Cuaderno de trabajo

Clustering: algoritmo C-medias

Albert Sanchis

Departamento de Sistemas
Informáticos y Computación

Objetivos formativos

- Aplicar el algoritmo C-medias de Duda y Hart

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Algoritmo C-medias de Duda y Hart [1]

Algorithm *C-means*

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

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

Algoritmo C-medias de Duda y Hart

- ***Entrada:*** una partición inicial, $\Pi = \{X_1, \ldots, X_C\}$
- ***Salida:*** una partición optimizada, $\Pi^* = \{X_1, \ldots, X_C\}$
- ***Método:***

  Calcular medias y J

  repetir

    para todo dato $\boldsymbol{x}$

      Sea i el clúster en el que se encuentra $\boldsymbol{x}$

      Hallar un $j^* \neq i$ que minimice $\triangle J$ al transferir $\boldsymbol{x}$ de i a j^*

      Si $\triangle J < 0$, transferir $\boldsymbol{x}$ de i a j^* y actualizar medias y J

  hasta no encontrar transferencias provechosas

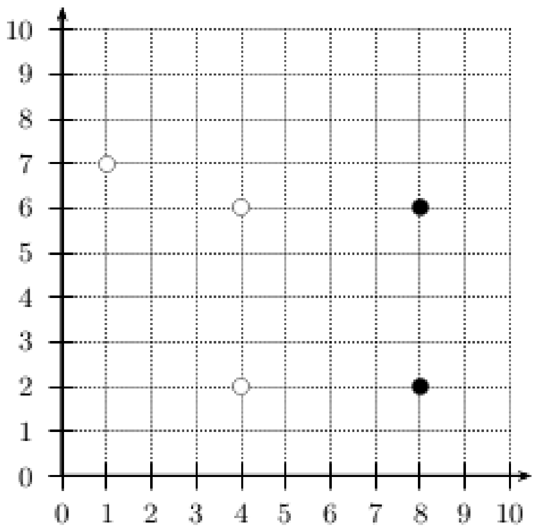
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

- ***Cuestión 1***: Dados los siguientes 5 vectores bidimensionales:

$$\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 7 \end{pmatrix} \quad \boldsymbol{x}_2 = \begin{pmatrix} 4 \\ 2 \end{pmatrix} \quad \boldsymbol{x}_3 = \begin{pmatrix} 4 \\ 6 \end{pmatrix} \quad \boldsymbol{x}_4 = \begin{pmatrix} 8 \\ 2 \end{pmatrix} \quad \boldsymbol{x}_5 = \begin{pmatrix} 8 \\ 6 \end{pmatrix}$$

  y la siguiente partición inicial en dos clústers:

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



  ¿Cuál es la partición Π^* resultante tras aplicar el algoritmo C-medias de Duda y Hart?

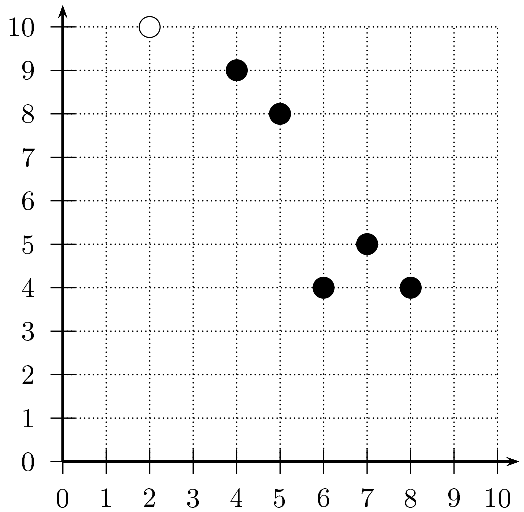
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

- ***Cuestión 2***: Dados los siguientes 6 vectores bidimensionales:

$$\boldsymbol{x}_1 = \begin{pmatrix} 2 \\ 10 \end{pmatrix} \boldsymbol{x}_2 = \begin{pmatrix} 8 \\ 4 \end{pmatrix} \boldsymbol{x}_3 = \begin{pmatrix} 5 \\ 8 \end{pmatrix} \boldsymbol{x}_4 = \begin{pmatrix} 7 \\ 5 \end{pmatrix} \boldsymbol{x}_5 = \begin{pmatrix} 6 \\ 4 \end{pmatrix} \boldsymbol{x}_6 = \begin{pmatrix} 4 \\ 9 \end{pmatrix}$$

y la siguiente partición inicial en dos clústers:

$$\Pi = \{X_1 = \{\boldsymbol{x}_1\}, X_2 = \{\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}\}$$



¿Cuál es la partición Π^* resultante tras aplicar el algoritmo C-medias de Duda y Hart?

Referencias

[1] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.