Sistemas Inteligentes

Escuela Técnica Superior de Informática

Universitat Politècnica de València

Tema B2T4:
Aprendizaje no supervisado: algoritmo k-medias.

Índice

Índice

Clustering

- Consiste en definir un agrupamiento sobre objetos no etiquetados
- Los objetos dentro de un grupo están fuertemente relacionados entre si
- Los objetos en grupos diferentes son muy distintos
- Ejemplo: clustering de vídeos para mostrar relacionados
  - Agrupar vídeos que son *similares* en temática
  - Idealmente los vídeos de un mismo grupo deberían estar relacionados
  - Vídeos en distintos grupos deberían tratar sobre diferentes temáticas
- Dos tipos de clustering: ***Particional*** y Jerárquico

Índice

Clustering particional

- Se dispone de N datos a particionar en C grupos (o clústers)

- Tenemos una ***función criterio*** J para evaluar una partición

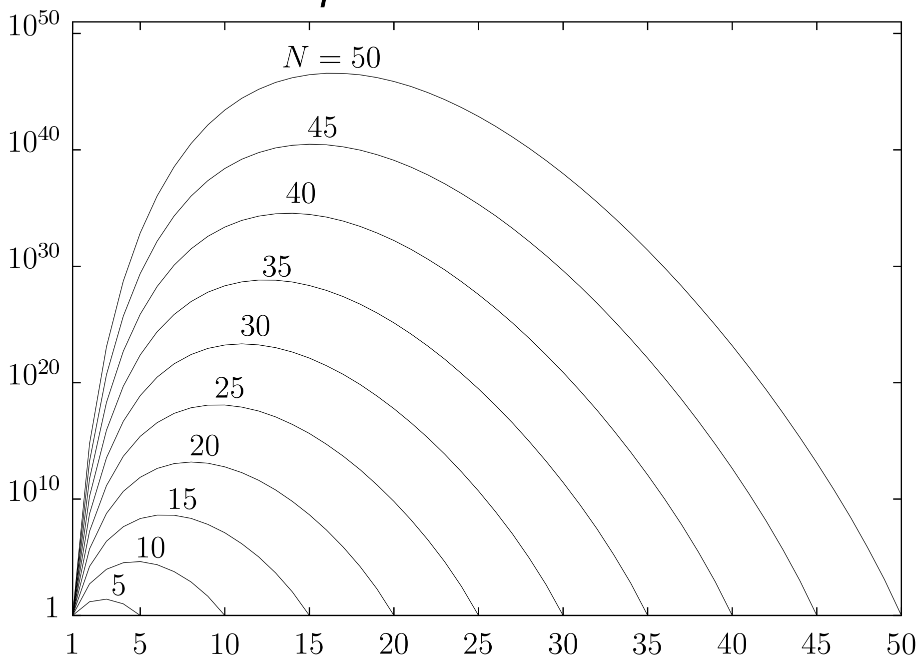
$$\Pi^* = \underset{\Pi=\{X_1,\ldots,X_C\}}{\arg\min} J(\Pi)$$

- Número de particiones es muy elevado

- No es factible evaluar todas las posibles particiones

- Solución subóptima mediante algoritmos aproximados: ***C-medias***

Número de particiones en función de C para varios N



Clustering particional: Función criterio "suma de errores cuadráticos" (SEC)

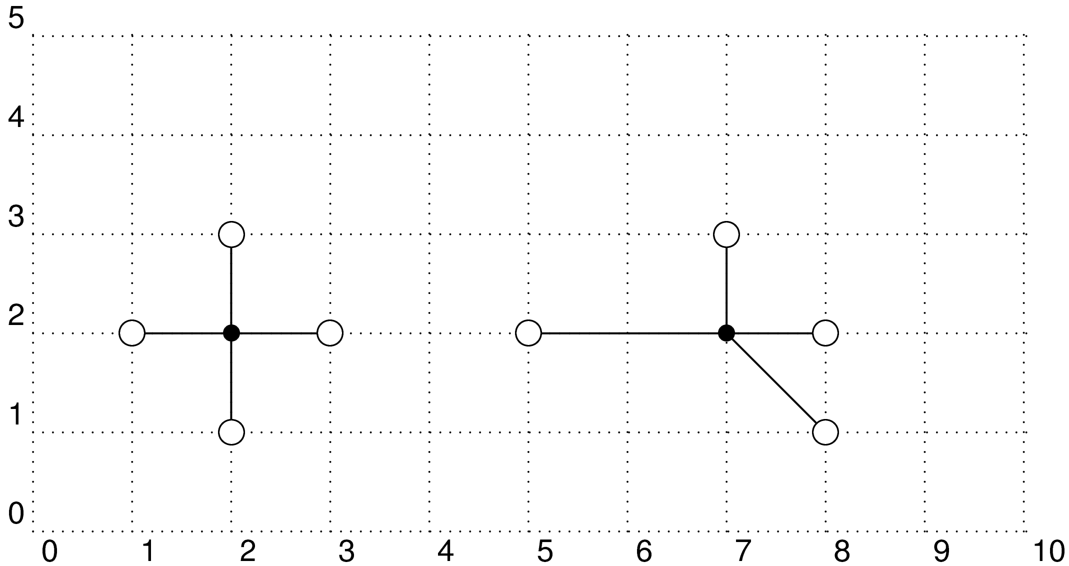
La SEC de una partición de N datos en C clusters, $\Pi = \{X_1, \ldots, X_C\}$, es:

$$J(\Pi) = J(X_1, \ldots, X_C) = \sum_c J_c\,, \quad J_c = \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{m}_c\|^2, \quad \boldsymbol{m}_c = \frac{1}{|X_c|} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$$

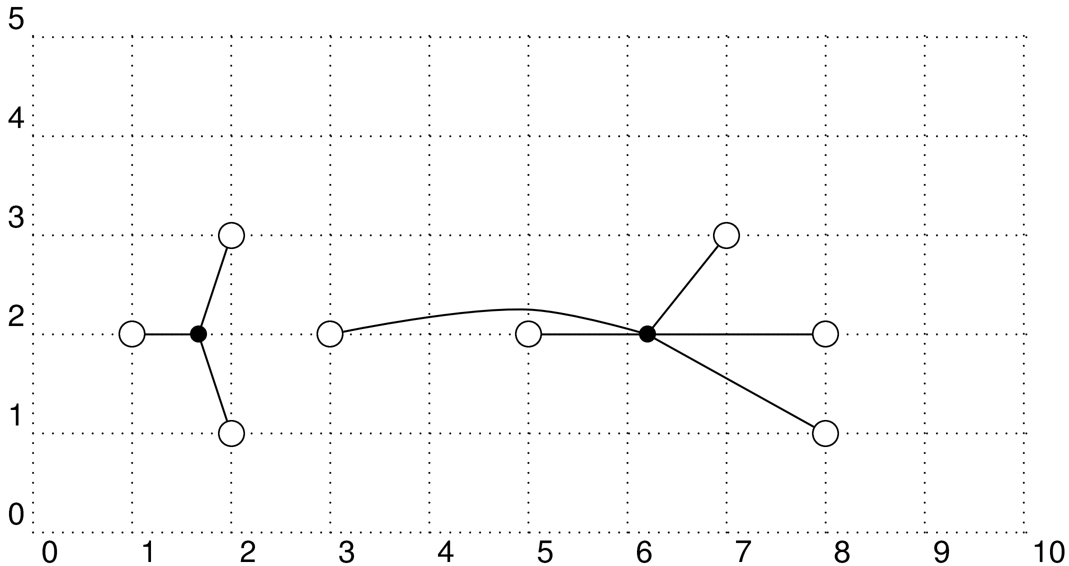
Interpretación:

- Para cada clúster, calculamos la distancia al cuadrado (error cuadrático) de cada dato $\boldsymbol{x} \in X_c$ a su *media*, $\boldsymbol{m}_c$
- Sumamos las distancias al cuadrado de cada dato a su media
- El objetivo es minimizar dicha suma

Ejemplo de clustering particional



SEC$=12$

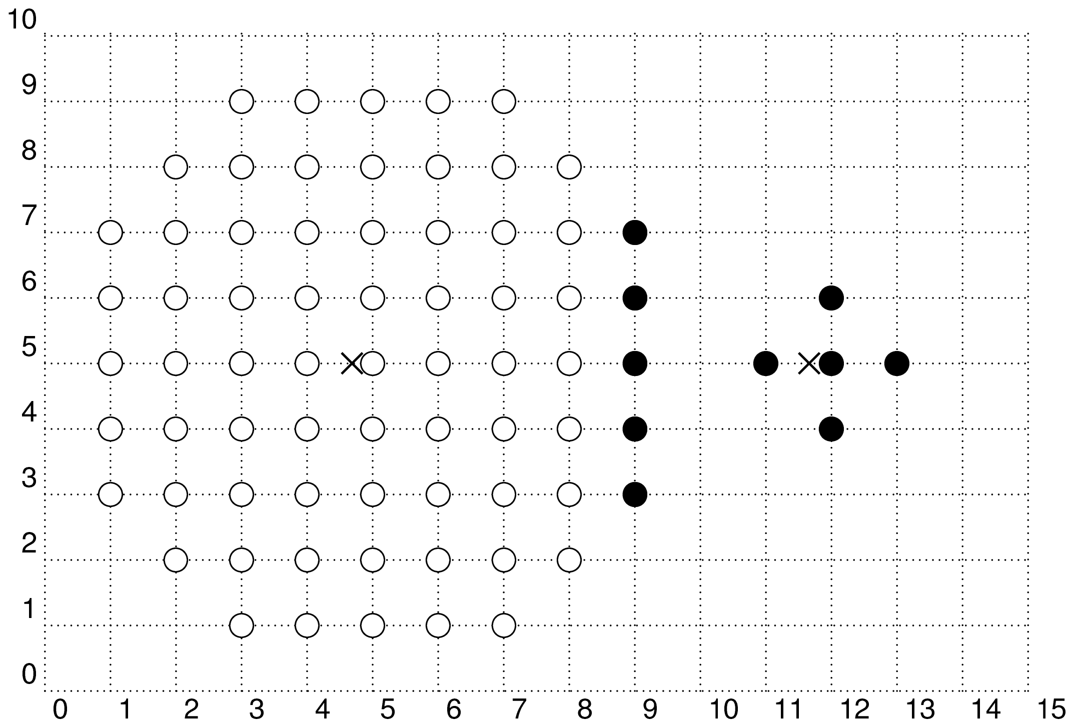


SEC>12

Funcionamiento práctico del criterio SEC

- Sólo es apropiado si los datos forman ***clusters hiperesféricos de tamaño similar.***

- Si los clusters tienen tamaños muy distintos, la agrupación *natural* no coincide con la SEC



Índice

Algoritmo C-medias

- Empieza de una partición inicial $\Pi_0 = \{X_1, \ldots, X_C\}$ que se mejora iterativamente
- *Mejora* entendida como minimización de la SEC
- Cada dato $\boldsymbol{x} \in X_i$ se prueba a cambiar a un clúster X_j
- Si la SEC con $\boldsymbol{x} \in X_j$ es menor que con $\boldsymbol{x} \in X_i$, se realiza el cambio
- Este cambio o transferencia conlleva actualización de X_i, X_j, $\boldsymbol{m}_i$, $\boldsymbol{m}_j$, J_i, J_j y J
- El algoritmo para cuando no hay transferencias al probar con todos los datos
- La versión de Duda & Hart de este algoritmo obtiene un *mínimo local*

Cálculo incremental de la SEC al transferir x del cluster X_i al X_j

$$X_i' = X_i - \{\boldsymbol{x}\} \qquad\qquad X_j' = X_j + \{\boldsymbol{x}\}$$

$$\boldsymbol{m}_i' = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} \qquad\qquad \boldsymbol{m}_j' = \boldsymbol{m}_j + \frac{\boldsymbol{x} - \boldsymbol{m}_j}{n_j + 1}$$

$$J_i' = J_i - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2 \qquad\qquad J_j' = J_j + \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2$$

$$\triangle J = \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

La transferencia será provechosa si el incremento de SEC es negativo; es decir:

$$\frac{n_j}{n_j + 1}\|\boldsymbol{x} - \boldsymbol{m}_j\|^2 < \frac{n_i}{n_i - 1}\|\boldsymbol{x} - \boldsymbol{m}_i\|^2 \qquad (1)$$

Estas ecuaciones permiten minimizar la SEC mediante refinamientos sucesivos a partir una partición inicial dada.

Optimización de la SEC: algoritmo C-medias

Algorithm C*-means* (versión “correcta” [Duda & Hart])

Input: X; C; $\Pi = \{X_1, \ldots, X_C\}$;

Output: $\Pi^* = \{X_1, \ldots, X_C\}$; $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_C$; J

for $c = 1$ **to** C **do** $\boldsymbol{m}_c = \frac{1}{n_c} \sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$ **endfor**

repeat

  $transfers =$ false

  forall $\boldsymbol{x} \in X$ (let $i : \boldsymbol{x} \in X_i$) **do**

    if $n_i > 1$ **then**

$$j^* = \arg\min_{j \neq i} \frac{n_j}{n_j + 1} \|\boldsymbol{x} - \boldsymbol{m}_j\|^2$$

$$\triangle J = \frac{n_{j^*}}{n_{j^*} + 1} \|\boldsymbol{x} - \boldsymbol{m}_{j^*}\|^2 - \frac{n_i}{n_i - 1} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2$$

      if $\triangle J < 0$ **then**

        $transfers =$ true

$$\boldsymbol{m}_i = \boldsymbol{m}_i - \frac{\boldsymbol{x} - \boldsymbol{m}_i}{n_i - 1} \qquad \boldsymbol{m}_{j^*} = \boldsymbol{m}_{j^*} + \frac{\boldsymbol{x} - \boldsymbol{m}_{j^*}}{n_{j^*} + 1}$$

$$X_i = X_i - \{\boldsymbol{x}\} \qquad X_{j^*} = X_{j^*} + \{\boldsymbol{x}\}$$

$$J = J + \triangle J$$

      endif

    endif

  endforall

until $\neg transfers$ // Coste por iteración: $O(N \cdot C \cdot D)$, $N = |X|$, D = dimensión

Ejercicio: algoritmo C-medias

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$
x_{n1}	1	4	4	8	8
x_{n2}	7	2	6	2	6

$$\Pi = \{X_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}, X_2 = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}\}$$

Solución ejercicio: Inicialización

$$\boldsymbol{m}_1 = \frac{1}{3}(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3) = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\boldsymbol{m}_2 = \frac{1}{2}(\boldsymbol{x}_4 + \boldsymbol{x}_5) = \begin{pmatrix} 8 \\ 4 \end{pmatrix}$$

$$\begin{aligned} J_1 &= \|\boldsymbol{x}_1 - \boldsymbol{m}_1\|^2 + \|\boldsymbol{x}_2 - \boldsymbol{m}_1\|^2 + \|\boldsymbol{x}_3 - \boldsymbol{m}_1\|^2 \\ &= 8 + 10 + 2 = 20 \\ J_2 &= \|\boldsymbol{x}_4 - \boldsymbol{m}_2\|^2 + \|\boldsymbol{x}_5 - \boldsymbol{m}_2\|^2 = 4 + 4 = 8 \\ J &= J_1 + J_2 = 28 \end{aligned}$$

Solución ejercicio: Iteración 1

¿$\boldsymbol{x}_1$ de X_1 a X_2? : $\Delta J = \frac{2}{3} \cdot 58 - \frac{3}{2} \cdot 8 = \frac{80}{3} > 0 \;\Rightarrow$ NO

¿$\boldsymbol{x}_2$ de X_1 a X_2? : $\Delta J = \frac{2}{3} \cdot 20 - \frac{3}{2} \cdot 10 = -\frac{5}{3} < 0 \;\Rightarrow$ SÍ

$$\boldsymbol{m}_1 = \boldsymbol{m}_1 - \frac{\boldsymbol{x}_2 - \boldsymbol{m}_1}{n_1 - 1} = \begin{pmatrix} 5/2 \\ 13/2 \end{pmatrix} = \begin{pmatrix} 2{,}5 \\ 6{,}5 \end{pmatrix}$$

$$\boldsymbol{m}_2 = \boldsymbol{m}_2 + \frac{\boldsymbol{x}_2 - \boldsymbol{m}_2}{n_2 + 1} = \begin{pmatrix} 20/3 \\ 10/3 \end{pmatrix} = \begin{pmatrix} 6{,}67 \\ 3{,}33 \end{pmatrix}$$

$$J = J + \Delta J = \frac{79}{3} = 26{,}33$$

¿$\boldsymbol{x}_3$ de X_1 a X_2? : $\Delta J = \frac{3}{4} \cdot \frac{128}{9} - \frac{2}{1} \cdot \frac{10}{4} = \frac{17}{3} = 5{,}67 > 0 \;\Rightarrow$ NO

¿$\boldsymbol{x}_4$ de X_2 a X_1? : $\Delta J = \frac{2}{3} \cdot \frac{151}{2} - \frac{3}{2} \cdot \frac{32}{9} = \frac{805}{16} = 50{,}31 > 0 \;\Rightarrow$ NO

¿$\boldsymbol{x}_5$ de X_2 a X_1? : $\Delta J = \frac{2}{3} \cdot \frac{61}{2} - \frac{3}{2} \cdot \frac{80}{9} = 7 > 0 \;\Rightarrow$ NO

Solución ejercicio: Iteración 2

¿$\boldsymbol{x}_1$ de X_1 a X_2? : $\Delta J = \frac{3}{4} \cdot \frac{410}{9} - \frac{2}{1} \cdot \frac{5}{2} = \frac{175}{6} = 29{,}17 > 0 \;\Rightarrow$ NO

¿$\boldsymbol{x}_2$ de X_2 a X_1? : $\Delta J = \frac{2}{3} \cdot \frac{45}{2} - \frac{3}{2} \cdot \frac{80}{9} = \frac{5}{3} = 1{,}67 > 0 \;\Rightarrow$ NO

Partición optimizada: $\Pi^* = \{X_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_3\}, X_2 = \{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_5\}\}$

Optimización de la SEC: otra versión de C-medias

Algorithm C*-means* (versión “popular”)

Input: X; C; $\Pi = \{X_1, \ldots, X_C\}$;

Output: $\Pi^* = \{X_1, \ldots, X_C\}$; $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_C$

repeat

$\quad transfers =$ false

$\quad$ **for** $c = 1$ **to** C **do** $\boldsymbol{m}_c = \dfrac{1}{n_c} \displaystyle\sum_{\boldsymbol{x} \in X_c} \boldsymbol{x}$ **endfor**

$\quad$ **forall** $\boldsymbol{x} \in X$ (let $i : \boldsymbol{x} \in X_i$) **do**

$\quad\quad$ **if** $n_i > 1$ **then**

$\quad\quad\quad j^* = \displaystyle\arg\min_{1 \leq j \leq C} \; d(\boldsymbol{x}, \boldsymbol{m}_j)$

$\quad\quad\quad$ **if** $j^* \neq i$ **then**

$\quad\quad\quad\quad transfers =$ true

$\quad\quad\quad\quad X_i = X_i - \{\boldsymbol{x}\}$; $X_{j^*} = X_{j^*} + \{\boldsymbol{x}\}$

$\quad\quad\quad$ **endif**

$\quad\quad$ **endif**

$\quad$ **endforall**

until $\neg transfers$

// Coste por iteración: $O(N \cdot C \cdot D)$, $N = |X|$, D = coste de $d(\cdot, \cdot)$

Optimalidad de los algoritmos C-medias

- Ninguna de las versiones del algoritmo C-*medias* garantiza la obtención de un mínimo global de la SEC
- La versión de Duda & Hart obtine un *mínimo local*
- La versión "popular" no garantiza la minimización local en algunos casos

  Ejemplo:

  $X = \{1, 3, 4.5\} \subset \mathbb{R}$; $\quad \Pi^0 = \{\{1, 3\}, \{4.5\}\}$; $\quad J^0 = 2.0$

  C-medias "popular": $\quad \Pi^\star = \Pi^0$; $\quad J^\star = J^0 = 2.0$

  C-medias Duda & Hart: $\quad \Pi^\star = \Pi^1 = \{\{1\}, \{3, 4.5\}\}$; $\quad J^\star = J^1 = 1.125$