniciar con N conglomerados unitarios (cada observación es un cluster)
2. En cada uno de los $N - 1$ pasos:
   - Encontrar los dos clusters más cercanos
   - Fusionarlos en un solo cluster
3. Registrar la disimilitud de cada fusión

Agrupamiento aglomerativo: Idea

1. Iniciar con N conglomerados unitarios (cada observación es un cluster)
2. En cada uno de los $N - 1$ pasos:
   - Encontrar los dos clusters más cercanos
   - Fusionarlos en un solo cluster
3. Registrar la disimilitud de cada fusión
4. Resultado: un **dendrograma**

Agrupamiento aglomerativo: Idea

1. Iniciar con N conglomerados unitarios (cada observación es un cluster)
2. En cada uno de los $N - 1$ pasos:
   - Encontrar los dos clusters más cercanos
   - Fusionarlos en un solo cluster
3. Registrar la disimilitud de cada fusión
4. Resultado: un **dendrograma**

Pregunta clave: ¿Cómo definimos la distancia **entre grupos**?

Algoritmo aglomerativo

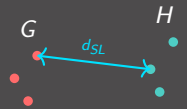
Entrada: Matriz de disimilitudes $\{d_{ii'}\}$, método de enlace
Salida : Dendrograma (árbol de fusiones)

Inicializar N clusters: $C_i = \{x_i\}$ para $i = 1, \ldots, N$
Calcular disimilitudes entre todos los pares de clusters
for $t = 1$ **to** $N - 1$ **do**
    Encontrar el par (C_a, C_b) con menor disimilitud intergrupal
    Fusionar: $C_{\text{new}} \leftarrow C_a \cup C_b$
    Registrar la fusión y su altura (disimilitud)
    Actualizar la matriz de disimilitudes con C_{new}
end
return *Dendrograma con las* $N - 1$ *fusiones*

Enlace simple (Single Linkage)

Disimilitud intergrupal = distancia entre el par **más cercano**:

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

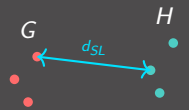


Enlace simple (Single Linkage)

Disimilitud intergrupal = distancia entre el par **más cercano**:

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

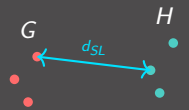
- Técnica del "vecino más cercano"



Enlace simple (Single Linkage)

Disimilitud intergrupal = distancia entre el par **más cercano**:

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

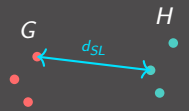


- Técnica del "vecino más cercano"
- Invariante frente a transformaciones monótonas

Enlace simple (Single Linkage)

Disimilitud intergrupal = distancia entre el par **más cercano**:

$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

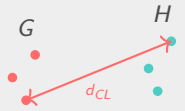


- Técnica del "vecino más cercano"
- Invariante frente a transformaciones monótonas
- Problema: encadenamiento

Enlace completo (Complete Linkage)

Disimilitud intergrupal = distancia entre el par **más lejano**:

$$d_{CL}(G, H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

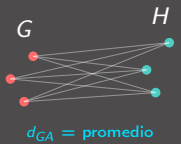


- Técnica del "vecino más lejano"
- Produce clusters **compactos** (diámetro pequeño)
- Invariante a transformaciones monótonas
- Puede violar "cercanía"

Enlace promedio (Group Average)

Disimilitud intergrupal = **promedio** de todas las distancias entre grupos:

$$d_{GA}(G, H) = \frac{1}{N_G \cdot N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

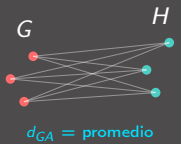


Enlace promedio (Group Average)

Disimilitud intergrupal = **promedio** de todas las distancias entre grupos:

$$d_{GA}(G, H) = \frac{1}{N_G \cdot N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

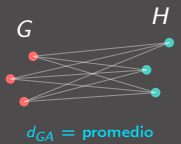
- Compromiso entre SL y CL



Enlace promedio (Group Average)

Disimilitud intergrupal = **promedio** de todas las distancias entre grupos:

$$d_{GA}(G, H) = \frac{1}{N_G \cdot N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

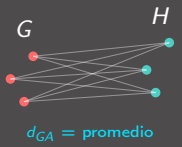


- Compromiso entre SL y CL
- **No** invariante a transformaciones monótonas

Enlace promedio (Group Average)

Disimilitud intergrupal = **promedio** de todas las distancias entre grupos:

$$d_{GA}(G,H) = \frac{1}{N_G \cdot N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$



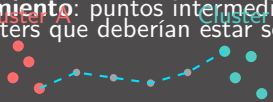
- Compromiso entre SL y CL
- **No** invariante a transformaciones monótonas
- Consistencia estadística: converge a una característica poblacional

Comparación de métodos de enlace

	Simple	Completo	Promedio
Fórmula	mín $d_{ii'}$	máx $d_{ii'}$	$\frac{1}{N_G N_H} \sum d_{ii'}$
Clusters	Alargados	Compactos	Intermedio
Defecto	Encadenamiento	Viola cercanía	Depende de escala
Invariancia	Sí (monótona)	Sí (monótona)	No
Consistencia	No ($\to 0$)	No ($\to \infty$)	Sí

Fenómeno de encadenamiento (Single Linkage)

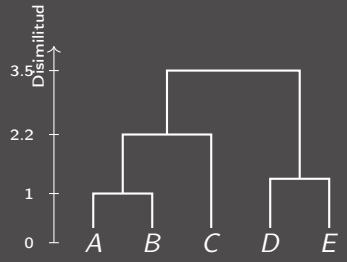
Encadenamiento: puntos intermedios conectan dos clusters que deberían estar separados



El enlace simple une clusters si **un solo par** de puntos es cercano $\Rightarrow$ produce clusters alargados y no compactos.

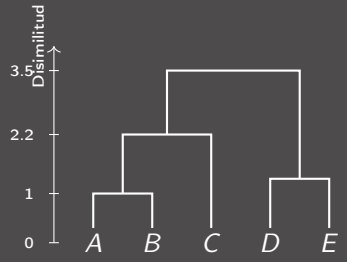
¿Qué es un dendrograma?

- Árbol binario que representa la jerarquía



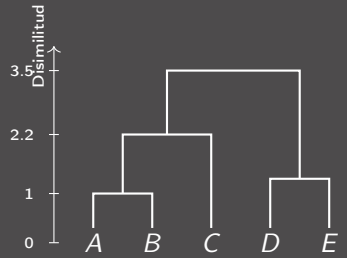
¿Qué es un dendrograma?

- Árbol binario que representa la jerarquía
- **Hojas**: observaciones individuales



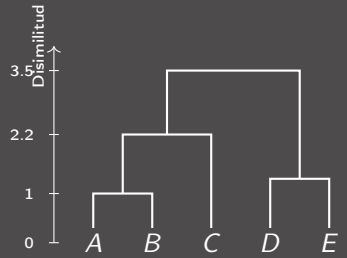
¿Qué es un dendrograma?

- Árbol binario que representa la jerarquía
- **Hojas**: observaciones individuales
- **Raíz**: cluster con todos los datos



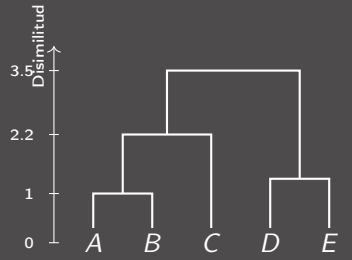
¿Qué es un dendrograma?

- Árbol binario que representa la jerarquía
- **Hojas**: observaciones individuales
- **Raíz**: cluster con todos los datos
- **Altura**: disimilitud de la fusión

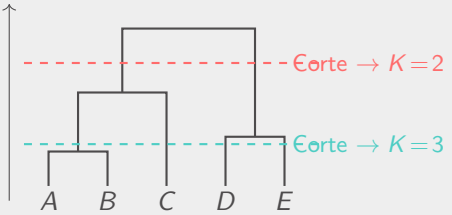


¿Qué es un dendrograma?

- Árbol binario que representa la jerarquía
- **Hojas**: observaciones individuales
- **Raíz**: cluster con todos los datos
- **Altura**: disimilitud de la fusión
- **Monotonicidad**: la altura crece con el nivel



Cómo leer un dendrograma



- Cortar horizontalmente a una **altura** → partición en K clusters
- Fusiones a alturas altas → candidatos a clusters "naturales"
- Los clusters quedan **anidados**: $K+1$ clusters son sub-divisiones de los K

Ejemplo: Matriz de distancias

Consideremos 5 puntos con la siguiente matriz de distancias:

	A	B	C	D	E
A	0				
B	2	0			
C	6	5	0		
D	10	9	4	0	
E	9	8	5	3	0

Aplicaremos enlace simple (single linkage) paso a paso.

Ejemplo: Paso 1 — Fusión de $\{A\}$ y $\{B\}$

Mínima distancia: $d(A, B) = 2 \Rightarrow$ Fusionar $\{A, B\}$

Actualizar distancias (enlace simple):

	AB	**C**	**D**	**E**
AB	0			
C	5	0		
D	9	4	0	
E	8	5	3	0

$d(AB, C) = \text{mín}(d(A, C), d(B, C)) = \text{mín}(6, 5) = 5$

Ejemplo: Paso 2 — Fusión de $\{D\}$ y $\{E\}$

Mínima distancia: $d(D, E) = 3 \Rightarrow$ Fusionar $\{D, E\}$

Actualizar distancias:

	AB	C	DE
AB	0		
C	5	0	
DE	8	4	0

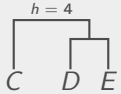
$d(DE, C) = \text{mín}(d(D, C), d(E, C)) = \text{mín}(4, 5) = 4$

Ejemplo: Paso 3 — Fusión de $\{C\}$ y $\{D, E\}$

Mínima distancia: $d(C, DE) = 4 \Rightarrow$ Fusionar $\{C, D, E\}$

Actualizar distancias:

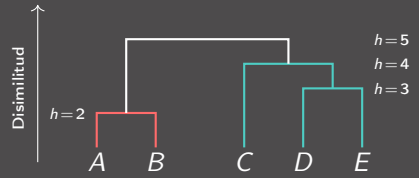
	AB	CDE
AB	0	
CDE	5	0



$d(AB, CDE) = \text{mín}(d(AB, C), d(AB, DE)) = \text{mín}(5, 8) = 5$

Ejemplo: Paso 4 — Fusión final

Única fusión restante: $d(AB, CDE) = 5 \Rightarrow$ Fusionar todo.



Dendrograma completo con enlace simple para los 5 puntos.

Mismo ejemplo con distintos enlaces

Resultados diferentes según el método de enlace:

Paso	Single	Complete
1	$\{A, B\}$ a $h = 2$	$\{A, B\}$ a $h = 2$
2	$\{D, E\}$ a $h = 3$	$\{D, E\}$ a $h = 3$
3	$\{C, D, E\}$ a $h = 4$	$\{C, D, E\}$ a $h = 5$
4	$\{A, B, C, D, E\}$ a $h = 5$	$\{A, B, C, D, E\}$ a $h = 10$

- Enlace simple: fusiones a alturas menores (efecto de encadenamiento)
- Enlace completo: mayor separación entre fusiones

Diámetro de un conglomerado

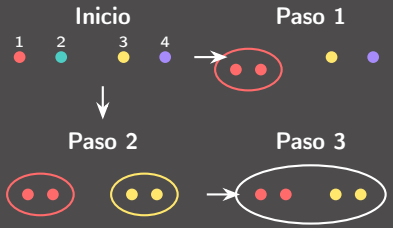
El **diámetro** D_G de un grupo G es la mayor disimilitud entre sus miembros:

$$D_G = \max_{\substack{i \in G \\ i' \in G}} d_{ii'}$$

Enlace simple:
Puede producir clusters con diámetros **muy grandes** (encadenamiento).

Enlace completo:
Tiende a producir clusters con diámetros **pequeños** (compactos).

Aglomerativo paso a paso (visual)



Agrupamiento divisivo (top-down)

1. Comenzar con todos los datos en un solo cluster G

Agrupamiento divisivo (top-down)

1. Comenzar con todos los datos en un solo cluster G
2. Encontrar la observación con mayor disimilitud promedio a las demás; moverla a un nuevo cluster H

Agrupamiento divisivo (top-down)

1. Comenzar con todos los datos en un solo cluster G
2. Encontrar la observación con mayor disimilitud promedio a las demás; moverla a un nuevo cluster H
3. Transferir a H las observaciones más cercanas (en promedio) a H que a G

Agrupamiento divisivo (top-down)

1. Comenzar con todos los datos en un solo cluster G
2. Encontrar la observación con mayor disimilitud promedio a las demás; moverla a un nuevo cluster H
3. Transferir a H las observaciones más cercanas (en promedio) a H que a G
4. Elegir qué cluster dividir a continuación (mayor diámetro)

Agrupamiento divisivo (top-down)

1. Comenzar con todos los datos en un solo cluster G
2. Encontrar la observación con mayor disimilitud promedio a las demás; moverla a un nuevo cluster H
3. Transferir a H las observaciones más cercanas (en promedio) a H que a G
4. Elegir qué cluster dividir a continuación (mayor diámetro)
5. Repetir hasta que todos sean singletons

Agrupamiento divisivo (top-down)

1. Comenzar con todos los datos en un solo cluster G
2. Encontrar la observación con mayor disimilitud promedio a las demás; moverla a un nuevo cluster H
3. Transferir a H las observaciones más cercanas (en promedio) a H que a G
4. Elegir qué cluster dividir a continuación (mayor diámetro)
5. Repetir hasta que todos sean singletons

Ventaja: más preciso cuando se buscan pocos clusters grandes.
Desventaja: menos estudiado; decisiones de división irreversibles.

Algoritmo de división (DIANA)

Entrada: Matriz de disimilitudes $\{d_{ii'}\}$
Salida : Dendrograma de divisiones

$\mathcal{C} \leftarrow \{\{x_1, \ldots, x_N\}\}$
while *algún cluster tenga más de 1 elemento* **do**
  Elegir $G \in \mathcal{C}$ con mayor diámetro D_G
  $i^* \leftarrow \arg\max_{i \in G} \frac{1}{|G|-1} \sum_{i' \in G, i' \neq i} d_{ii'}$
  $H \leftarrow \{x_{i^*}\}$; $G \leftarrow G \setminus \{x_{i^*}\}$
  repeat
    for $i \in G$ **do**
      if $\bar{d}(i, H) < \bar{d}(i, G \setminus \{i\})$ **then**
        $H \leftarrow H \cup \{i\}$; $G \leftarrow G \setminus \{i\}$
      end
    end
  until *no haya más transferencias*
  $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{G_{\text{old}}\}) \cup \{G, H\}$
end

Aglomerativo vs. divisivo

Aspecto	Aglomerativo	Divisivo
Dirección	Bottom-up	Top-down
Inicio	N singletons	1 cluster global
Ventaja	Más estudiado	Mejor para pocos clusters
Decisión	Fusionar 2 clusters	Dividir 1 cluster
Complejidad	$O(N^2 \log N)$	$O(2^N)$ (óptimo); $O(N^2)$ heurístico

Caso real: Microarreglos de tumores humanos

Datos: 6830 genes $\times$ 64 muestras.

- Enlace promedio y completo: resultados **similares**

Caso real: Microarreglos de tumores humanos

Datos: 6830 genes $\times$ 64 muestras.

- Enlace promedio y completo: resultados **similares**
- Enlace simple: grupos desbalanceados, encadenamiento

Caso real: Microarreglos de tumores humanos

Datos: 6830 genes $\times$ 64 muestras.

- Enlace promedio y completo: resultados **similares**
- Enlace simple: grupos desbalanceados, encadenamiento
- Cortar el dendrograma a distintas alturas $\rightarrow$ diferentes K

Caso real: Microarreglos de tumores humanos

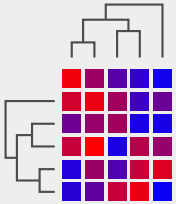
Datos: 6830 genes $\times$ 64 muestras.

- Enlace promedio y completo: resultados **similares**
- Enlace simple: grupos desbalanceados, encadenamiento
- Cortar el dendrograma a distintas alturas $\rightarrow$ diferentes K

Ventajas sobre K-means en este caso:

- Estructura **anidada** (subtipos dentro de tipos)
- **Ordenamiento** de muestras en el dendrograma
- Los biólogos pueden interpretar clusters de genes

Heatmap con dendrograma bidireccional



El reordenamiento bidireccional:

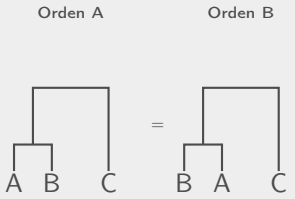
- Filas (genes) y columnas (muestras) reordenadas por dendrograma
- Revela **patrones** de co-expresión
- Más informativo que un ordenamiento aleatorio

Agrupamiento jerárquico vs. K-means

Aspecto	Jerárquico	K-means
Requiere K	No	Sí
Resultado	Dendrograma	Partición plana
Anidamiento	Sí	No
Determinístico	Sí	No (dep. de inicio)
Complejidad	$O(N^2 \log N)$	$O(NKpI)$
Escalabilidad	Limitada (N^2)	Buena
Distancia	Cualquiera	Euclidiana

Orientación de ramas en el dendrograma

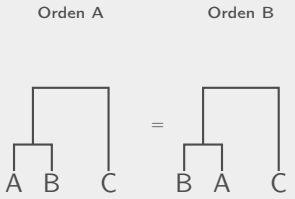
Al invertir las ramas de cualquier fusión, el dendrograma sigue siendo válido.



Orientación de ramas en el dendrograma

Al invertir las ramas de cualquier fusión, el dendrograma sigue siendo válido.

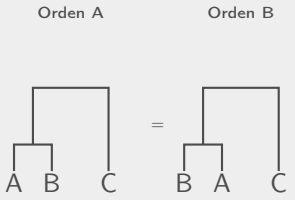
- Para N fusiones: 2^{N-1} ordenamientos de hojas posibles



Orientación de ramas en el dendrograma

Al invertir las ramas de cualquier fusión, el dendrograma sigue siendo válido.

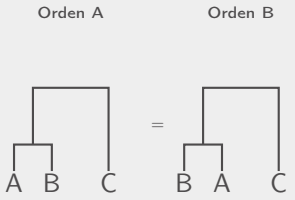
- Para N fusiones: 2^{N-1} ordenamientos de hojas posibles
- Se añade una **regla de ordenamiento** (ej.: subárbol más compacto a la izquierda)



Orientación de ramas en el dendrograma

Al invertir las ramas de cualquier fusión, el dendrograma sigue siendo válido.

- Para N fusiones: 2^{N-1} ordenamientos de hojas posibles
- Se añade una **regla de ordenamiento** (ej.: subárbol más compacto a la izquierda)
- Alternativa: MDS para ordenar las hojas



Consejos prácticos

1. **Elegir el enlace:**
   - Enlace promedio o Ward para la mayoría de casos
   - Evitar enlace simple si se sospecha encadenamiento

Consejos prácticos

1. **Elegir el enlace:**
   - Enlace promedio o Ward para la mayoría de casos
   - Evitar enlace simple si se sospecha encadenamiento
2. **Estandarizar variables** si tienen escalas diferentes

Consejos prácticos

1. **Elegir el enlace:**
   - Enlace promedio o Ward para la mayoría de casos
   - Evitar enlace simple si se sospecha encadenamiento
2. **Estandarizar variables** si tienen escalas diferentes
3. **Evaluar calidad:** silueta

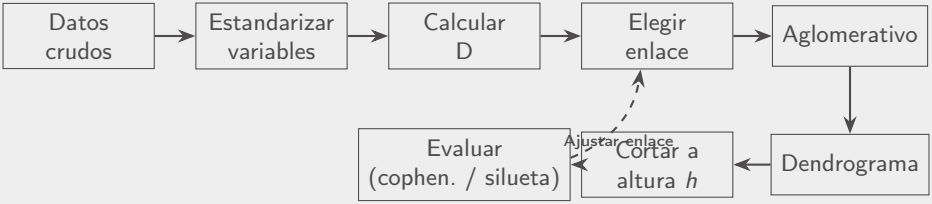
Consejos prácticos

1. **Elegir el enlace:**
   - Enlace promedio o Ward para la mayoría de casos
   - Evitar enlace simple si se sospecha encadenamiento
2. **Estandarizar variables** si tienen escalas diferentes
3. **Evaluar calidad:** silueta
4. **Cortar el dendrograma:**
   - Buscar saltos grandes en las alturas de fusión

Consejos prácticos

1. **Elegir el enlace:**
   - Enlace promedio o Ward para la mayoría de casos
   - Evitar enlace simple si se sospecha encadenamiento
2. **Estandarizar variables** si tienen escalas diferentes
3. **Evaluar calidad:** silueta
4. **Cortar el dendrograma:**
   - Buscar saltos grandes en las alturas de fusión
5. **Combinar métodos:** jerárquico para explorar estructura, luego K-means para particionar

Pipeline de agrupamiento jerárquico



Conclusiones

Resumen y reflexiones **Próximos pasos**

Resumen: Agrupamiento jerárquico

1. **No requiere fijar** K: el dendrograma muestra toda la jerarquía

Resumen: Agrupamiento jerárquico

1. **No requiere fijar** K: el dendrograma muestra toda la jerarquía
2. **Aglomerativo**: bottom-up, fusiona los 2 clusters más cercanos

Resumen: Agrupamiento jerárquico

1. **No requiere fijar** K: el dendrograma muestra toda la jerarquía
2. **Aglomerativo**: bottom-up, fusiona los 2 clusters más cercanos
3. **Divisivo**: top-down, divide el cluster más heterogéneo

Resumen: Agrupamiento jerárquico

1. **No requiere fijar** K: el dendrograma muestra toda la jerarquía
2. **Aglomerativo**: bottom-up, fusiona los 2 clusters más cercanos
3. **Divisivo**: top-down, divide el cluster más heterogéneo
4. **Enlace**: simple, completo, promedio, Ward

Resumen: Agrupamiento jerárquico

1. **No requiere fijar** K: el dendrograma muestra toda la jerarquía
2. **Aglomerativo**: bottom-up, fusiona los 2 clusters más cercanos
3. **Divisivo**: top-down, divide el cluster más heterogéneo
4. **Enlace**: simple, completo, promedio, Ward
5. **Dendrograma**: representación visual con estructura anidada

Resumen: Agrupamiento jerárquico

1. **No requiere fijar** K: el dendrograma muestra toda la jerarquía
2. **Aglomerativo**: bottom-up, fusiona los 2 clusters más cercanos
3. **Divisivo**: top-down, divide el cluster más heterogéneo
4. **Enlace**: simple, completo, promedio, Ward
5. **Dendrograma**: representación visual con estructura anidada
6. **Evaluación**: coeficiente cophenético, estadístico Gap

Ventajas y desventajas

Ventajas

- No requiere K a priori
- Estructura anidada
- Dendrograma interpretable
- Determinístico
- Funciona con matrices de disimilitud

Desventajas

- $O(N^2)$ en espacio
- No escala bien ($N > 10^4$)
- Fusiones/divisiones irreversibles
- Sensible a la elección de enlace
- Impone estructura jerárquica exista o no

¿Cuándo usar agrupamiento jerárquico?

Escenario	Recomendación
Explorar estructura sin fijar K	Jerárquico
Datos genómicos / taxonómicos	Jerárquico
$N > 10{,}000$	K-means / Mini-batch
Clusters esféricos, N grande	K-means
Clusters no convexos	DBSCAN
Solo matrices de proximidad	Jerárquico o K-medoides

Referencias

- Hastie, Tibshirani & Friedman (2009). *The Elements of Statistical Learning*, Cap. 14.

UTEC Posgrado

UTEC Posgrado