Clase N°10. Técnicas de aprendizaje automático no supervisado.

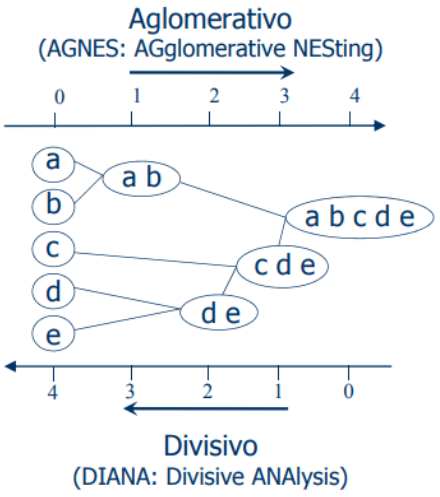
A diferencia del aprendizaje supervisado donde las variables observables y la variable predictora son conocidas de antemano, en el aprendizaje no supervisado, si bien se conocen los datos de entrada se desconoce cuál es la variable de salida, dado que los *datos no están etiquetados.* Es así que el objetivo del aprendizaje no supervisado, consiste en encontrar patrones o tendencias de las variables observables.

Modelo de Clustering jerárquico.

Consiste en establecer una jerarquía u orden de agrupamiento entre las variables mediante grupos de datos, llamados clústers. Entre las técnicas más utilizadas, existen la aglomerativa y la divisiva.

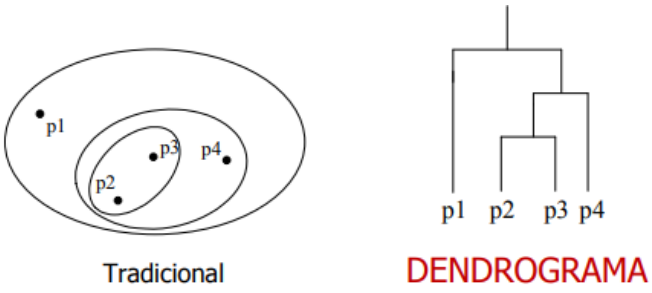
Aglomerativo: Establece diferentes grupos de datos (ab, de, cde) y los agrupa en función de su cercanía o distancias hasta que solo quede un único clúster (abcde).

Divisivo: Parte de un clúster único (abcde) que contenga todos los casos y lo divide en función a las diferentes distancias entre los datos generando k clústers (ab, cde, de). Estas técnicas se pueden visualizar en el siguiente gráfico:

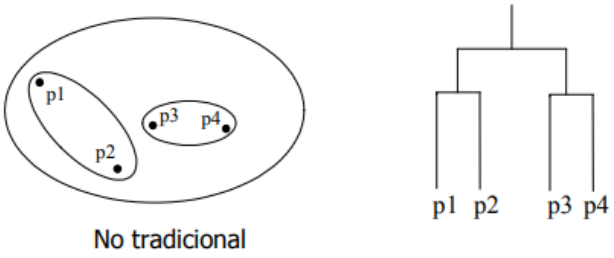


Por otra parte, para visualizar el grado de asociación entre las variables se puede utilizar un ***dendograma*****,** el cual mediante un esquema de árbol, muestra cómo se relacionan las variables desde un nodo central hacia nodos secundarios aumentando su nivel de agrupamiento.

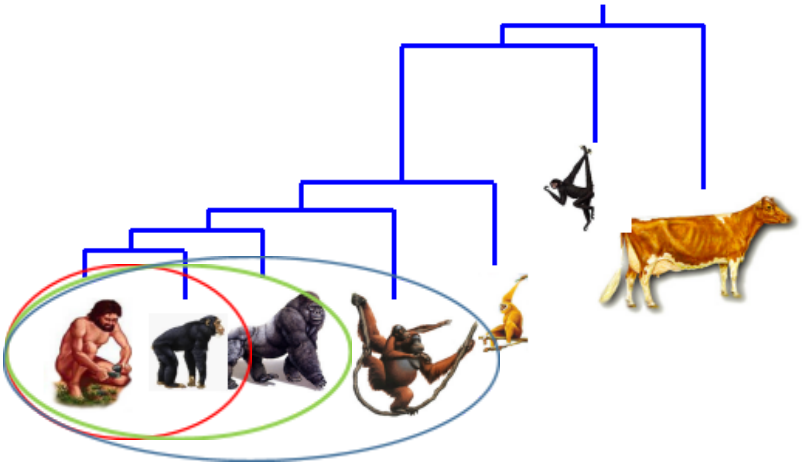
Ejemplo 1: Método tradicional



Ejemplo 2: Método no tradicional



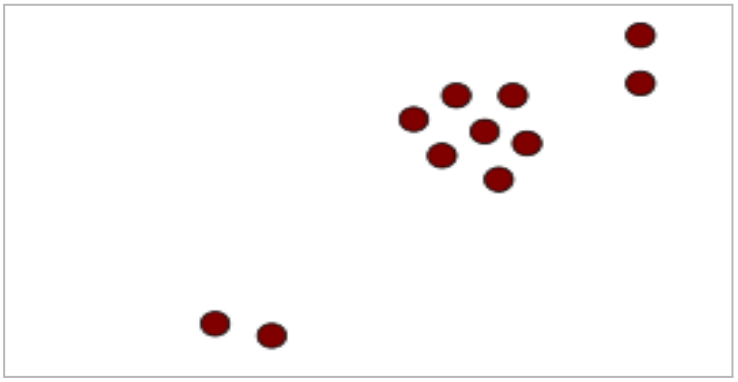
Ejemplo 3: Dendograma de evolución de especies.



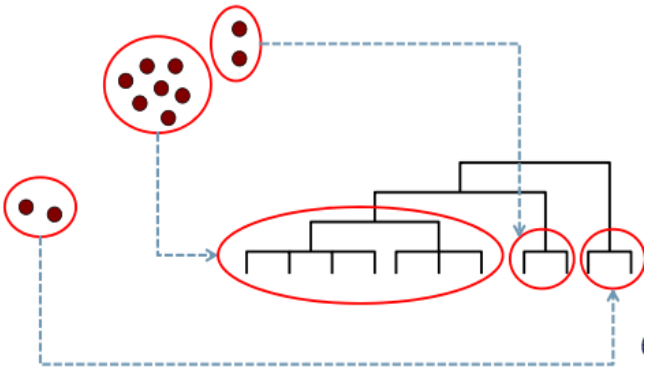
La similitud entre las especies está determinada por la altura del nodo más cercano.

Actividad 1:

Determinar el número de clústers que se podrían formar con los siguientes puntos y representarlos en un dendograma.

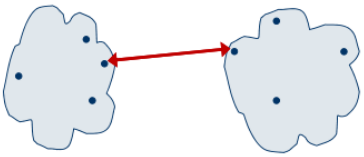
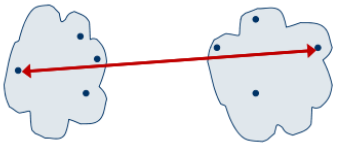


Solución:



Algoritmo de clustering jerárquico (aglomerativo)

1. Construir la ***matriz de distancias*** entre todos los puntos.
   La matriz es simétrica, cuadrada, y se verifica la propiedad conmutativa entre los puntos.
2. Combinar los ***clústers más cercanos***. (Método de **single-link**, **complete-link**).
3. Single-link considera la distancia mínima entre cada grupo, y complete-link la distancia máxima entre los agrupamientos.

single-link	complete-link
	

4. Aplicar estrategia de ***control irrevocable***.(Al unir dos grupos de elementos se descartan nuevas posibles uniones)

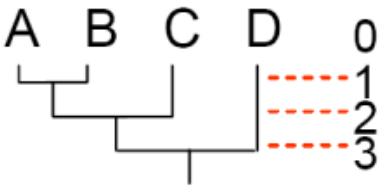
Ejemplo 4: Agrupar los puntos A, B, C, D mediante un algoritmo de clustering jerárquico.

***A*B *C *D**

1. Se construye la sig. matriz de distancias entre cada punto:

	A	**B**	**C**	**D**
A	0	1	4	5
B	1	0	2	6
C	4	2	0	3
D	5	6	3	0

2,3. Construir clústers más cercanos: **AB**, **BA**, CB (B ya formó grupo con A, por lo tanto por el principio de control irrevocable no puede formar un nuevo grupo con otro elemento. Se descarta esta unión y se continúa con la segunda distancia más próxima que es **CD**). El último clúster está formado por los puntos **DC.**



Desventajas del clustering jerárquico

- Baja escalabilidad, dado que se suele utilizar una muestra y no la totalidad del universo de datos.
- El método single-link se ve afectado ante la presencia de outliers (ruido).
- El método complete-link si bien es más robusto frente a la presencia de ruido, tiende a dividir los clústers grandes en clústers más pequeños o segmentados.

Actividad: Preguntas de opción múltiple.

Responder con verdadero o falso a las siguientes afirmaciones:

1. En el aprendizaje no supervisado los datos están etiquetados.
2. En un dendograma la similitud entre los datos está determinada por la altura del nodo más cercano.
3. La matriz de distancias es: simétrica, cuadrada, y se verifica la propiedad conmutativa entre los puntos.
4. El método single-link considera la distancia máxima entre cada grupo.