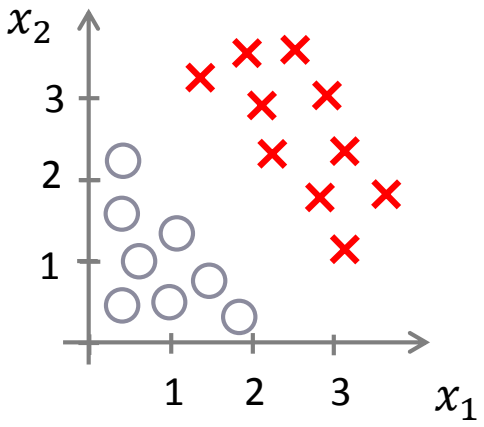
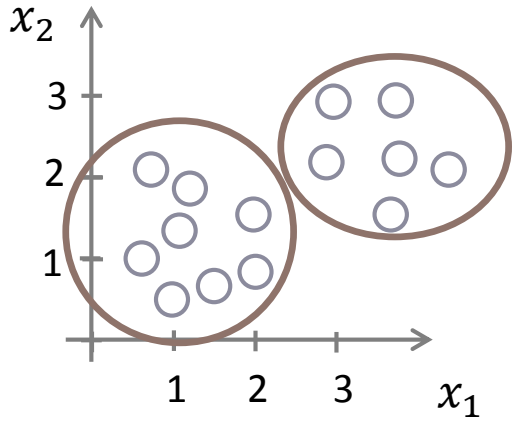
CLUSTERING. APRENDIZAJE NO SUPERVISADO

Miguel Pagola Barrio

Aprendizaje no supervisado



Aprendizaje supervisado: los ejemplos están previamente clasificados



Aprendizaje no supervisado: los ejemplos no están etiquetados

Aprendizaje no supervisado

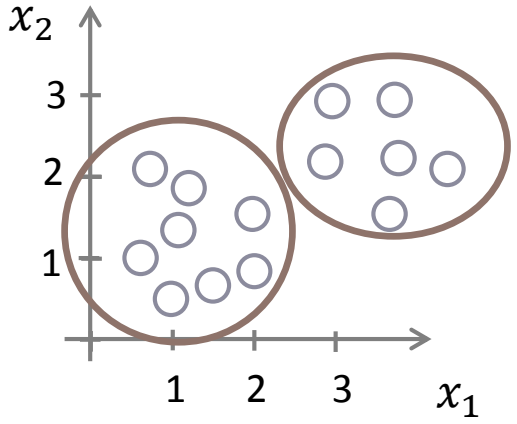
- Problema: tenemos muchos datos y queremos sacar información de ellos.
- Solución: reducirlos!
- **Clustering:** agrupar los ejemplos y quedarnos con los más característicos.
- **Reducción de la dimensionalidad:** reducir el número de dimensiones.

Clustering

- Dados una serie de ejemplos:
- Dividirlos en subconjuntos de ejemplos que son similares entre si.
- ¿Cómo medimos esa similitud?
- La noción de cluster, o grupo, puede ser:
  - un grupo de elementos entre los cuales hay poca distancia,
  - areas del espacio con mucha densidad de elementos
  - un grupo de elementos que sigan una distribución de probabilidad particular.

Clustering

- La idea principal:
  - Los elementos dentro del cluster tienen que ser similares, de tal forma que a nivel global la suma **de las "distancias" intra cluste**r debe ser **pequeña**.
  - Los clústeres tienen que ser diferenciados, es decir **la "distancia" inter-cluster** debe ser **grande**.



  - Por lo tanto es un problema de optimización multiobjetivo.

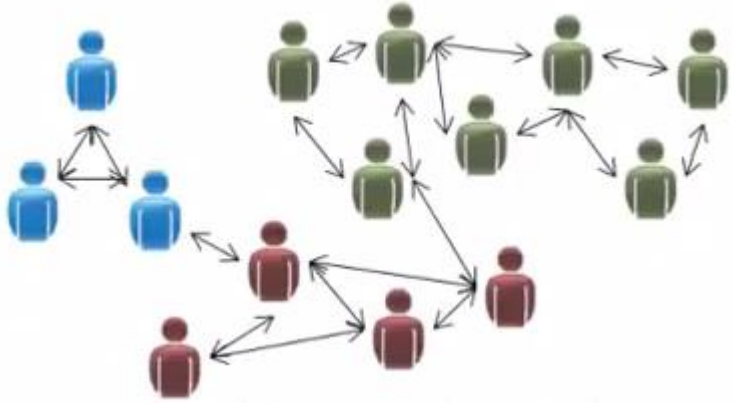
Aprendizaje no supervisado

- Clustering
  - K-means (K-medias)
  - Mixture Models, Expectation-Maximization
  - Hierarchical clustering
- Reducción de dimensionalidad
  - Análisis de los componentes principales. PCA

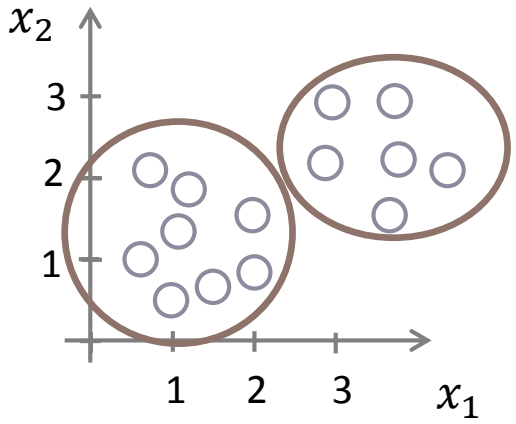
Aplicaciones del clustering

Segmentación de mercado

Análisis de redes sociales



Análisis de imágenes



Análisis de datos, detección de anomalías, análisis de secuencias, etc..

Aplicaciones del clustering

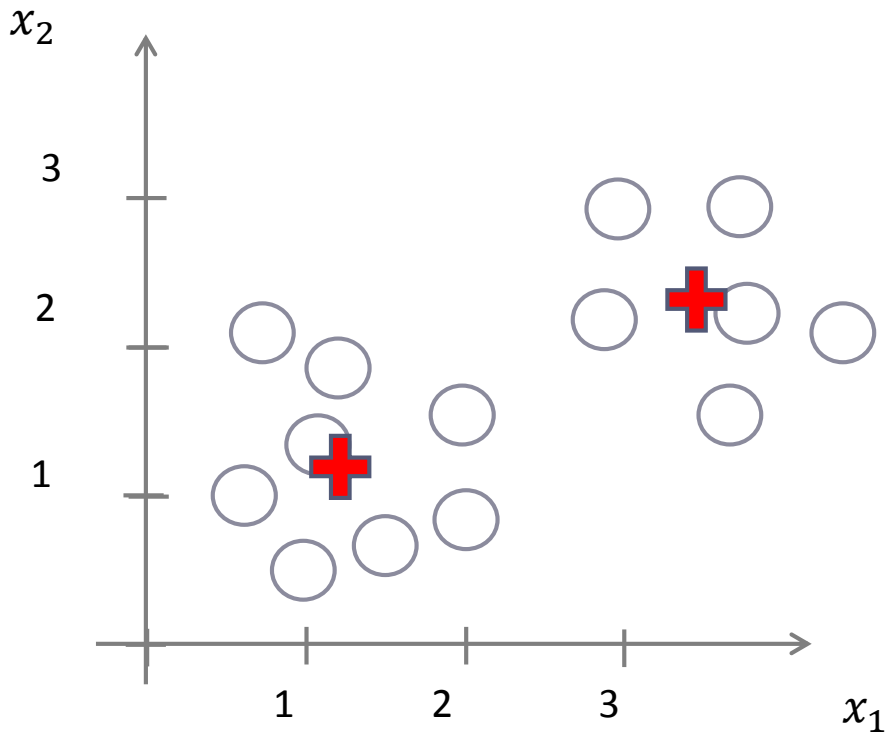
- Visualización de la estructura de los datos
- Dividir el problema original en subproblemas

- Obtención de nuevas características (Bag of Words)

K-means

- Clustering basado en centroides. Los clusteres están representados por un element central, que no tiene porque pertenecer al dataset.
- Buscamos un número fijo de clusteres.
- El problema de optimización se convierte en encontrar los k centroides de los k clusteres y asignar los elementos al centroide más cercano, de tal forma que la suma de las distancias a los centroides sea mínima.

K-means



Función objetivo del algoritmo K-means

- Data set X, con $x^{(i)}$ cada elemento
- $c^{(i)}$ = índice del clúster (1, 2, 3, …, K) en el que está asignado $x^{(i)}$
- μ_k = centroide del clúster k (tiene las mismas dimensiones que los ejemplos)
- $\mu_{c^{(i)}}$ = centroide del clúster al cual el ejemplo $x^{(i)}$ ha sido asignado
- Función objetivo:

$$J\left(c^{(1)}, \ldots, c^{(m)}, \mu_1, \mu_2, .., \mu_k\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

K-means. K-medias

- Inicializar de forma aleatoria los centros de los K clústeres (centroides), $\mu_1, \mu_2, .., \mu_k$.

Repetir hasta converger{

For i = 1..m

$c^{(i)} :=$ índice el centroide más cercano a $x^{(i)}$

For k = 1..K

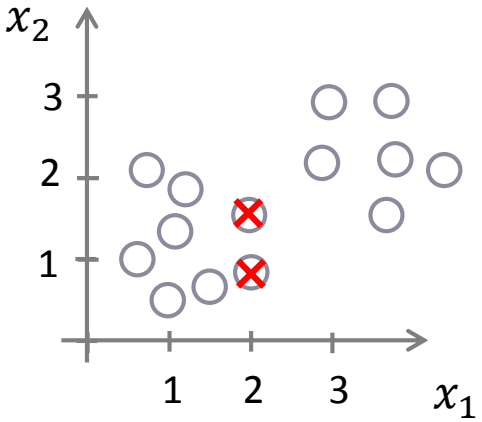
μ_k := media de los ejemplos asignados al clúster k

}

http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Inicialización aleatoria

- Como cualquier algoritmo de optimización, la solución del algoritmo k-means depende de las condiciones iniciales.
- Tres formas inicialización aleatoria:
  1. Seleccionar de forma aleatoria k elementos para que sean los centroides iniciales.
  2. Crear los centroides de forma aleatoria
  3. Generar $c^{(1)}, \dots, c^{(m)}$ de forma aleatoria



- TODOS tienen el peligro para caer en mínimos locales.

Inicialización aleatoria

For i = 1:100{

Inicializar aleatoriamente K-means

Ejecutar K-means y obtener $c^{(1)}, \dots, c^{(m)}, \mu_1, \mu_2, .., \mu_k$

Calcular $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \mu_2, .., \mu_k)$

}

Seleccionar la partición de menor J.

¿Cuantos clústeres hay en el dataset?

- Evaluación interna, evaluar la solución del algoritmo de clustering con los propios datos.
- "Cluster validity", Davies–Bouldin index
  - Queremos que los clústeres sean compactos y estén separados entre sí
  - Compacidad del cluster j: $C_j = \frac{1}{|C_j|} \sum_{x^{(i)} \in C_j} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$
  - Separación entre los clústeres i,j: $S_{i,j} = \left\| \mu_i - \mu_j \right\|^2$
  - $R_i = \max_{i,j \; i \neq j} \left\{ \frac{C_i + C_j}{S_{i,j}} \right\}$
  - $BD = \frac{1}{K} \sum_i^K R_i$ El K óptimo será el de menor valor de DB

Cluster validity

- **Índice de Dunn**. Trata de identificar clústeres muy densos y completamente separados. Es el ratio entre la mínima distancia inter-cluster y la máxima distancia intra-cluster.

$$D = \frac{\min\limits_{1<i<j\leq k} \left\| \mu_i - \mu_j \right\|}{\max\limits_{1\leq i\leq k, 1\leq i\leq m,} \left\| x_j^{(i)} - x_j^{(i)} \right\|}$$

- **El coeficiente de silueta**
  - **a**: La distancia media entre un ejemplo y el resto de elementos del mismo cluster.
  - **b**: La distancia media entre un ejemplo y el resto de elementos en cluster más cercano.
- El coeficiente de silueta es la media del valor s para todos los ejemplos del dataset.

$$s = \frac{b - a}{\max(a, b)}$$

¿Cómo sé si la partición que he obtenido es correcta?

- Evaluación externa, cuando tenemos información sobre las clases de los ejemplos.
- En la práctica imposible (sólo para comparar algoritmos)
- Rand índex:
  - Dadas dos particiones X= $\{c^{(1)}, \dots, c^{(m)}\}$ e $Y = \{c^{(1)}, \dots, c^{(m)}\}$ de los mismos datos:
    - a = número de pares de elementos que están en el mismo clúster en X y que están en el mismo clúster en Y
    - b = número de pares de elementos que están en diferentes clústeres en X y que están en diferentes clústeres en Y
    - c = número de pares de elementos que están en el mismo clúster en X y que están en diferentes clústeres en Y
    - d = número de pares de elementos que están en diferentes clústeres en X y que están en el mismo clúster en Y

$$RI = \frac{a + b}{a + b + c + d}$$

K-means funciona si:

- Los clústeres son esféricos.
- Los clústeres están separados (no están muy solapados).
- Los clústeres tienen volumen similar.
- Los clústeres tienen un número de elementos parecidos.