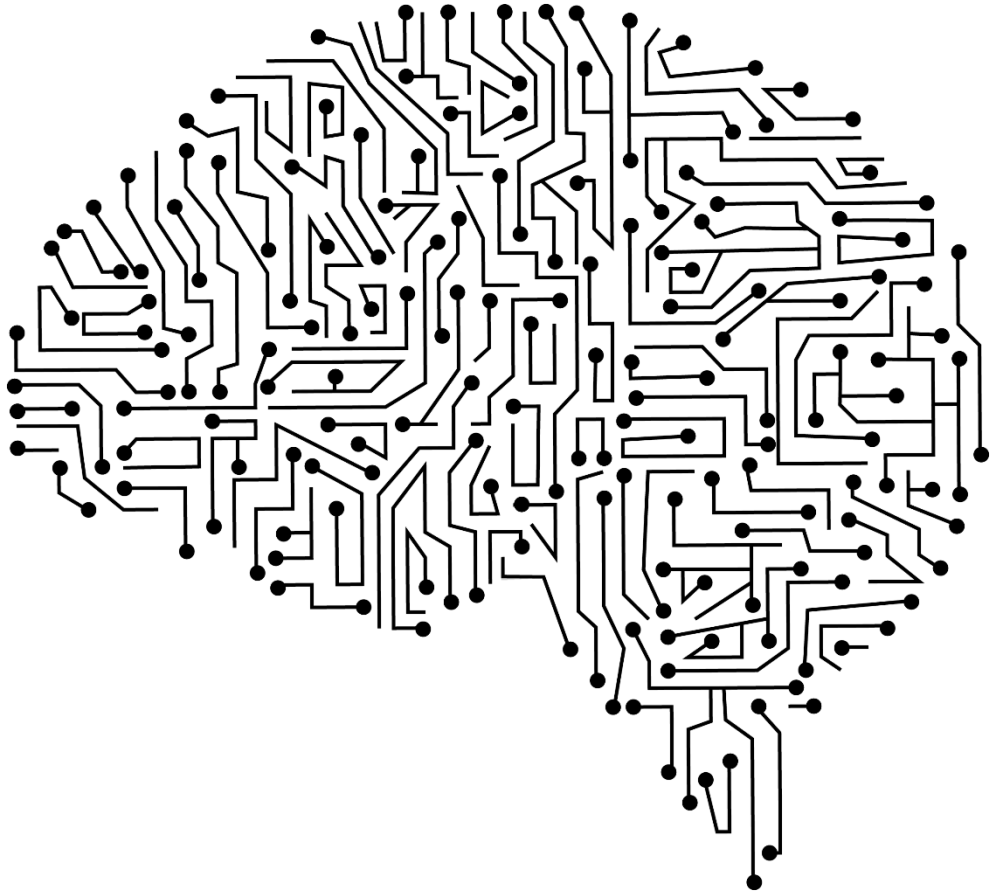
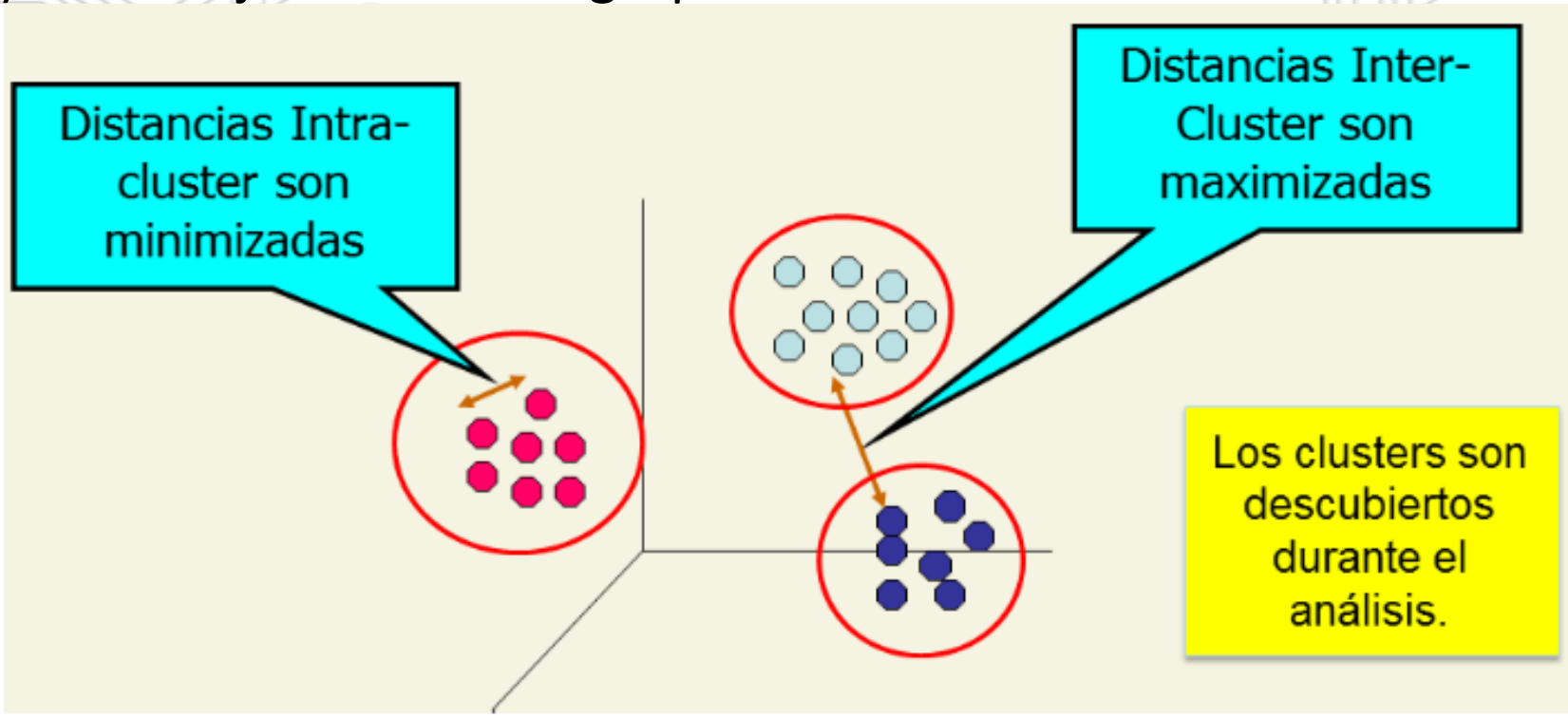
Introducción al análisis de clúster



Clustering

¿Que es el análisis de clúster?

Encontrar grupos de objetos de tal forma que los objetos en un grupo sean similares (o relacionados) unos a otros y distintos de (o poco relacionados) otros objetos en otros grupos



¿Qué es el análisis de clúster?

- El análisis de clúster o conglomerados tiene como objetivo agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos.

- Normalmente se agrupan las observaciones, aunque puede también aplicarse para agrupar variables.

- Estos métodos se conocen también con el nombre de métodos de clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión.

Métodos

- **Partición de los datos-métodos particionales:**
  - Algoritmo k-medias
  - **pam**
- **Métodos jerárquicos:**
  - Métodos aglomerativos
  - Métodos divisivos
- **Clúster basado en modelos**

¿Qué estudia?

El análisis de conglomerados estudia tres tipos de problemas:

- **Partición de los datos:** Disponemos de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:
  1. cada elemento pertenezca a uno y solo uno de los grupos.
  2. todo elemento quede clasificado.
  3. cada grupo sea internamente homogéneo.

  Por ejemplo: se dispone de una base de datos de compras de clientes y se desea hacer una tipología de estos clientes en función de sus pautas de consumo.

¿Qué estudia?

El análisis de conglomerados estudia tres tipos de problemas:

- **Construcción de jerarquías:** Deseamos estructurar los elementos de un conjunto de forma jerárquica por su similitud.

  Por ejemplo: tenemos una encuesta de atributos de distintas profesiones y queremos ordenarlas por similitud. Una clasificación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores.

  Este tipo de clasificación es muy frecuentes en biología, al clasificar animales, plantas etc. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos. Sin embargo, como veremos, la jerarquía construida permite obtener también una partición de los datos en grupos.

¿Qué estudia?

El análisis de conglomerados estudia tres tipos de problemas:

- **Clúster basado en modelos:** En este caso, se parte de la hipótesis de que los datos han sido generados de una mixtura de distribuciones.

- Entonces lo que hay que estimar son los parámetros desconocidos de esta mixtura usando el método de máxima verosimilitud.

- Las observaciones son asignadas al grupo que tiene mayor probabilidad de haberlas generado.

¿Qué necesitamos?

- Los métodos de partición utilizan la matriz de datos.
- Los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos.
- Los algoritmos basados en un modelo utilizan algoritmos de optimización para estimar los parámetros desconocidos.