Clustering

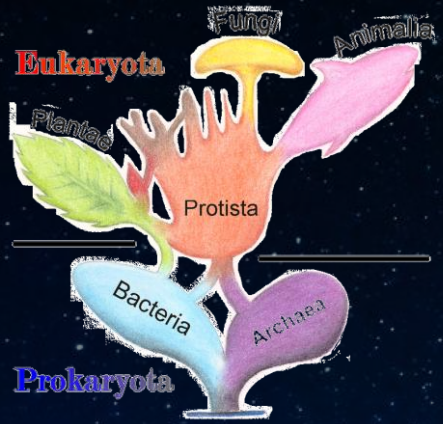
Aprendizaje No Supervisado

Aprendizaje No Supervizado: Generalmente se habla de algoritmos de Agrupamiento (Clustering), que tiene como objetivo realizar una separación de elementos en diferentes grupos que poseen ciertas características en común.

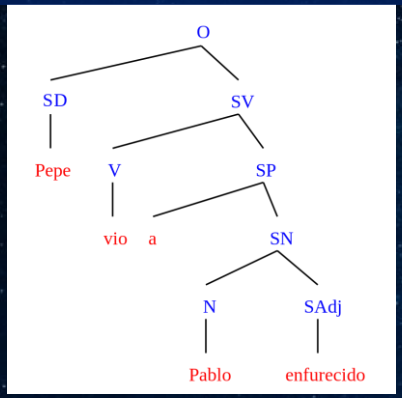
Datos: El algoritmo de entrenamiento no utiliza una variable objetivo (target)

Aprendizaje No Supervisado

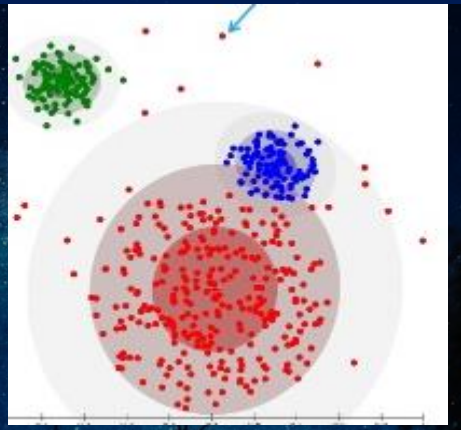
Clasificación de especies



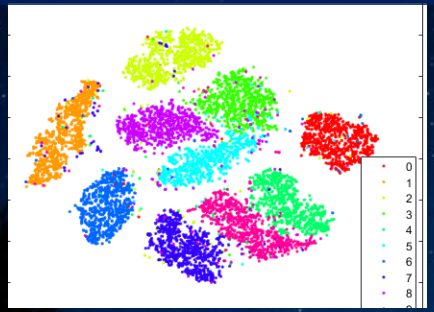
Proc. lenguaje natural



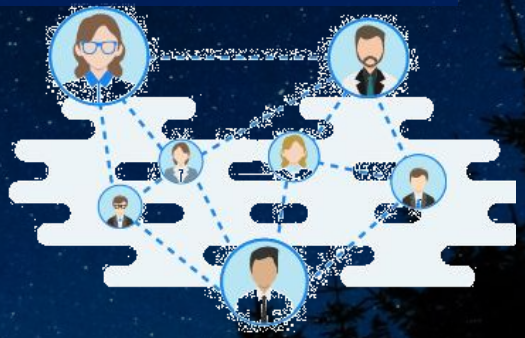
Detección de anomalías



Reducción de Dimensionalidad



Redes sociales/Mercado



Datos astronómicos

Agrupamiento (clustering)

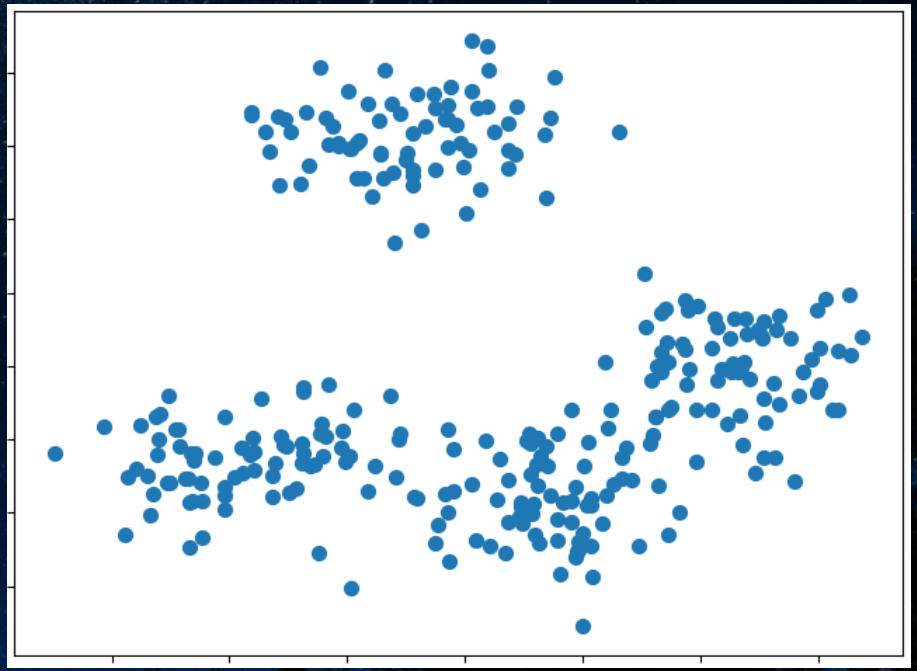
¿cómo agruparía las bolas del juego de Pool en diferentes conjuntos?

¿Qué característica necesito?
¿Cuántos conjuntos necesito?

Podrían agruparse por color (rojas, amarillentas, azuladas).
En orden por sus valores.
Por su función en el juego (lisas, rayadas).

Agrupamiento (clustering)

¿cómo agruparía los siguientes puntos?

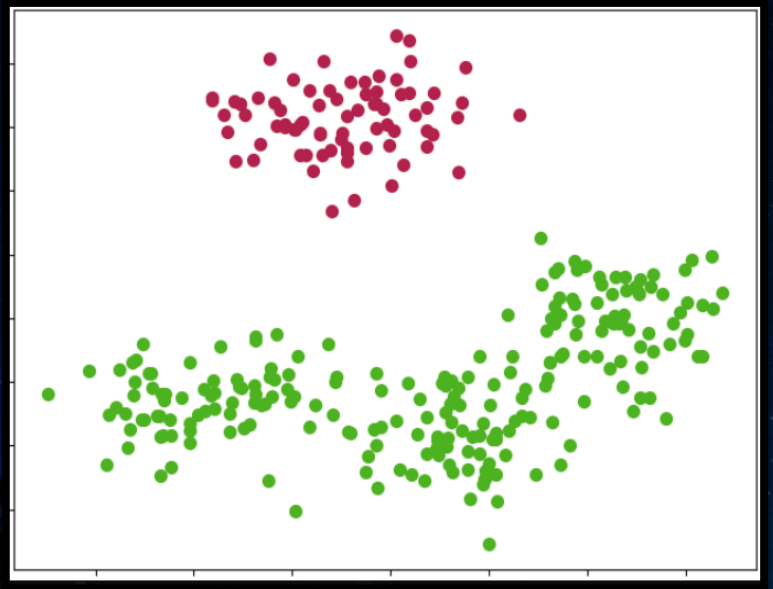


Agrupamiento (clustering)

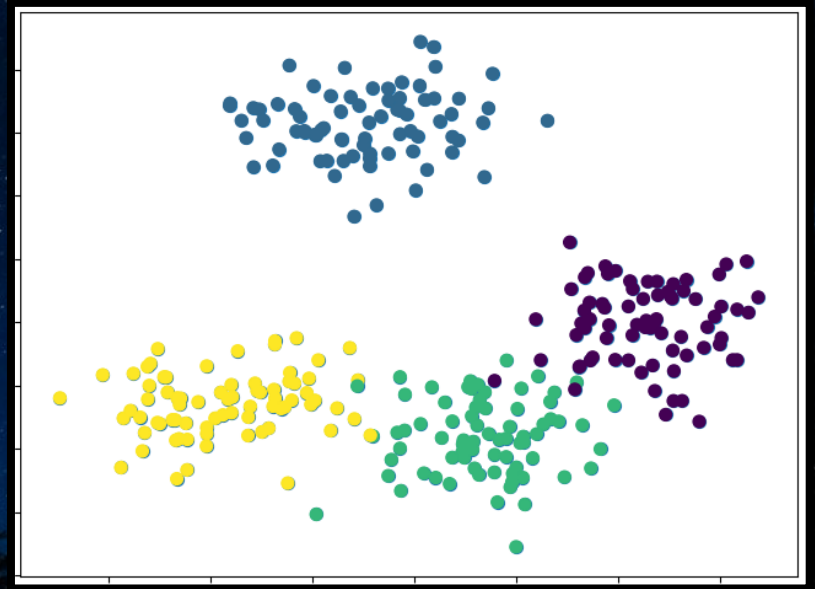
Diferentes grupos sobre los mismos datos

¿Uno es mejor que el otro?

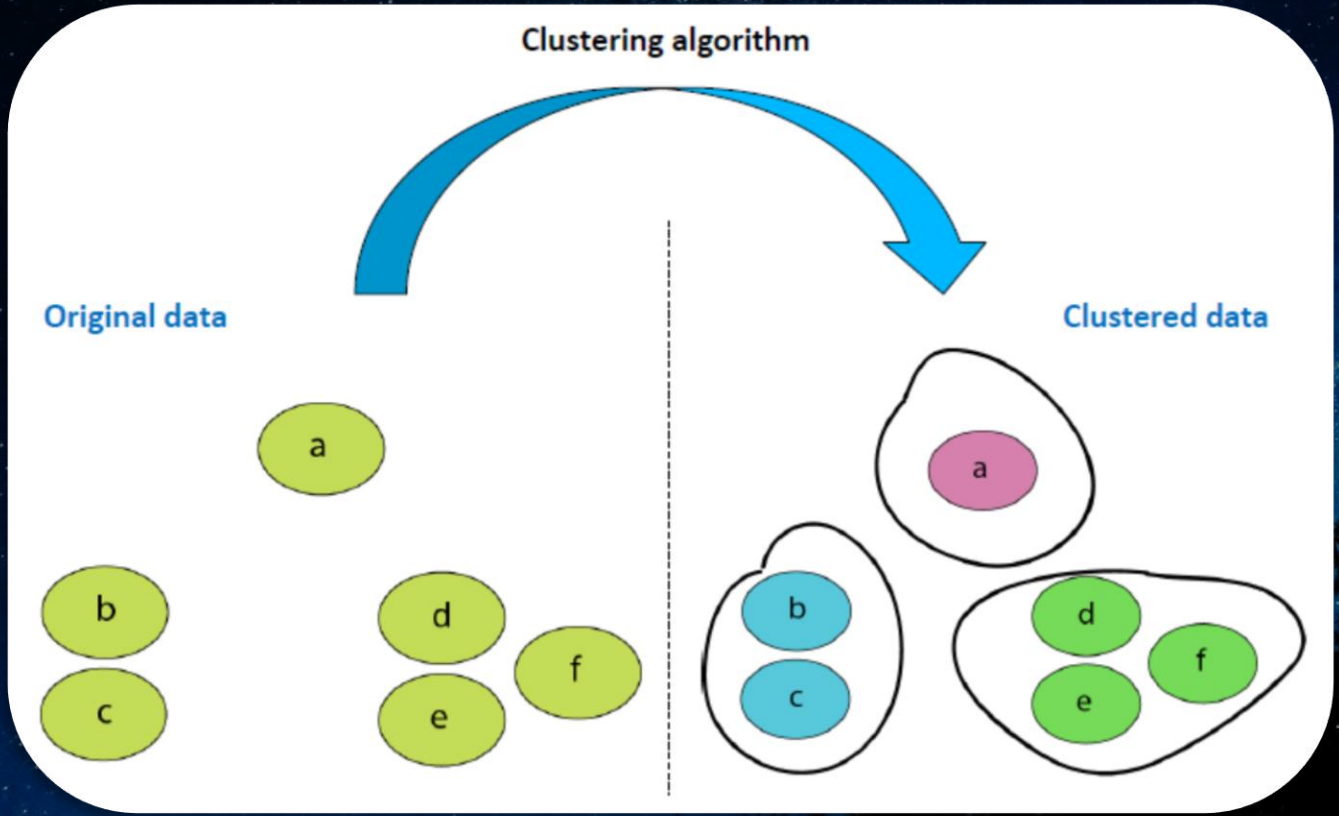
Ejemplo con 2 grupos



Ejemplo con 4 grupos



Objetivo del Clustering

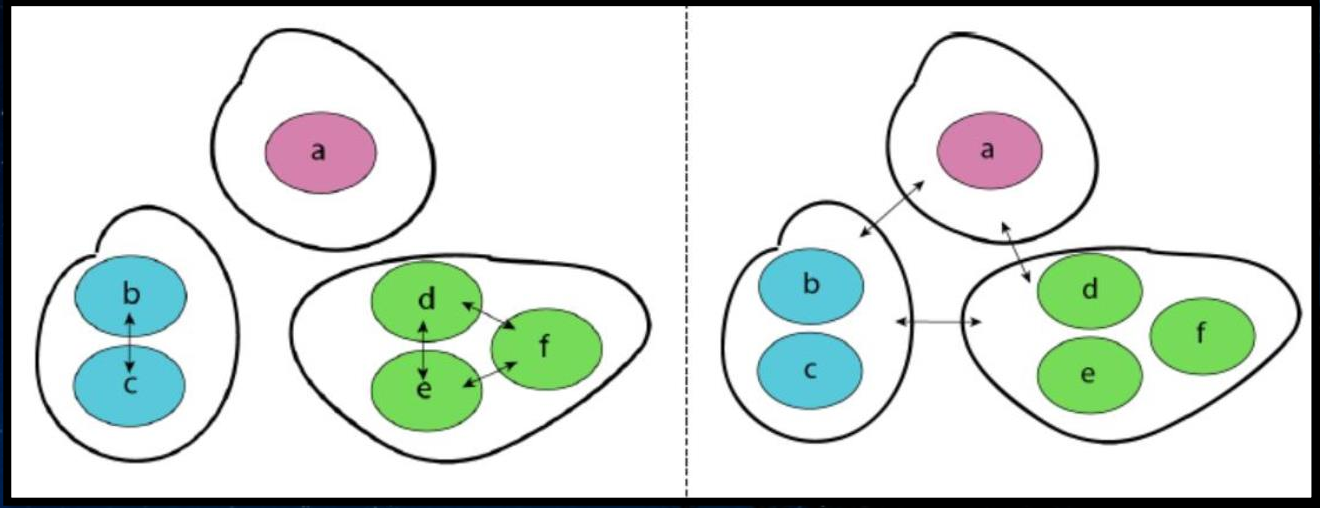


Objetivo del Clustering

Objetivo general:

Minimizar distancia intra-cluster (ejemplos de un mismo cluster)

Maximizar distancia inter-cluster (ejemplos de distintos clusters)



Distancias del Clustering

Distancia Euclidea

- Euclidea(X,Y)= $\sqrt{[\sum_i^n (x_i-y_i)^2]}$
- La más utilizada

Distancia Manhattan

- Manhattan(X,Y)=$\sum_i^n |x_i-y_i|$
- Específica para algunos problemas

Otras

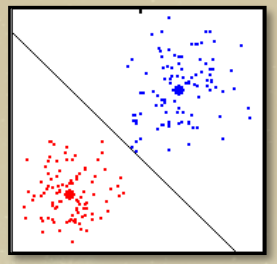
- Chebyshev: Ejemplo Rey ajedrez.

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

Tipos de algoritmos de Clustering

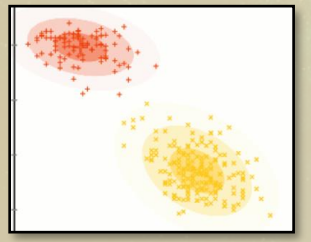
Partitivos (Duros, hard)

- Particionan los datos en K grupos.
- Una instancia pertenece a un único grupo.



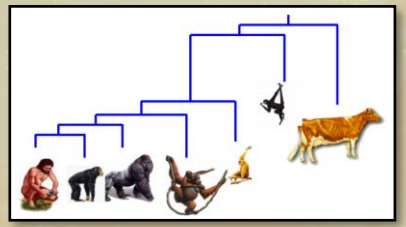
Probabilísticos (blandos, fuzzy)

- Modelan densidad de los ejemplos
- P(x|C): prob. de que x pertenezca a C



Jerárquicos

- Dendograma: Grupos -> Subgrupos -> Subgrupos…



Basados en densidad

- Ej: DBSCAN