UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Agrupamiento jerárquico

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 18) 08.marzo.2021

Clasificación

Consideramos el problema de clasificación en un conjunto $\mathbb{X} \subseteq \mathbb{R}^d$. Si $\mathbb{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Este problema consiste en construir una función de asignación

$$h : \mathbb{X} \to C, \quad (\mathbf{x}_i) = c_i,$$

donde $C = \{c_1, c_2, \ldots, c_k\}$ es un conjunto finito de categorías o clases. El

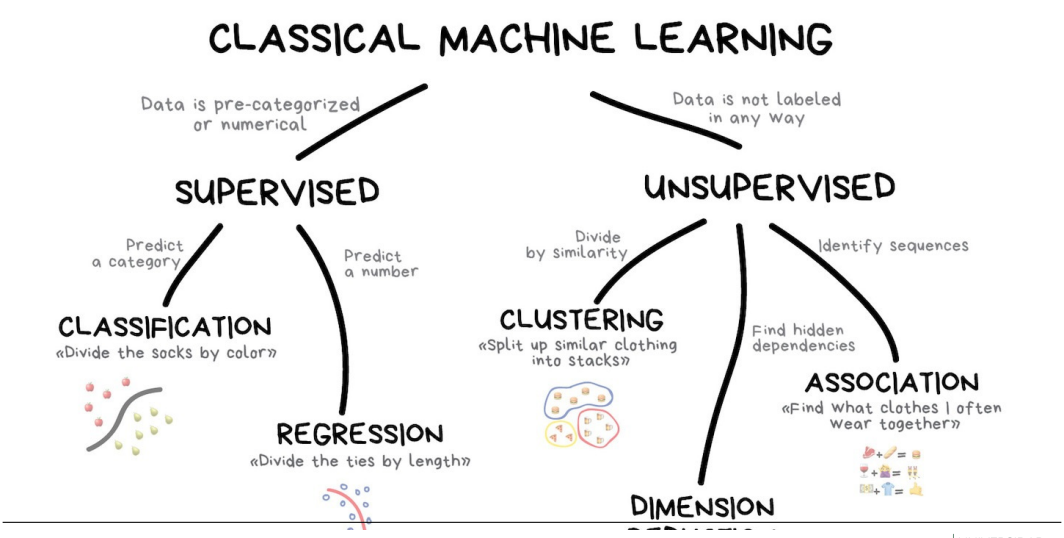
problema de clasificación consiste en asignarle a cada $\mathbf{x}_i$ una clase correspondiente c_i utilizando algún criterio específico.

En casos más generales, construimos u mapa

$$h : \mathbb{X} \to \mathbb{R}^k, \quad h(\mathbf{x}_i) = (p_1(\mathbf{x}_i), p_2(\mathbf{x}_i), \ldots, p_k(\mathbf{x}_i)),$$

donde cada $p_j(\mathbf{x}_i) = \mathbb{P}(\mathbf{x}_i \in C_j)$, esto es, $p_j(\mathbf{x}_i)$ es la probabilidad de que $\mathbf{x}_i$ pertenezca a la categoría j.

Clasificación supervisada y no supervisada



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Clasificación supervisada y no supervisada

- En la clasificación supervisada, además del conjunto de datos $\mathbb{X} = \{\mathbf{x}_i\}_{i=1}^n$, se cuenta adicionalmente con un conjunto de etiquetas ya pre-definidas $\mathbf{y} = \{y_i\}_{i=1}^n$.
  El conjunto C y la cantidad de etiquetas las da el contexto.
  Tales pares $\{(\mathbf{x}_i, y_i)\}$ se utilizan para construir un modelos de clasificación $h : \widetilde{\mathbb{X}} \to C$, el cual permite clasificar elementos en un superconjunto mayor a $\mathbb{X}$. (Típicamente $\mathbb{X}$ es el conjunto de entrenamiento).

- En el caso no supervisado (clustering), no se cuenta con el conjunto de etiquetas $\mathbf{y}$. En ese caso, se utiliza la misma estructura de los datos para detectar o agrupar los datos en conglomerados.
  El conjunto C y la cantidad de etiquetas k no se conoce.

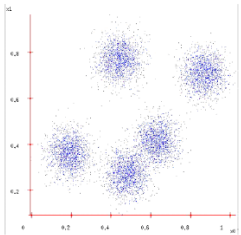
Clasificación supervisada y no supervisada

Clasificación no supervisada = Agrupamiento (*clustering*).

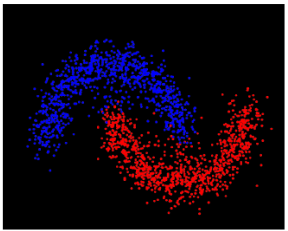
Problema:

- Segmentar datos en subgrupos homogéneos.
- Encontrar grupos en base de semejanza.

Situación idonea:

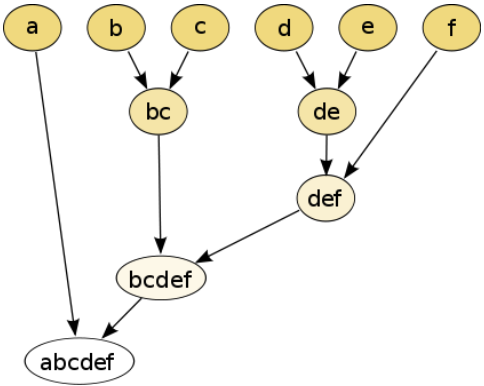


dolor de cabeza:



Agrupamiento jerárquico

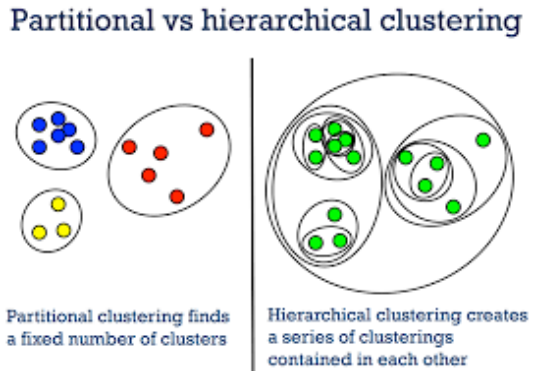
Agrupamiento jerárquico Definimos distancias entre grupos de observaciones a partir de distancias entre puntos $\mathbf{x}_i$.



Agrupamiento jerárquico

Diferencias entre agrupamiento jerárquico y agrupamiento particional.

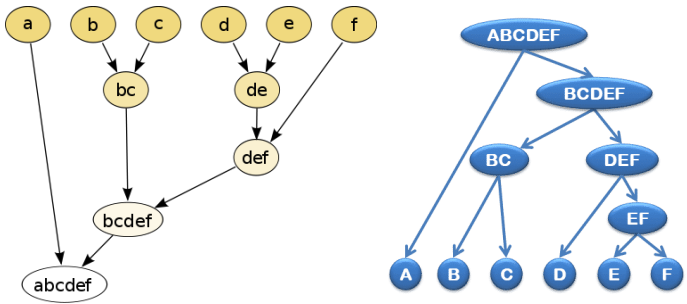
- En el particional, los grupos son (o suelen ser) disjuntos.
- En el esquema jerárquico, los grupos están encadenados.



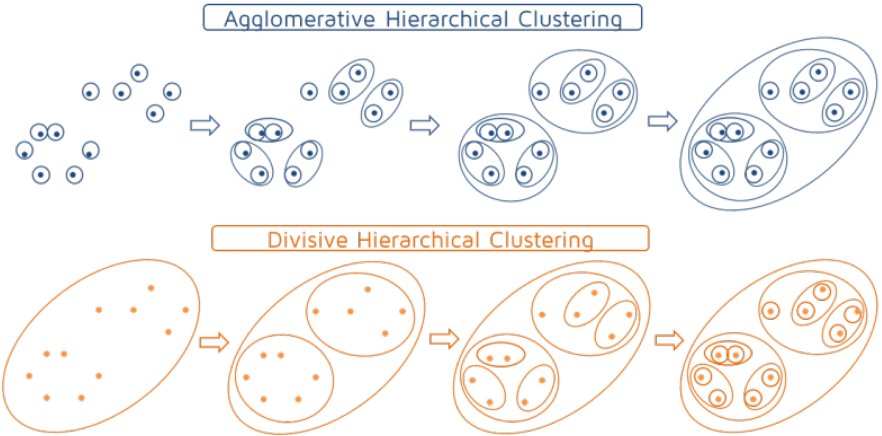
Agrupamiento jerárquico

Hay dos esquemas

- Aglomerativo o *bottom up*: se inicia con cada punto es un clúster, y en cada iteración se van agrupando.
- Divisivo o *top down*: se inicia con únicos cluster conteniendo todos los puntos, y en cada iteración se van separando.



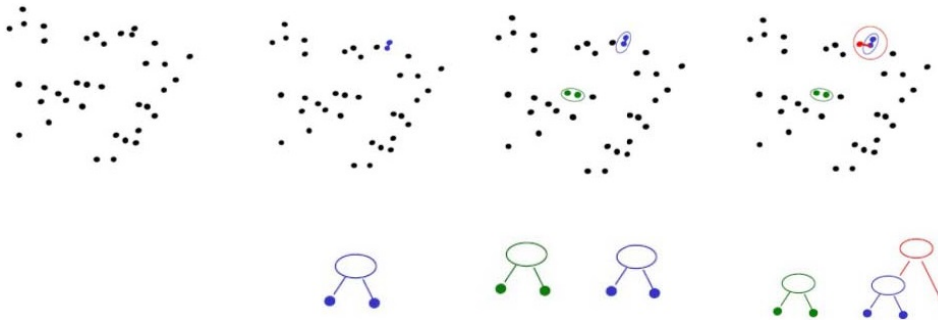
Agrupamiento jerárquico



Agrupamiento jerárquico

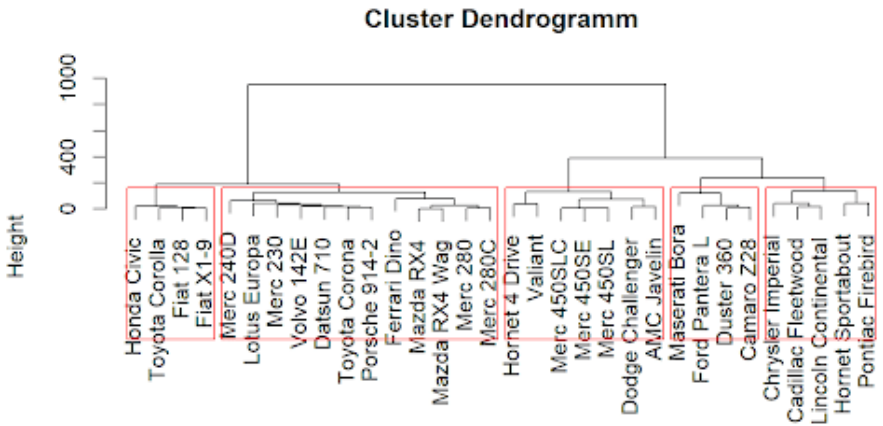
Idea en el esquema aglomerativo:

- Definir cada dato como un cluster.
- Repetir hasta tener un sólo cluster: Unir los dos clústers más cercanos según $d(\cdot, \cdot)$ en un sólo clúster nuevo.



Agrupamiento jerárquico

La salida típica viene en forma de un dendrograma.



Agrupamiento jerárquico

En general $d(\cdot, \cdot)$ puede ser una distancia, o una versión más relajada (e.g., discrepancia, disimilitud).

Cuidado! No siempre se obtiene un dendrograma. La función $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ debe cumplir ciertas condiciones:

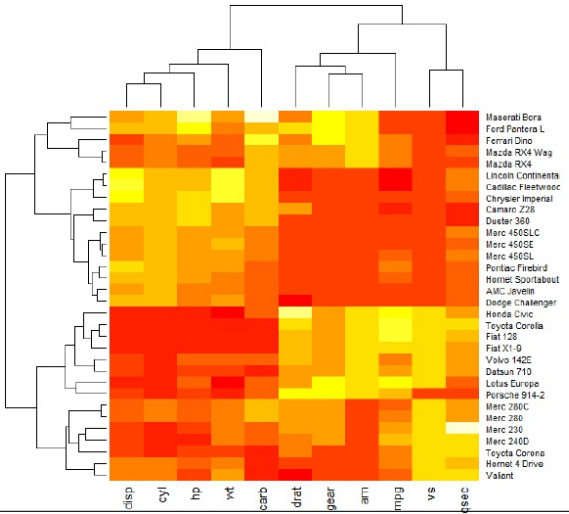
1. (simetría) $d(a, b) = d(b, a)$
2. (desigualdad triangular) $d(a, c) < \max\{d(a, b), d(b, c)\}$
3. (no negatividad) $d(a, b) \geq$ o ?? No es necesario. Pero sí debe cumplir que d está limitada inferiormente.

Agrupamiento jerárquico

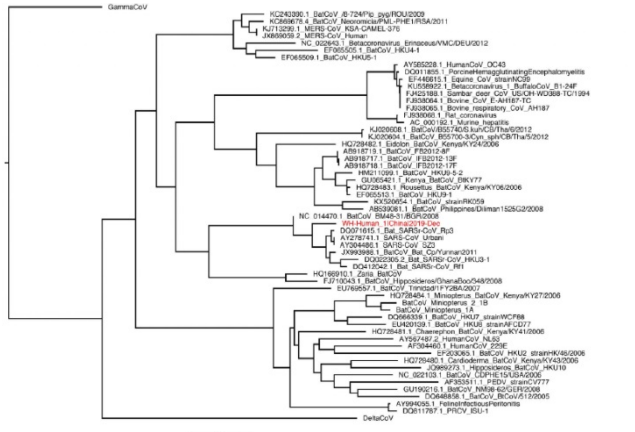
Métricas comunes:

- distancia euclideana $d(a,b) = ||a-b||_2 = \Big(\sum_{i=1}^{d}(a_i - b_i)^2\Big)^{1/2}$.
- distancia euclideana cuadrada $d(a,b) = ||a-b||_2^2 = \sum_{i=1}^{d}(a_i - b_i)^2$.
- distancias de Minkowski $d(a,b) = ||a-b||_p = \Big(\sum_{i=1}^{d}(a_i - b_i)^p\Big)^{1/p}$.
- Norma 1 o distancia *Manhattan* $d(a,b) = ||a-b||_1 = \sum_{i=1}^{d}|a_i - b_i|$.
- distancia de Mahalanobis $d(a,b) = \big((\mathbf{a}-\mathbf{b})^T\Sigma^{-1}(\mathbf{a}-\mathbf{b})\big)^{1/2}$.

Agrupamiento jerárquico

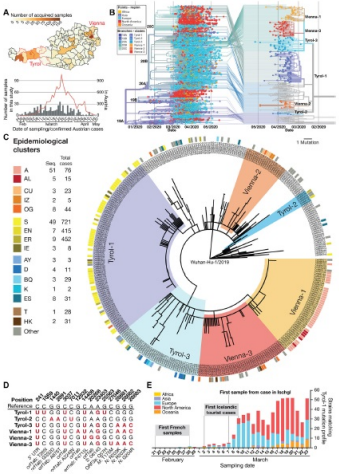


Agrupamiento jerárquico



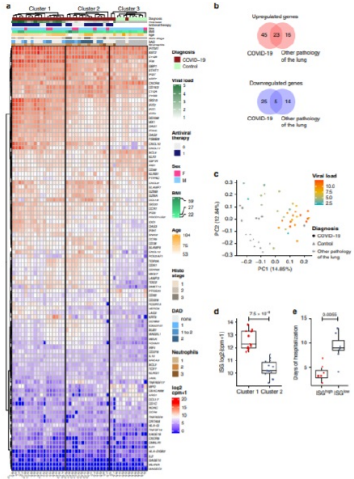
Preliminary maximum likelihood phylogenetic analysis of novel Wuhan, China human CoV in red, GenBank (accession MN908947)
Novel CoV seq data from: http://virological.org.t/initial-genome-release-of-novel-coronavirus/319. The Shanghai Public Health Clinical Center & School of Public Health, in collaboration with the Central Hospital of Wuhan, Huazhong University of Science and Technology, the Wuhan Center for Disease Control and Prevention, the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control, and the University of Sydney, Sydney, Australia.

Agrupamiento jerárquico



https://stm.sciencemag.org/content/12/573/eabe2555/tab-figures-data

Agrupamiento jerárquico



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Agrupamiento jerárquico

¿Cómo elegir $d(\cdot, \cdot)$? Existen muchos métodos

- *Maximum or complete-linkage*

$$d(A, B) = \max\{d(a, b) : \ a \in A, \ b \in B\}.$$

- *Minimum or single-linkage*

$$d(A, B) = \min\{d(a, b) : \ a \in A, \ b \in B\}.$$

- *Averaged linkage*

$$d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b).$$

Agrupamiento jerárquico

¿Cómo elegir $d(\cdot,\cdot)$? Existen muchos métodos

- *Weighted averaged linkage*

$$d(\{i\} \cup \{j\}, k) = \frac{d(i,k) + d(j,k)}{2}.$$

- *Centroid linkage*

  $d(A,B) = ||c_A - c_B||$, c_A, c_B son los centroides de los clúster A, B, resp.

- *Minimum energy clustering*

$$d(A,B) = \frac{2}{|A| \cdot |B|} \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} d(a_i, b_j) - \frac{1}{|A|^2} \sum_{i,j=1}^{|A|} d(a_i, a_j) - \frac{1}{|B|^2} \sum_{i,j=1}^{|B|} d(b_i, b_j).$$

Agrupamiento jerárquico

¿Cómo elegir $d(\cdot, \cdot)$? Existen muchos métodos

- *Maximum or complete-linkage*

$$d(A, B) = \max\{d(a, b) : \ a \in A, \ b \in B\}.$$

- *Minimum or single-linkage*

$$d(A, B) = \min\{d(a, b) : \ a \in A, \ b \in B\}.$$

- *Averaged linkage*

$$d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b).$$

Agrupamiento jerárquico

Otros criterios de *linkage* incluyen:

- La suma de todas la varianzas intragrupos.
- El aumento de la varianza para el grupo que se fusiona (criterio de Ward).
- La probabilidad de que los grupos candidatos se generen a partir de la misma función de distribución (*V-linkage*).
- El producto de grados de entrada y de salida en un grafo de *k*-vecinos más cercanos.
- El incremento de algún descriptor de conglomerado (es decir, una cantidad definida para medir la calidad de un conglomerado) después de fusionar dos conglomerados.

Agrupamiento jerárquico

El método de Ward:
El criterio de varianza mínima de Ward minimiza la varianza total intragrupos: en cada paso encuentra el par de grupos que conduzca a un aumento mínimo en la varianza total dentro del grupo después de la fusión. Este aumento es una distancia al cuadrado ponderada entre los centros de los conglomerados. AL inicio, se la distancia inicial entre objetos individuales como la norma euclidiana al cuadrado.

En la práctica, se utiliza el algoritmo de Lance-Williams:

- Suponga que los clústers C_i, C_j son los siguientes a fusionarse. La siguiente fórmula produce la actualización de las distancias $d_{ij} = d(C_i, C_j)$.
- $d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d_{ik} + \frac{n_j + n_k}{n_i + n_j + n_k} d_{jk} - \frac{n_i + n_j}{n_i + n_j + n_k} d_{ij}.$

Ejemplo

Ejemplo 1:

	a	b	c	d
a	0	5	6.1	7
b	5	0	4	6.2
c	6.1	4.0	0	6
d	7	6.2	6	0

Agrupamiento jerárquico

Ejemplo 2:

	a	b	c	d
a	0	5.0	5.6	7.2
b	5.0	0	4.6	5.7
c	5.6	4.6	0	4.9
d	7.2	5.6	4.9	0