DIPLOMADO
Ciencia de Datos

Módulo 7
Aprendizaje de máquina no supervisado
1. Clustering, aglomerativo

Eduardo Espinosa Avila

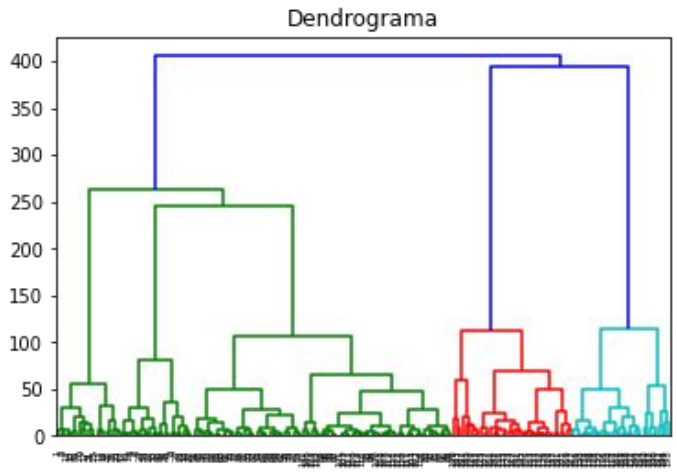
DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Agrupamiento (*clustering*)

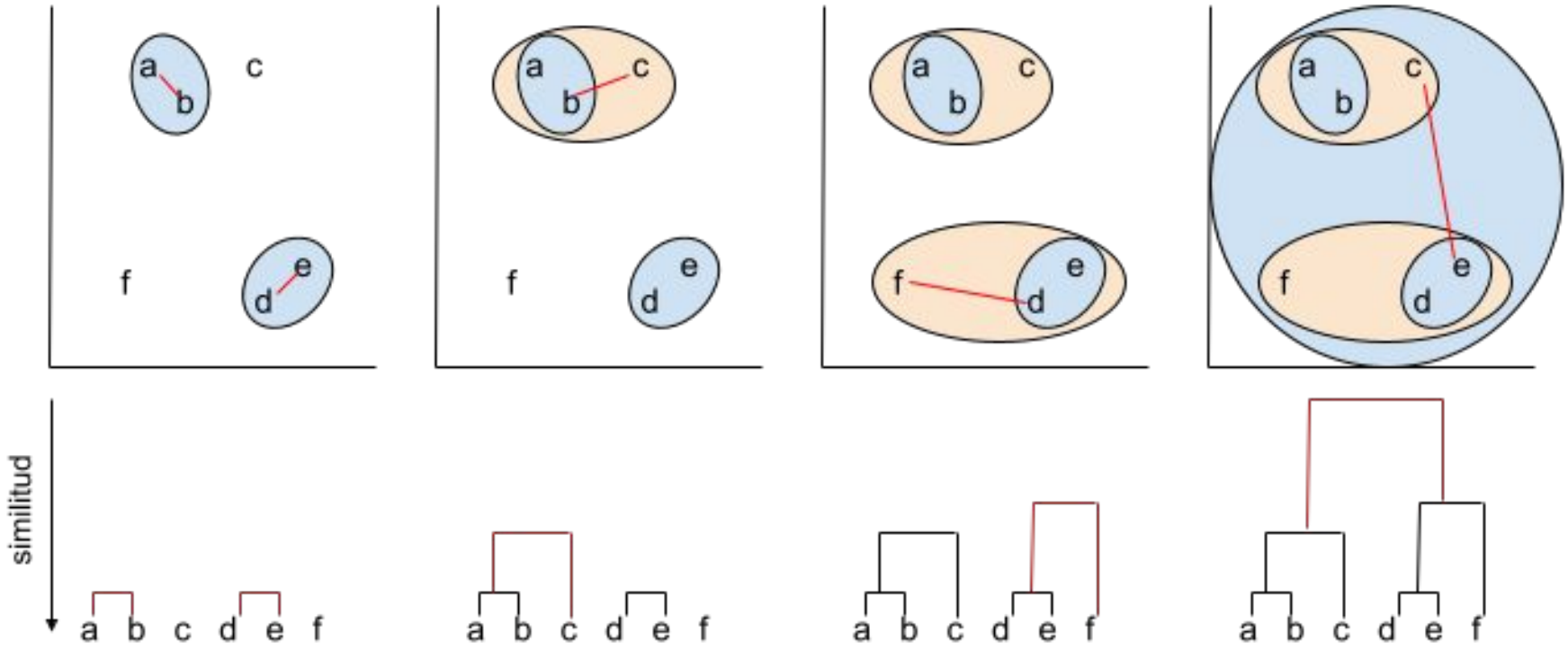
En ocasiones es posible dividir una colección de observaciones en distintos subgrupos, basados únicamente en los atributos de las observaciones.

La intención es realizar división de datos en grupos (clusters) de observaciones que son más similares dentro de un grupo que entre varios grupos.

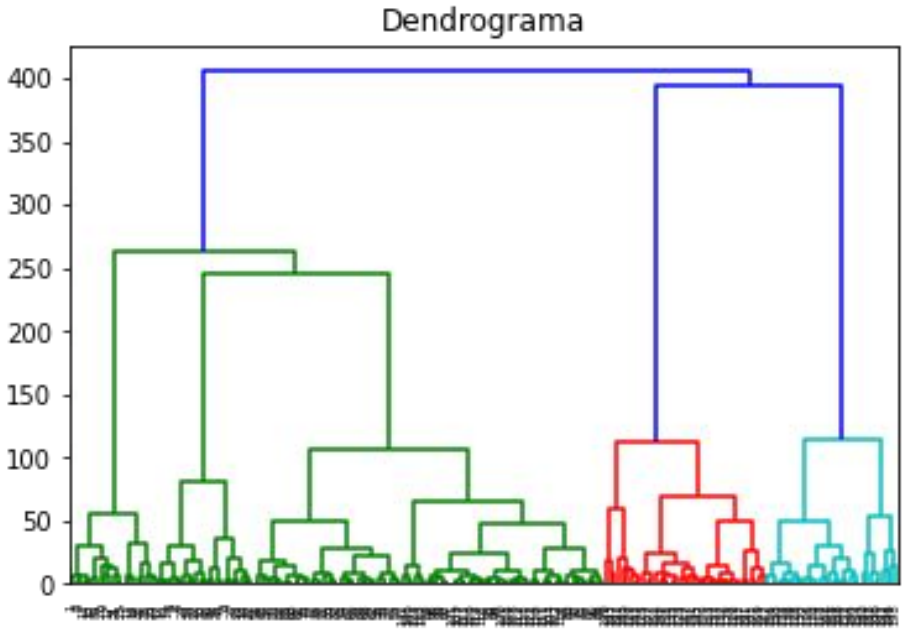
Los grupos son formados ya sea agregando observaciones o dividiendo un gran grupo de observaciones en una colección de grupos más pequeños.

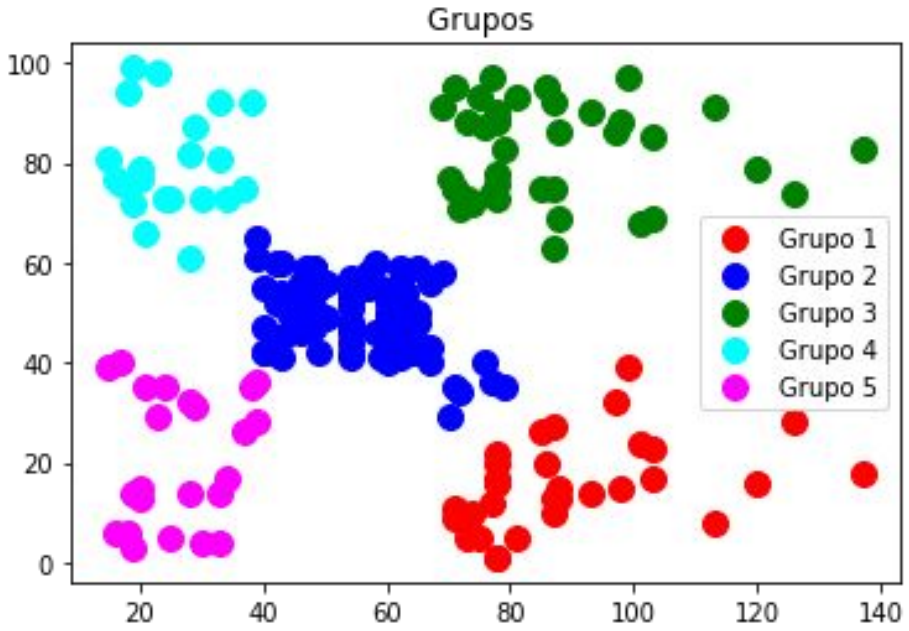


Agrupamiento aglomerativo



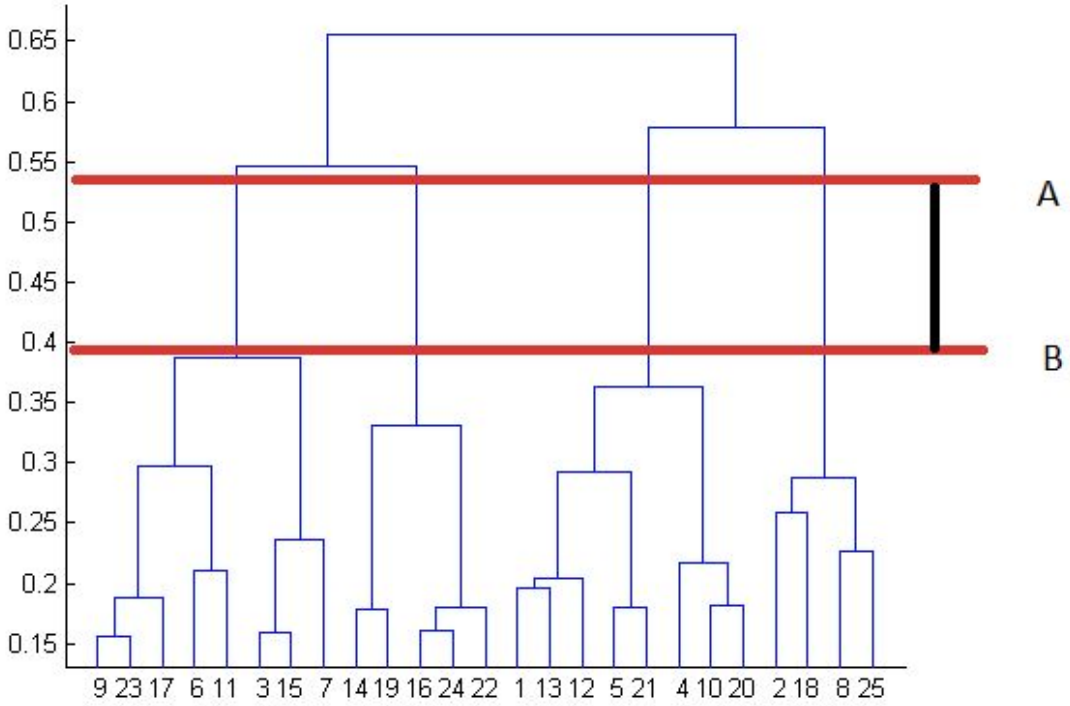
Agrupamiento aglomerativo, definición





A diferencia de *k-means*, este algoritmo no se establece una cantidad *k* de grupos: reduce iterativamente el número de grupos al combinarlos.

Agrupamiento aglomerativo, definición

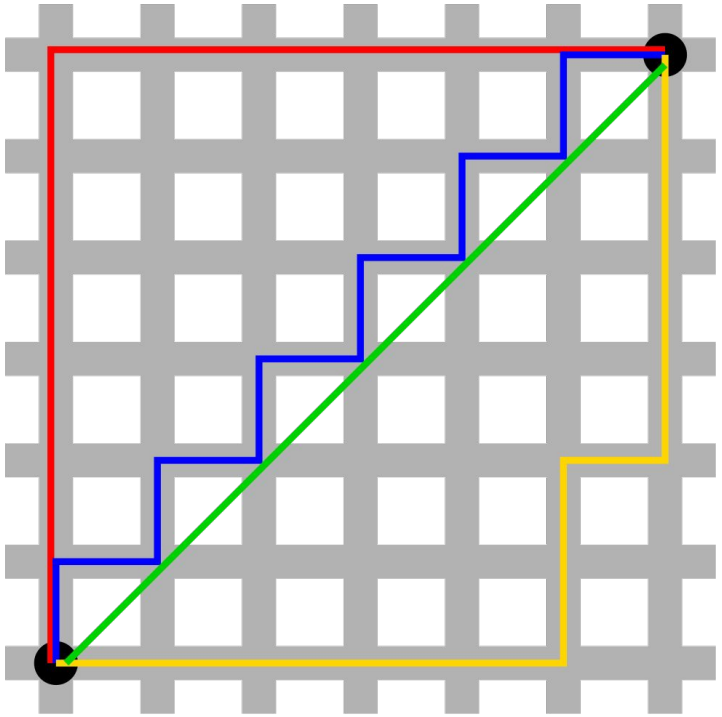


El historial de los agrupamientos se puede observar directamente en el *dendrograma*.

Con este diagrama se obtiene el agrupamiento óptimo.

Agrupamiento aglomerativo, definición

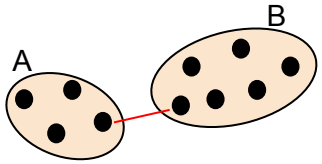
Al igual en que *k-means* se puede usar la distancia euclidiana y la métrica Manhattan para determinar los grupos que habrán de formarse.



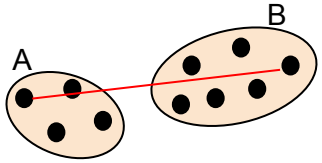
Agrupamiento aglomerativo, enlazamientos

Existen cuatro criterios de enlazamiento (*linkage*) que se utilizan para combinar los grupos:

- *Single*: distancia menor.



- *Complete*: distancia mayor.



- *Average*: distancia promedio. $L(A,B) = \frac{1}{n_A n_B} \sum_{k=1}^{n_A} \sum_{l=1}^{n_B} d_v(\mathbf{x}_k, \mathbf{x}_l), \mathbf{x}_k \in A, \mathbf{x}_l \in B$

- *Ward*: suma de las diferencias cuadráticas. $d_v(A,B) = (\overline{\mathbf{x}}_A - \mathbf{x}_k)^2 + (\overline{\mathbf{x}}_B - \mathbf{x}_l)^2, \mathbf{x}_k \in A, \mathbf{x}_l \in B$

Agrupamiento aglomerativo, aplicaciones

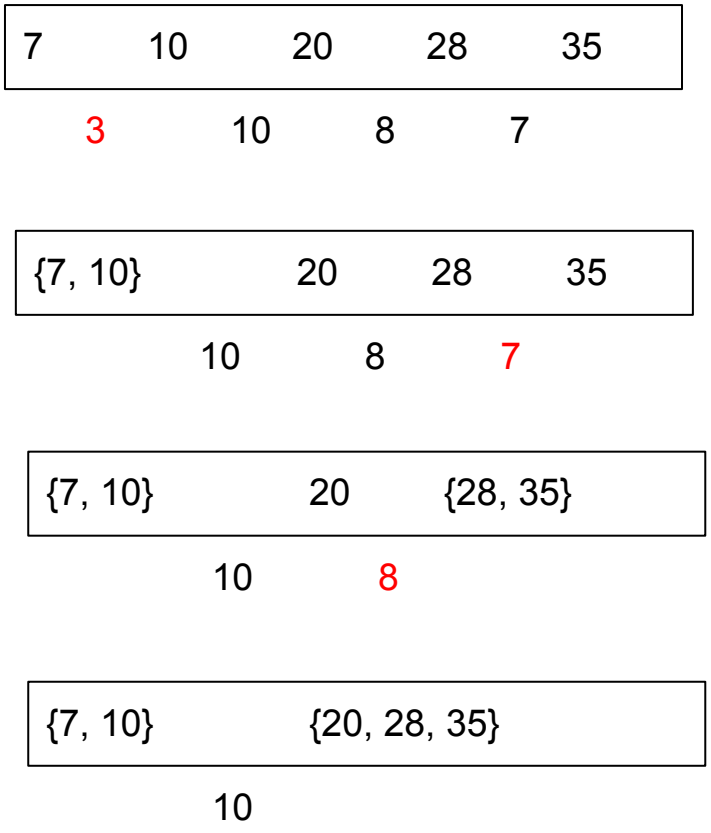
El agrupamiento aglomerativo puede utilizarse en muchas situaciones, por ejemplo:

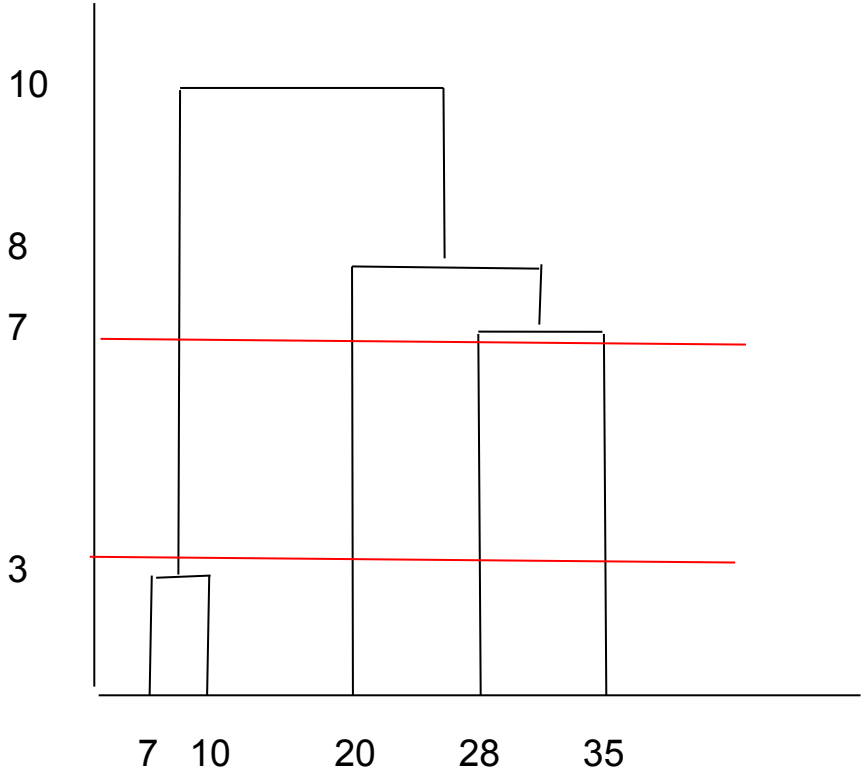
- Identificación de *fake news*
  - https://stanford.io/3osDY9T
  - https://bit.ly/3whr5ll
- Mercadotecnia y ventas
- Identificación de actividades fraudulentas
  - https://bit.ly/3oqTGm7

- Análisis y organización de documentos, Filtrado de *spam*

Agrupamiento aglomerativo, ejemplo

Para el conjunto de datos unidimensionales {7,10,20,28,35} , obtener el agrupamiento jerárquico y dibujar el dendrograma correspondiente.





Contacto

Dr. Eduardo Espinosa Avila

laloea@fisica.unam.mx

Tels: 5556225000 ext. 5003

Redes sociales:

https://twitter.com/laloea

https://www.linkedin.com/in/eduardo-espinosa-avila-84b95914a/

Referencias

- Steele, Brian, Chandler, John & Reddy, Swarna
  Algorithms for Data Science / Brian Steele, John Chandler and Swarna Reddy -- Switzerland : Springer, 2016
  1 recurso en línea (430 páginas)
  https://link-springer-com.pbidi.unam.mx:2443/book/10.1007/978-3-319-45797-0

- Seyedmehdi Hosseinimotlagh, Evangelos E. Papalexakis
  Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles
  http://snap.stanford.edu/mis2/files/MIS2_paper_2.pdf

- Yong Ge, Hui Xiong, Chuanren Liu, Zhi-Hua Zhou
  A Taxi Driving Fraud Detection System
  https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm11.pdf