Sistemas Inteligentes

Escuela Técnica Superior de Informática

Universitat Politècnica de València

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

Índice

Índice

Clustering

Clustering es un problema generalmente mal definido: no existe una definición precisa y universalmente aceptada.

Según [Anderberg,1973], el objetivo del clustering es:

> Agrupar objetos en clases tales que los de una misma clase presenten un alto grado de ***asociación natural*** entre sí, mientras que las clases sean relativamente distintas unas de otras.

Dicho de otra manera:

> Encontrar ***agrupamientos naturales*** en un conjunto de objetos, de forma que la descripción de éstos se realice en términos de clases o grupos de objetos con fuertes semejanzas internas.

Dos tipos de clustering: ***Particional*** y ***Jerárquico***.

Índice

Clustering particional

Problema genérico:

Asumimos disponible una ***función criterio*** J para evaluar la calidad de cualquier partición de N datos en C clases. De este modo, el problema del clustering puede verse como uno de búsqueda del tipo:

$$\Pi^* = \operatorname*{arg\,min}_{\Pi=\{X_1,\dots,X_C\}} J(\Pi) \tag{1}$$

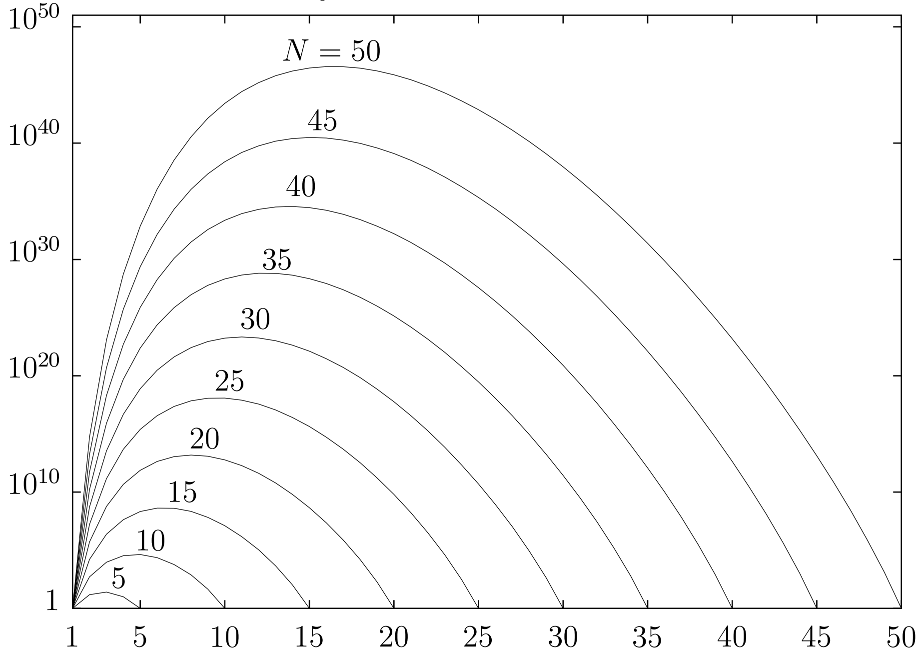
Dificultad:

El número de particiones a explorar es muy elevado incluso para valores pequeños de N y C (ver dcha.). No es factible buscar soluciones globalmente óptimas mediante técnicas de enumeración completa (explícita o implícita) salvo en casos particulares.

Solución:

Soluciones subóptimas obtenidas mediante algoritmos aproximados.

Número de particiones en función de C para varios N



Clustering particional: Criterio "suma de errores cuadráticos" (SEC)

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

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

Interpretación:

- En cada cluster X_c su *media*, $\boldsymbol{m}_c$, se interpreta como el "prototipo natural" de X_c. Cada dato $\boldsymbol{x} \in X_c$, se interpreta como una "versión distorsionada" de $\boldsymbol{m}_c$ y la distorsión de $\boldsymbol{x}$ se caracteriza por el *vector error* $\boldsymbol{x} - \boldsymbol{m}_c$.
- Como su nombre indica, el criterio SEC mide la suma (o media) de los cuadrados de las magnitudes de estos vectores error y, obviamente, es un criterio a minimizar.
- La *media* de cada cluster es el punto que representa los datos del cluster con menor SEC.

Ejemplo de clustering particional

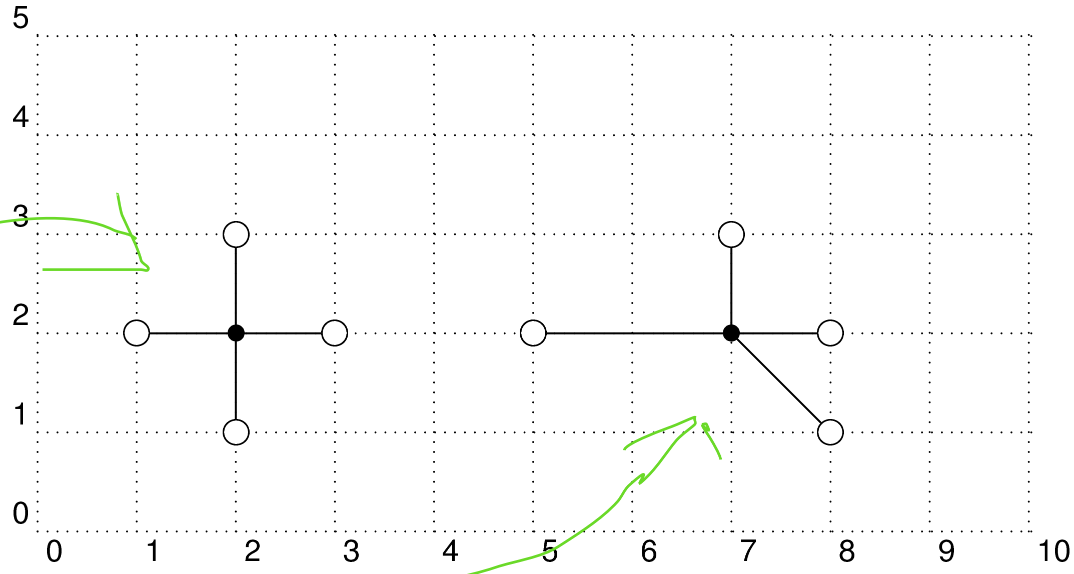
SEC = 1^2+1^2
+1^2+1^2

+

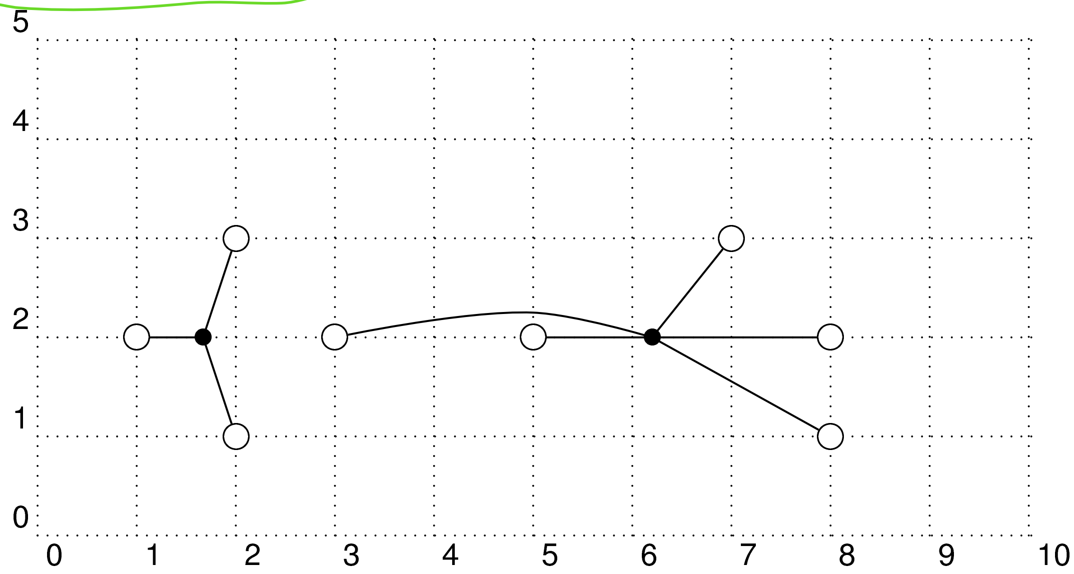
2^2+1^2+1^2
+sqrt(2)^2

=

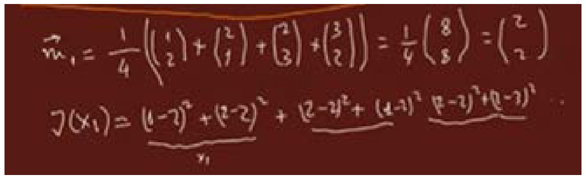
4 + 8 = 12



SEC$= 12$



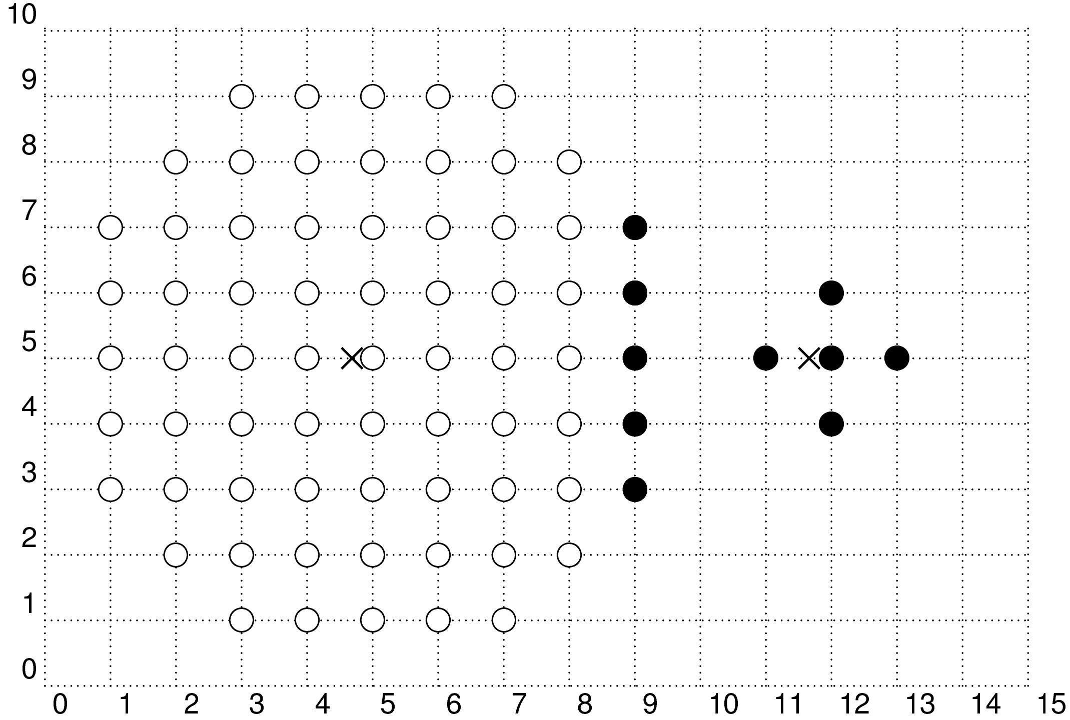
SEC> 12



Bondad del criterio SEC

El criterio SEC es apropiado sólo si los datos forman ***clusters hiperesféricos de tamaño similar.***

Si los tamaños de los clusters son muy distintos, es posible la agrupación natural *no* tenga el mínimo SEC:



Índice

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

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

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

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

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

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

$$\frac{n_j}{n_j + 1}\|\boldsymbol{x} - \boldsymbol{m}_j\|^2 < \frac{n_i}{n_i - 1}\|\boldsymbol{x} - \boldsymbol{m}_i\|^2 \qquad (3)$$

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; $\Pi = \{X_1, \dots, X_C\}$;

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

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

  $transfers =$ false

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

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

    if $n_i > 1$ **then**

      $j^* = \underset{1 \leq j \leq C}{\arg\min}\ d(\boldsymbol{x}, \boldsymbol{m}_j)$

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

        $transfers =$ true

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

      endif

    endif

  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:

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

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

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

El criterio SEC y Cuantificación Vectorial

Los siguientes criterios a minimizar son equivalentes:

$$J(X_1, \ldots, X_C) = \sum_c \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{m}_c\|^2 \qquad (4)$$

$$J(X_1, \ldots, X_C; \boldsymbol{r}_1, \ldots, \boldsymbol{r}_C) = \sum_c \sum_{\boldsymbol{x} \in X_c} \|\boldsymbol{x} - \boldsymbol{r}_c\|^2 \qquad (5)$$

$$J(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C) = \sum_{\boldsymbol{x}} \min_c \|\boldsymbol{x} - \boldsymbol{r}_c\|^2 \qquad (6)$$

Justificación:

- (5) equivale a (4) pues, para toda partición $X_1, \ldots, X_C$, los *representantes (de cluster)* $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C$ que minimizan (5) son las medias de los clusters.
- (5) equivale a (6) ya que, para todo conjunto de representantes $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_C$, la partición que minimiza (5) es aquella en la que cada dato se asigna al cluster de su representante más próximo.
- (6) se conoce como el problema del *diseño de un cuantificador vectorial* en *teoría de la información*

Otra interpretación del criterio SEC

El criterio SEC se puede reescribir sin incluir las medias de los clusters:

$$J(X_1, \ldots, X_C) = \frac{1}{2} \sum_c n_c \, \bar{s}_c \tag{7}$$

donde n_c es el número de datos en X_c y $\bar{s}_c$ es la media de las distancias Euclídeas al cuadrado entre todos estos datos:

$$\bar{s}_c = \frac{1}{n_c^2} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in X_c} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \tag{8}$$

Luego la SEC se puede interpretar como una suma ponderada de medias de distancias al cuadrado “intra-cluster”.

Con base en esta interpretación, podemos redefinir $\bar{s}_c$ para obtener criterios parecidos a la SEC (válidos incluso con datos *no-vectoriales*):

$$\bar{s}_c = \frac{1}{n_c^2} \sum_{x, x' \in X_c} d(x, x') \qquad\qquad \bar{s}_c = \frac{1}{n_c^2} \max_{x, x' \in X_c} d(x, x') \tag{9}$$