Clasificación y agrupamiento

Métodos para formar colecciones de objetos.

Contenidos

- Agrupamiento
  - Descripción
  - Ejercicio con K-means
- Clasificación

Agrupamiento (*clustering*)

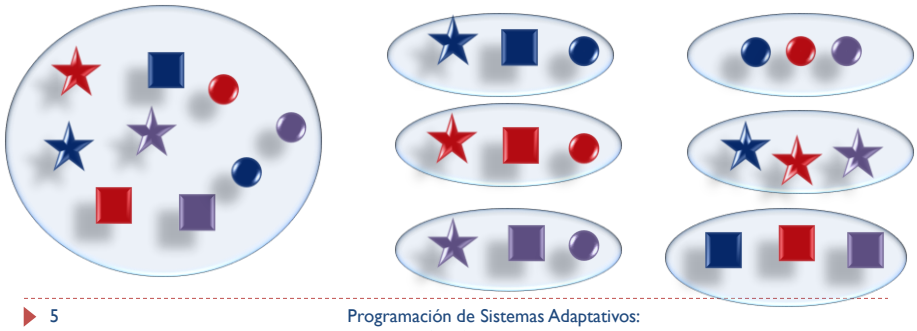
- Consiste en acomodar elementos en grupos.
  - Elementos en el mismo grupo deben ser más similares entre ellos
    - … y menos similares con respecto a otros grupos.
- Los grupos **no** se conocen de antemano.
- En los enfoques clásicos, la información se representa con vectores de características
  - ej. coordenadas en un plano cartesiano, vectores binarios, pesos reales, etc.

Datos

- Los datos pueden representar lo que Uds. quieran.
  - Películas
  - Páginas Web (documentos)
  - Superhéroes
  - Proteínas
  - Canciones
  - Personas
  - El límite es la creatividad.

Teorema del patito feo

- Necesitamos un criterio para agrupar
  - i.e., nunca tenemos grupos libres de sesgos.
- Existen diferentes métricas y enfoques.



Distancia y similitud

- Similitud = parecido entre un par de elementos
- Distancia = disimilitud entre un par de elementos
- Métricas comunes
  - Índice Jaccard → Conjuntos
  - Distancia euclidiana → Espacios geométricos
  - Coseno → Vectores de pesos

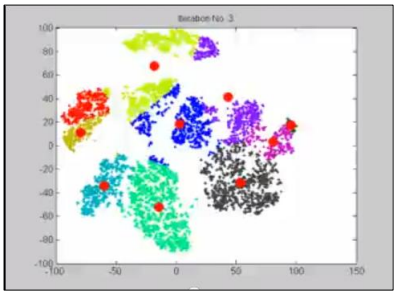
Tipos de clusters de acuerdo a contención de elementos

Algoritmo K-means

- Algoritmo clásico de agrupamiento
  - Particional
- Inspiración para muchos otros algoritmos
- Introducido en los 70's
- Español: K-medias

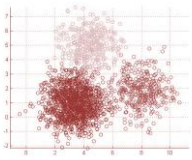
K-means

- http://www.youtube.com/watch?v=74rv4snLl70&feature=endscreen&NR=1



K-means: ¿Cómo funciona?

- Dado un conjunto de datos (vectores)
  - en N dimensiones
- Colocar **k** centroides.
  - al azar
- Asignar cada vector al centroide más cercano.
- Con los grupos formados, calcular nuevos centroides.
  - Que queden en el centro del grupo recién formado…
- Repetir esto hasta que el *error* sea menor a un umbral.



Recalculando centroides

- Obtener el promedio por cada dimensión del vector.
- Ejemplo (3D):

Ci = (5, 7, 10)
pa = (1, 2, 5)
pb = (3, 4, 6)
pc = (8, 12, 20)
pd = (4, 17, 5)

$$Ci_x = \frac{pa_x + pb_x + pc_x + pd_x}{\text{cant. puntos en } Ci} = \frac{1+3+8+4}{4} = \frac{16}{4} = 4$$

$$Ci_y = \frac{pa_y + pb_y + pc_y + pd_y}{\text{cant. puntos en } Ci} = \frac{2+4+12+17}{4} = \frac{35}{4} = 8.75$$

Recalculando centroides

Ci = (5, 7, 10)
pa = (1, 2, 5)
pb = (3, 4, 6)
pc = (8, 12, 20)
pd = (4, 17, 5)

$$Ci_z = \frac{pa_z + pb_z + pc_z + pd_z}{\text{cant. puntos en } Ci} = \frac{5+6+20+5}{4} = \frac{36}{4} = 9$$

Ci = ~~(5, 7, 10)~~

Ci = (4, 8.75, 9)

Ejemplo

Puntos (datos)
P1=(1,7)
P2=(2,6)
P3=(4,8)
P4=(10,19)
P5=(11,15)

Centroides
C1=(3,5)
C2=(12,13)

Cálculo de distancia a cada centroide (para asignar al más cercano)

Punto	Distancia euclidiana a C1 (3, 5)	Distancia C2 (12, 13)
(1,7)	2.8	12.5
(2,6)	1.4	12.2
(4,8)	3.2	9.4
(10,19)	15.7	6.3
(11,15)	12.8	2.2

Ejemplo de cálculo distancia euclidiana

$$dist(P1, C1) = \sqrt{(1-3)^2 + (7-5)^2}$$
$$dist(P1, C1) = \sqrt{4+4} = \sqrt{8}$$
$$dist(P1, C1) = 2.8$$

Asignación de puntos a grupos

Grupo 1 (C1)	Grupo 2 (C2)
(1,7)	(10, 19)
(2,6)	(11, 15)
(4,8)	

Centroides recalculados

C1 (x) = (1+2+4)/3 = 7/3 = 2.3
C1 (y) = (7+6+8)/3 = 21/3 = 7
C1 = (2.3, 7)

C2 (x) = (10+11)/2 = 21/2 = 10.5
C2 (y) = (19+15)/2 = 34/2 = 17
C2 = (10.5, 17)

Ejemplo

Puntos (datos)
P1=(1,7)
P2=(2,6)
P3=(4,8)
P4=(10,19)
P5=(11,15)

Centroides
C1=(3,5)
C2=(12,13)

Ejemplo de cálculo distancia euclidiana

$$dist(P1, C1) = \sqrt{(1-3)^2 + (7-5)^2}$$
$$dist(P1, C1) = \sqrt{4+4} = \sqrt{8}$$
$$dist(P1, C1) = 2.8$$

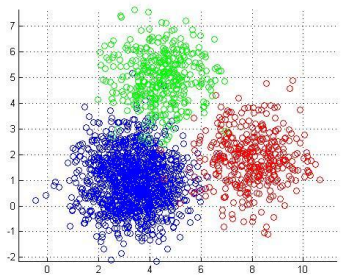
Centroides recalculados

C1 (x) = (1+2+4)/3 = 7/3 = 2.3
C1 (y) = (7+6+8)/3 = 21/3 = 7
C1 = (2.3, 7)

C2 (x) = (10+11)/2 = 21/2 = 10.5
C2 (y) = (19+15)/2 = 34/2 = 17
C2 = (10.5, 17)

K-means: ejercicio

- Puntos:
  - p1= (1,1),
  - p2=(2,4),
  - p3=(3,2),
  - p4=(7,2),
  - p5=(8,3)
- Centroides
  - k=2
  - c1=(3,5)
  - c2= (9,1)



K-means: ejercicio

- Utiliza la distancia euclidiana como métrica de *disimilitud*.
- Realiza dos iteraciones del algoritmo.

K-means: OJO

- No siempre es fácil colocar los centroides iniciales.
  - También es difícil saber cuántos poner.
- Susceptible a intrusos (*outliers*)

Otros tipos de clustering

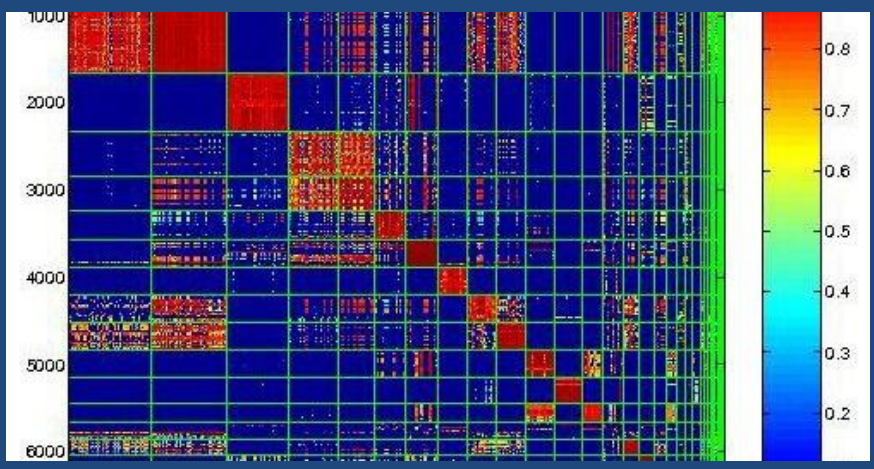
- Basado en grafos
- Basado en densidad
- Kernels
- Co-clustering

¿Cómo evaluar un agrupamiento?

- Visualmente
  - Matrices de similitud
- Precisión y recuerdo (*precision and recall*)
  - Correctitud y completez
  - Medidas integradoras: F

Matriz de similitud

Fuente: http://blogoutthemaps.blogspot.com/2011/01/similarity-matrix.html

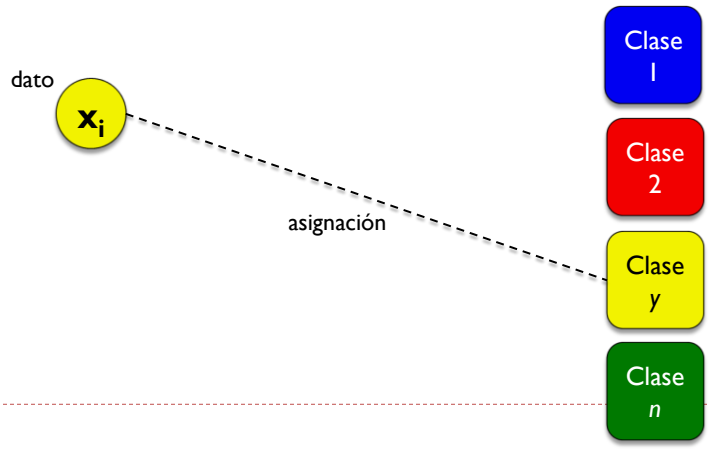


Clasificación

- Igual que agrupamiento, pero los grupos (llamados clases o categorías) ya se encuentran pre-establecidos.

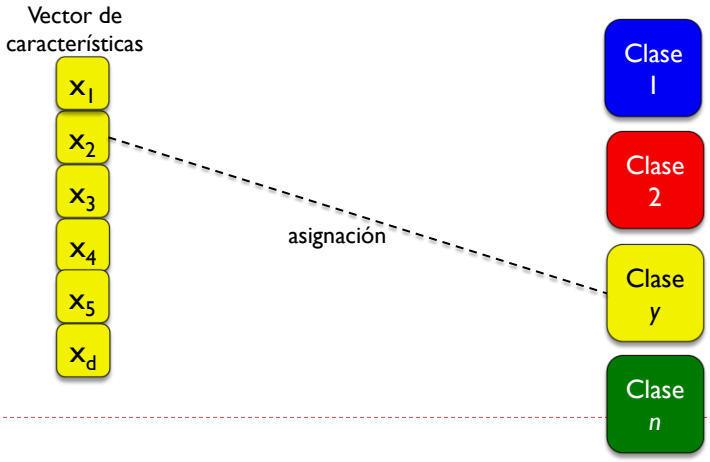
Clasificación

- También conocida como categorización.

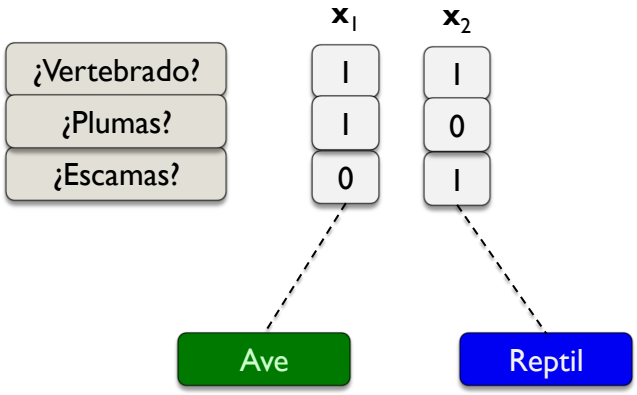


Clasificación

- También conocida como categorización.



Ejemplo con vectores binarios



Clasificación: Análisis de sentimiento

- Una aplicación reciente es el análisis de sentimiento para documentos (opiniones).
  - Tres clases: Positivo, negativo, neutro.
  - Ejemplo **positivo**: "La película estuvo muy divertida."
  - Ejemplo **negativo**: "El nuevo dispositivo es poco funcional."
  - Ejemplo **neutro**: "El motor viene con dos sensores."

Resumen

- Agrupamiento = colocar datos en grupos de acuerdo a su parecido.
- Tipos de grupos
  - Sin traslape
  - Con traslape
- K-medias (k-means)
  - Asignar el dato al centroide más cercano. Al final, centroides.

Resumen

- Agrupamiento = colocar datos en grupos de acuerdo a su parecido.
- Tipos de grupos
  - Sin traslape
  - Con traslape

- K-medias (k-means)
  - Asignar el dato al centroide más cercano. Al final, recalcular los centroides.

Referencias

- Tan et al. Introduction to Data Mining. Addison Wesley, EUA, 2006.
- A Tutorial on Clustering Algorithms. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html