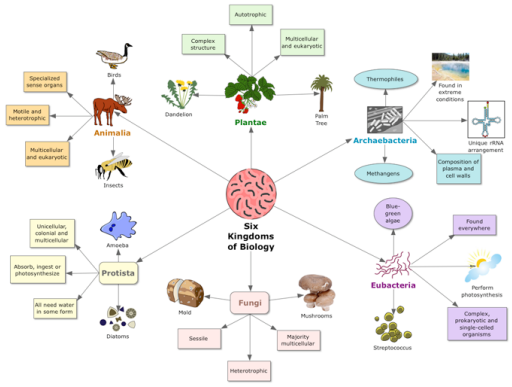
Clustering

Parte I: Introducción, Ejemplos, Caso de uso real

Bárbara Poblete

Inspiración histórica (taxonomía)



¿Qué es el clustering?

- Encontrar grupos de objetos especificando 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

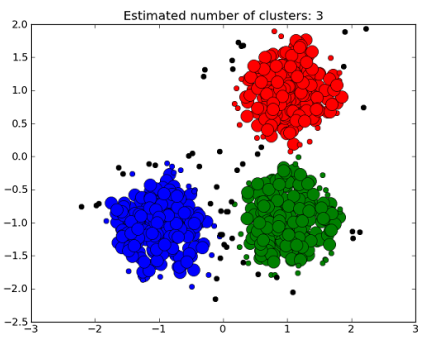
¿Cuándo y para qué usar clustering?

- 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

Busca capturar agrupaciones naturales en los datos



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 aleatorios 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

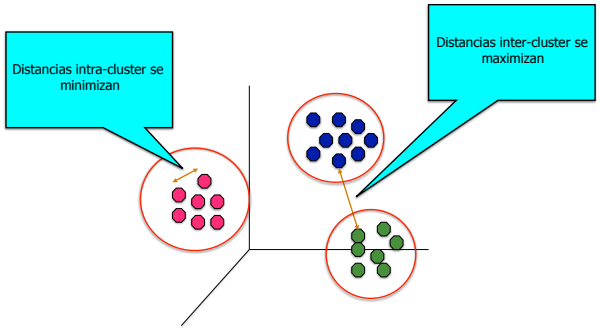
Aplicaciones

- WWW (e.j. clasificación de documentos en buscadores)
- Reconocimiento de patrones (e.j. agrupar series de tiempo)
- Procesamiento de imágenes
- Recomendación
- etc.

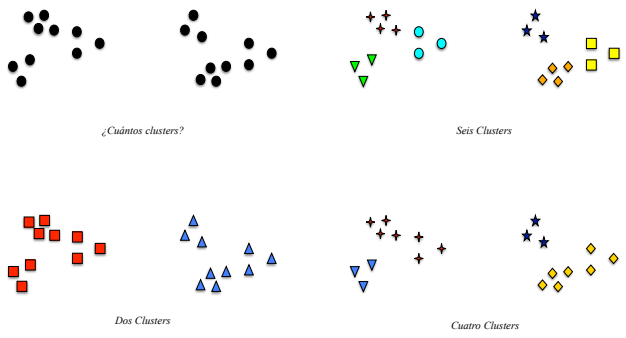
más aplicaciones

- Imagenes
- Web
- Películas
- Marketing

¿Qué es análisis de clusters?



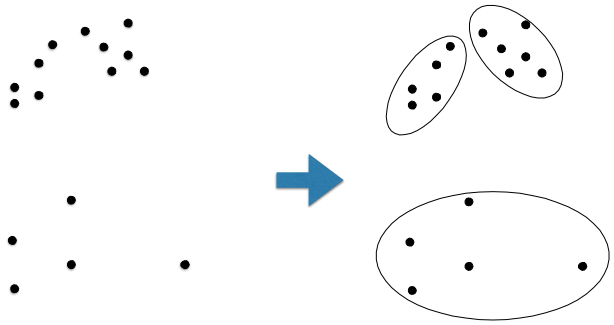
La noción de cluster puede ser ambigua



Tipos de clustering

- Un clustering es un conjunto de clusters
- Distinción importante entre conjuntos de clusters jerárquicos y particionales
- Clustering Particional
  - Divide los datos en subconjuntos sin traslape (clusters), tal que cada dato está en un solo subconjunto
- Clustering Jerárquico
  - Un conjunto de clusters anidados, organizados como un árbol

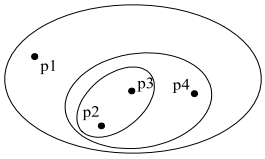
Clustering particional



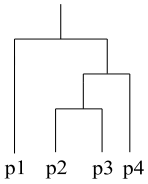
Puntos originales

Clustering Particional

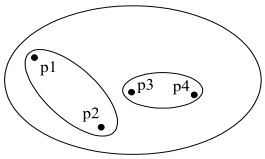
Clustering jerárquico



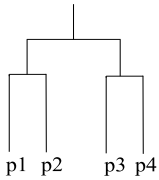
Clustering Jerárquico tradicional



Dendograma tradicional



Clustering Jerárquico no tradicional



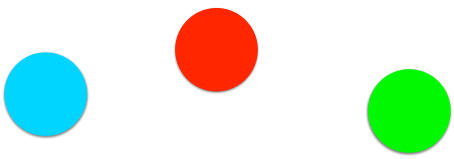
Dendograma no tradicional

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

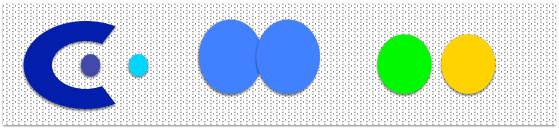
- Un cluster es un conjunto de objetos, tal que: un objeto dentro del cluster está más cerca (es más similar) **al centro de este cluster** que al centro de cualquier otro.
- El centro de un cluster puede ser el centroide, el promedio de todos los puntos en el cluster, o el medioide, el punto más “representativo” del cluster

Clusters contiguos (vecino más cercano o transitivo)

- Un cluster es un conjunto de puntos, tal que: un 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

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



Algoritmo de Clustering: K-means

- De los métodos más populares
- Particional
  - Cada cluster se asocia a un centroide
  - Cada punto se asigna al cluster cuyo centroide sea el más cercano
  - Parámetro, K = número de clusters

Algoritmo de Clustering: K-means

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Complejidad es O(n * K * I * d)
n puntos, K centros, I iteraciones, d dimensiones

Sum of Squared Error (SSE)

- Medida más común para evaluar clusters
- Por cada punto, error es la distancia al cluster más cercano
- x: punto en el cluster Ci, mi: centroide Ci
- Dados 2 clusters se escoge el que tiene menos error

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

Variante: Bisecting K-means

- Puede producir clustering jerárquico o particional

Algorithm 3 Bisecting K-means Algorithm.

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters