Curso DM

Clustering

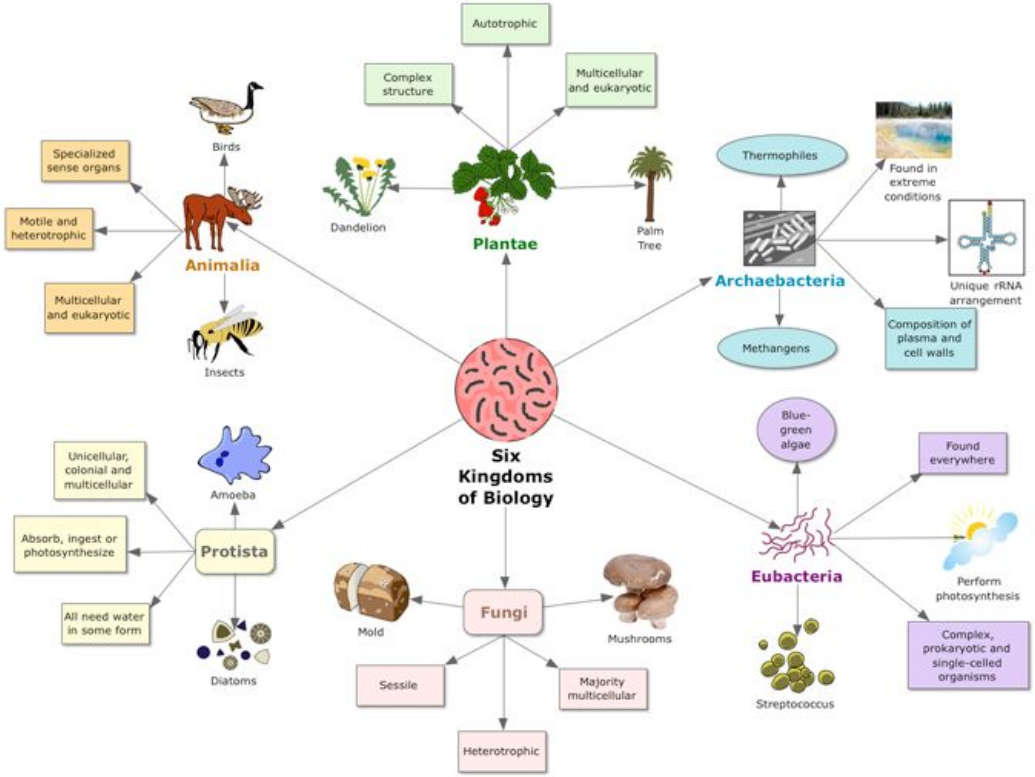
Parte 1: Introducción

Primavera 2023

Profesoras: Jazmine Maldonado y Cinthia Sánchez

Basada en versiones previas de Felipe Bravo y Bárbara Poblete

Inspiración histórica (taxonomía)



¿Qué es el clustering?

Técnica para encontrar grupos de objetos tal que:

- Los objetos en un grupo sean similares (o relacionados) entre sí y
- que sean diferentes (o no relacionados) a los objetos en otros grupos

¿Cuándo y para qué usar clustering?

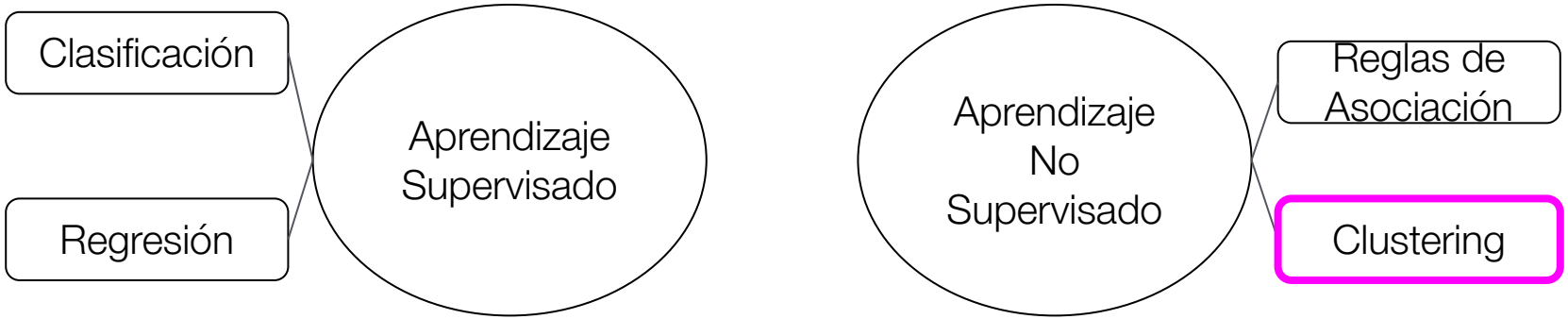
Cuando necesitemos dividir nuestros datos en grupos que sean: **significativos y/o útiles**

- Debemos preocuparnos de capturar la estructura natural de los datos
- A veces es sólo un punto de partida

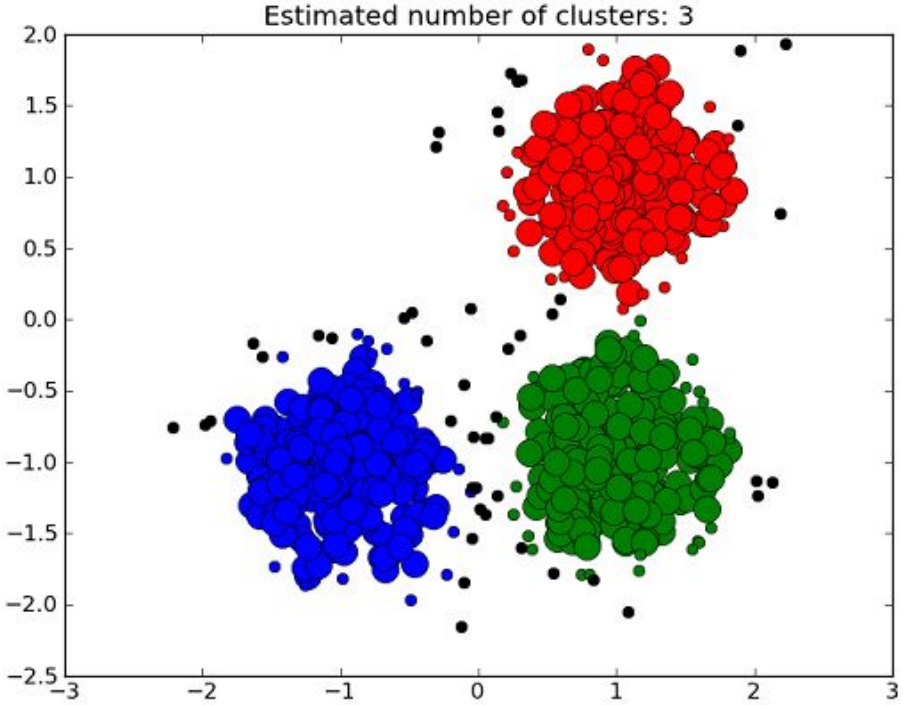
Clasificación vs. Clustering

- Clasificación: aprendizaje supervisado
- Clustering: aprendizaje no-supervisado (no requiere etiquetas)

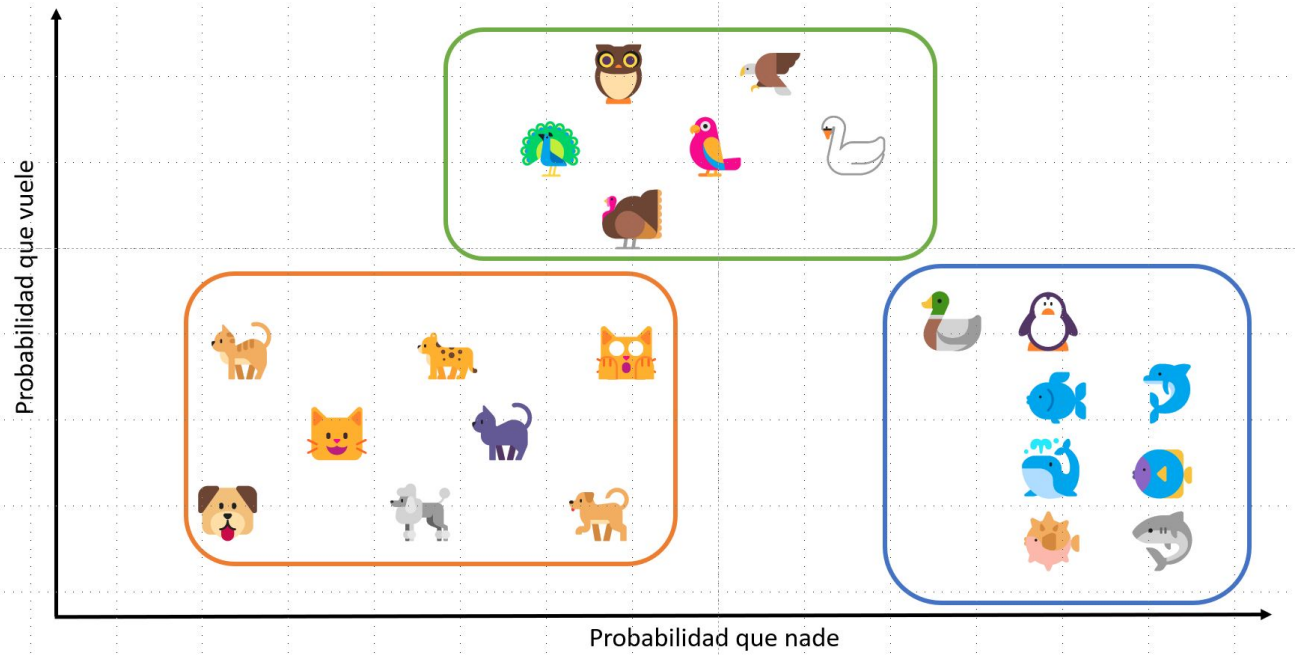
Tipos de aprendizaje y sus tareas:



Busca capturar agrupaciones naturales en los datos



Busca capturar agrupaciones naturales en los datos



Fuente: Repositorio Github Pablo Badilla. https://github.com/pbadillatorrealba/MDS7202

Análisis de clusters es una tarea esencial para muchas aplicaciones

Por ej:

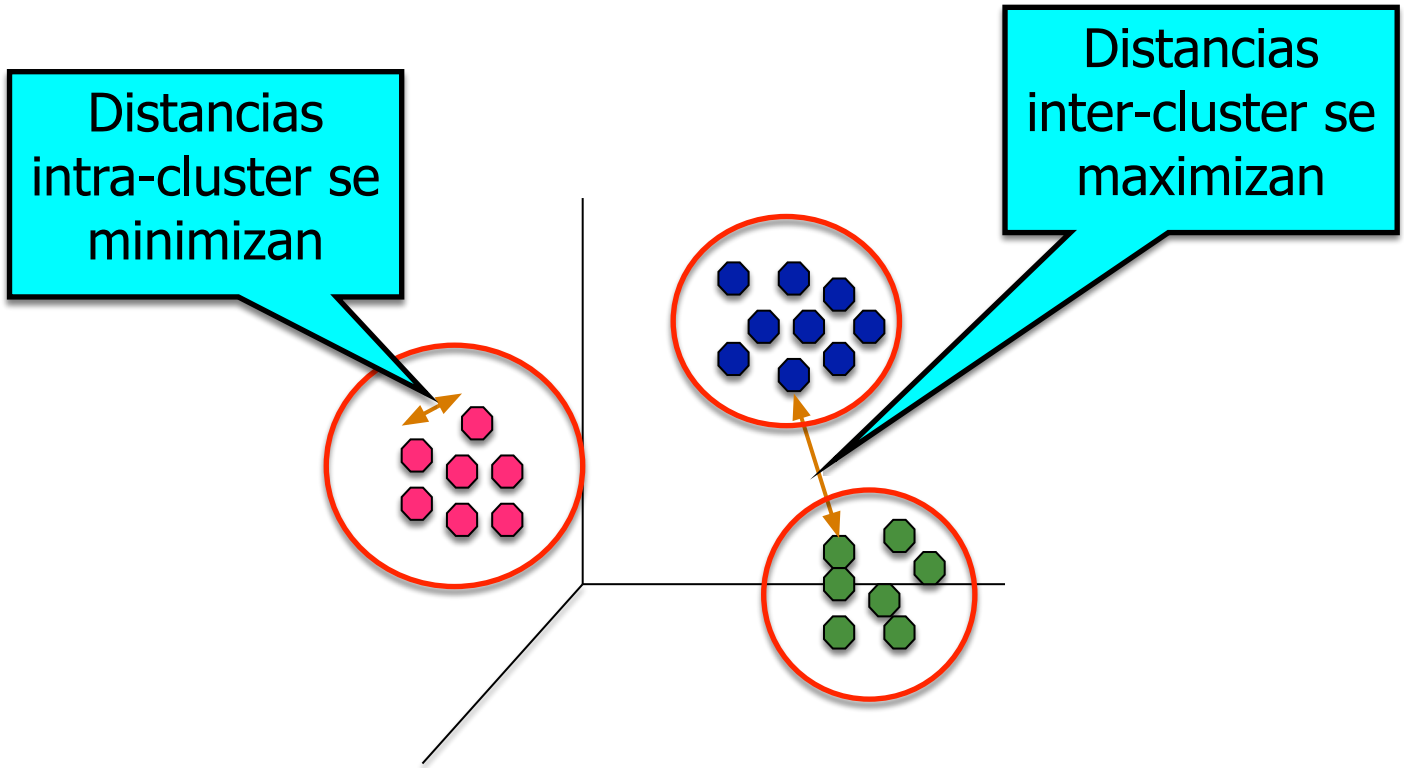
- Encontrar clusters naturales y describir sus propiedades (**data understanding**)
- Encontrar agrupamientos útiles (**data class identification**)
- Encontrar representantes para grupos homogéneos (**data reduction**)
- Encontrar objetos inusuales (**outliers detection**)
- Encontrar perturbaciones aleatorias de los datos (**noise detection**)

Formulación del problema

- Dado un conjunto de puntos, organizarlos en clusters (grupos, clases).
- Clustering: el proceso de agrupar objetos físicos en clases de objetos similares

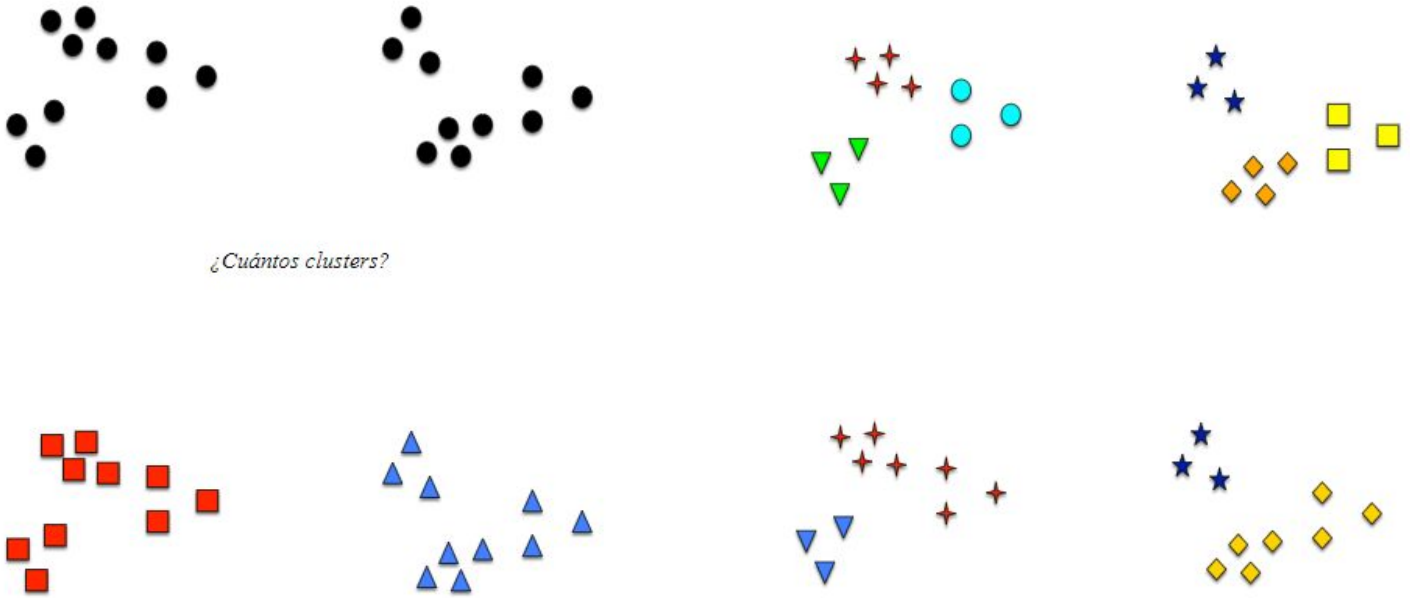
source: http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

¿Qué es análisis de clusters?



Los grupos de elementos se identifican sólo en base a las características que tienen los datos.

La noción de cluster puede ser ambigua



La noción de cluster puede ser ambigua

¿Cuántos clusters?

Seis Clusters

Dos Clusters

Cuatro Clusters

Tipos de clustering

- ¿Qué es un clustering? Es un conjunto de clusters.

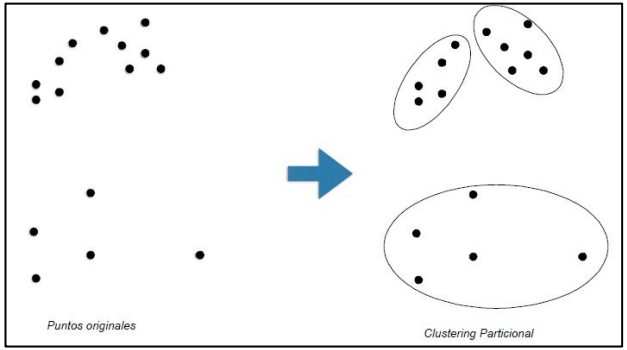
- Distinción importante entre conjuntos de clusters jerárquicos y particionales

Tipos de clustering

Tipos de clustering

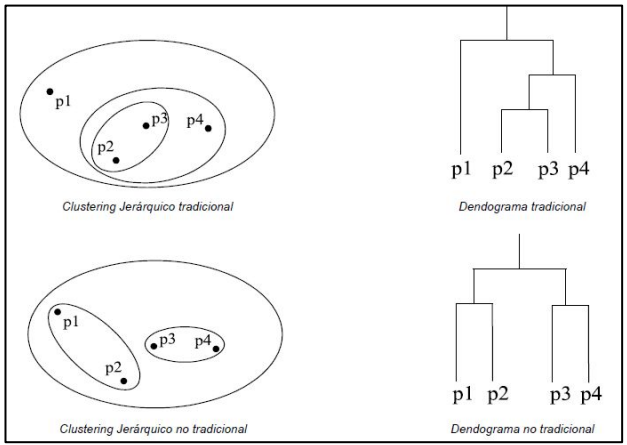
Particional

Divide los datos en subconjuntos sin traslape (clusters), tal que cada dato está en un solo subconjunto.



Jerárquico

Un conjunto de clusters anidados, organizados como un árbol.

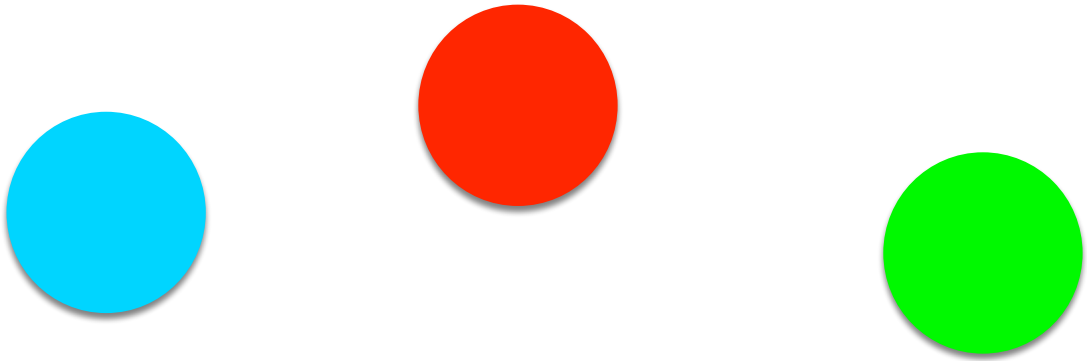


Tipos de clusters

- Bien separados
- Basados en un centro
- Contiguos
- Basados en densidad
- Propiedad o Conceptual

Clusters bien separados

Un cluster es un conjunto de puntos, tal que: cualquier punto en un cluster está más cerca (es más similar) **a cualquier otro punto en el mismo cluster** que a cualquier punto fuera de este.



Clusters basados en un centro

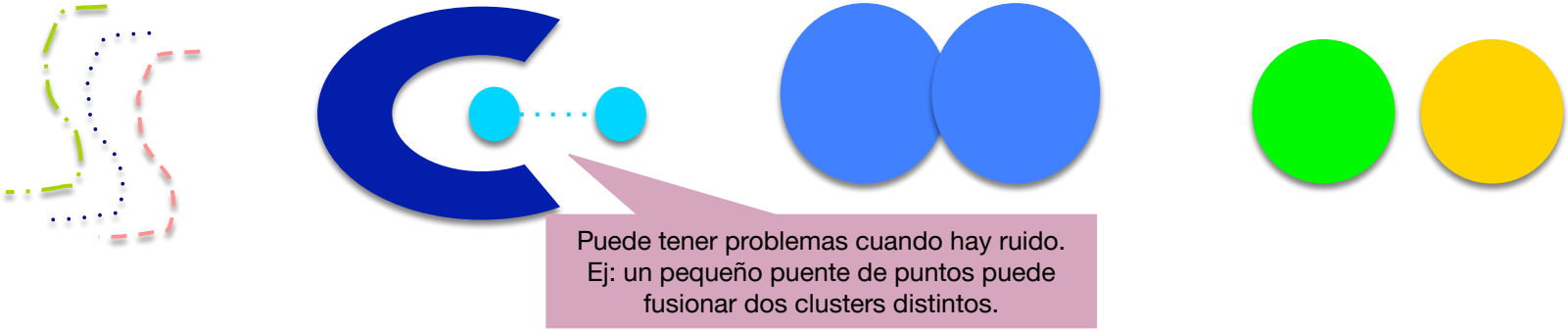
Un cluster es un conjunto de objetos, tal que: un objeto dentro del cluster está más cerca (es más similar) **al centro de este cluster** que al centro de cualquier otro.

El centro de un cluster puede ser el centroide, el promedio de todos los puntos en el cluster, o el medioide, el punto más “representativo” del cluster

Clusters contiguos
(vecino más cercano o transitivo)

Un cluster es un conjunto de puntos, tal que: cada punto en un cluster está más cerca (es más similar) **a uno o más puntos en el cluster** que a cualquier punto no en el cluster. Esta definición es útil cuando los clusters son **irregulares** o están **entrelazados**.

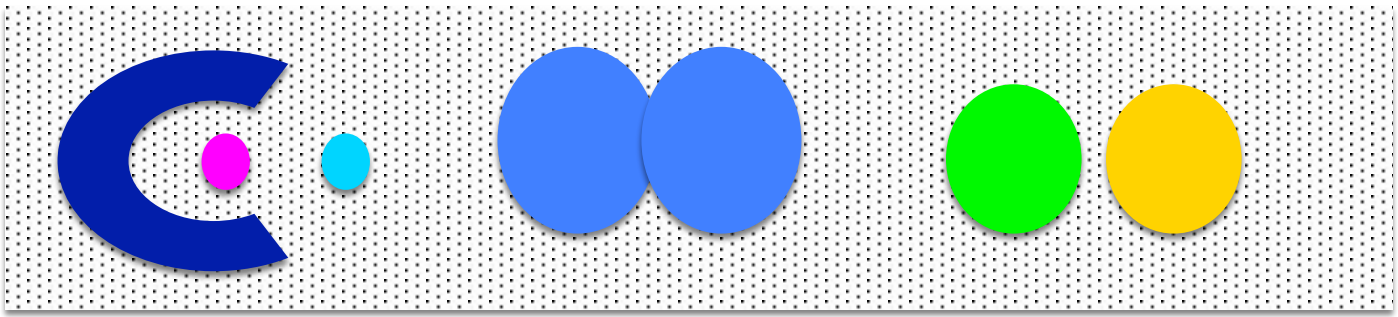
- Un cluster puede definirse como un componente conectado (dos objetos están conectados sólo si se encuentran a una distancia determinada el uno del otro).



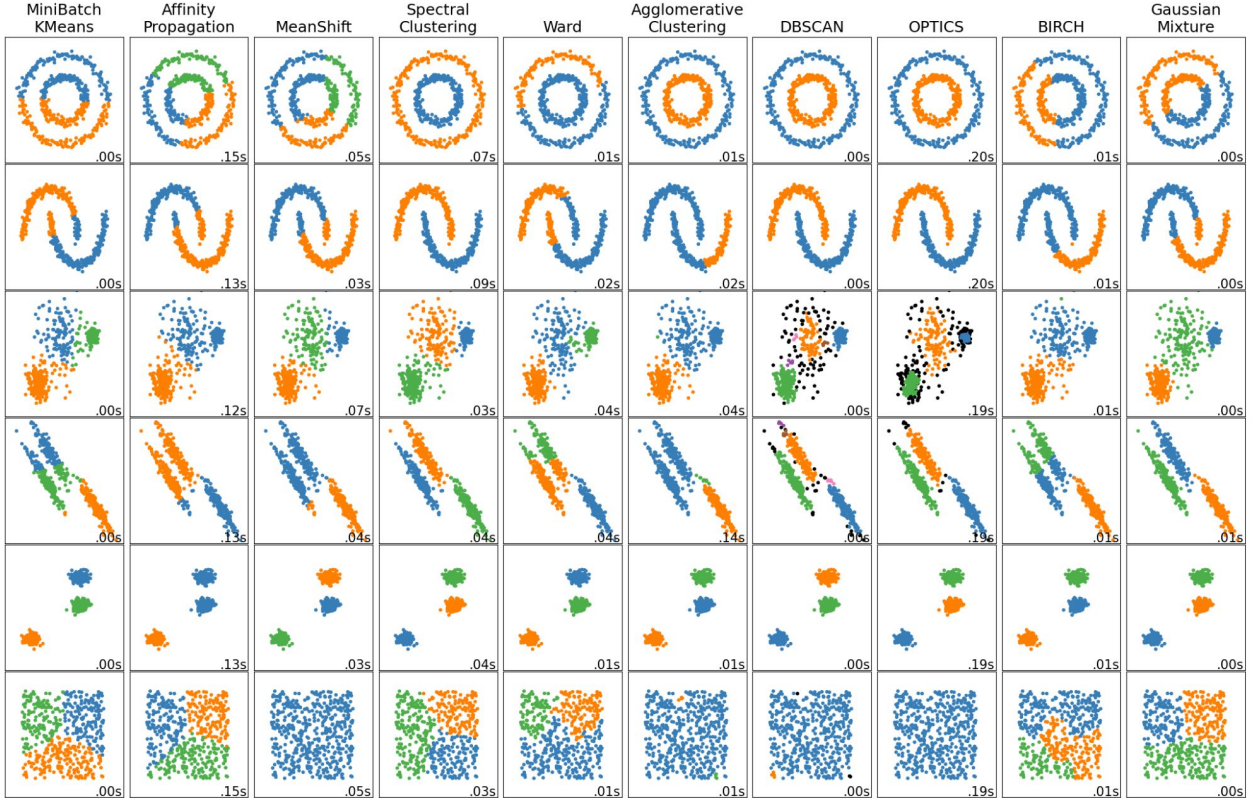
Clusters basados en densidad

Un cluster es una región densa de puntos, separada por regiones de baja densidad de otras regiones de alta densidad.

Usado cuando los clusters son irregulares o están entrelazados, y cuando hay ruido y outliers



Algoritmos de clustering



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Ejemplos y Casos de Uso

Plataformas Digitales

- Agrupar usuarios para generar recomendaciones
- Agrupar contenido para facilitar la navegación a los usuarios
- Detectar grupos en secuencias de interacciones (clicks) para detectar mejoras que optimicen el sitio
- Detección de spam o detección de comportamientos maliciosos

Ciencias

- Agrupar estrellas en base a su brillo
- Explorar muestras genéticas detectando grupos y analizandolos para identificar patrones de expresión relacionados
- Agrupar sonidos de ballenas para detectar patrones y analizar su forma de comunicación
- Análisis de imágenes para agruparlas y detectar outliers que pueden ser asociados, por ejemplo, a enfermedades

Retail y ventas

- Detectar grupos de clientes para caracterizarlos y definir estrategias que mejoren su experiencia de compra
- Detectar grupos de productos en base a volumen de ventas y frecuencia de compra para tomar decisiones enfocándose en lo más importante
- Agrupar tiendas para detectar patrones de compra que dependen de la localidad

Sociedad

- Análisis de grupos en redes sociales para caracterizar fenómenos sociales
- Detectar temáticas principales presentes en medios de comunicación en un periodo determinado
- Agrupar personas que responden una encuesta para detectar estructuras naturales e implícitas

Política

- Agrupar representantes en base a sus votaciones en proyectos de ley para detectar alianzas entre partidos
- Agrupar discursos políticos de presidentes y presidentas para detectar las temáticas principales y elementos diferenciadores

dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl