Reconocimiento de patrones

Clase 11: ISODATA

Para el día de hoy…

- ISODATA

ISODATA

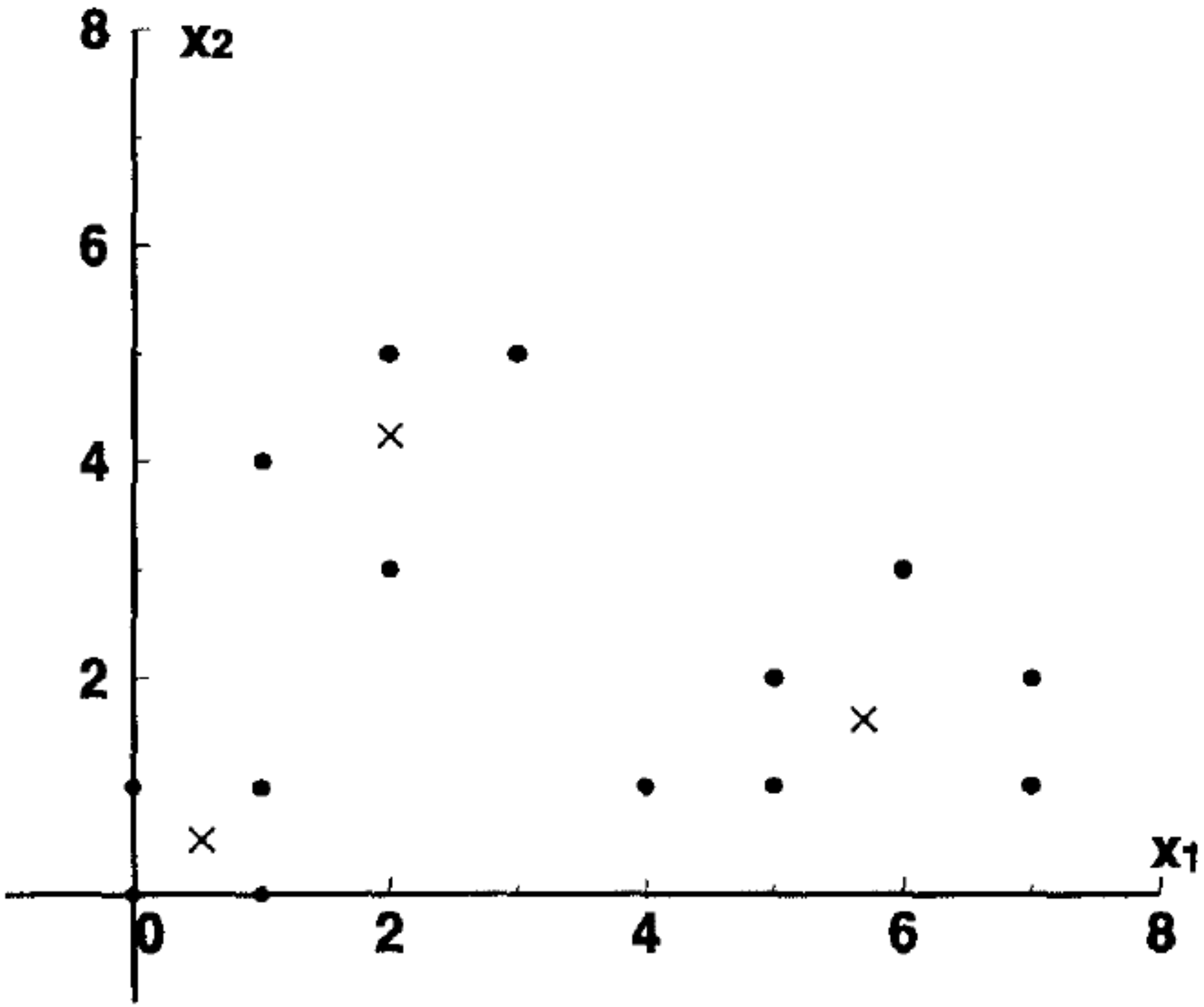
- Determina los centros de los grupos de forma iterativa como la media de sus muestras
- Incorpora varias heurísticas
- El usuario debe tener una idea del número de grupos
  - La solución no excederá dos veces la estimación inicial
  - Ni será menos de la mitad

Esqueleto del algoritmo

- Dados el número de grupos deseados, el mínimo de elementos por grupo, parámetros de agrupamiento, desviación estándar, grupos a agrupar e iteraciones
- Mientras no se cumpla el número de iteraciones
  - Distribuye las observaciones en los centros
  - Descarta los grupos con "pocas" muestras
  - Actualiza el centro de los grupos de acuerdo a la media
  - Calcula la distancia promedio de cada grupo y la distancia global
  - Si el número de grupos es menor a $\frac{k+1}{2}$ o impar menor a $2k$ intenta dividir un grupo
  - Detectar si se puede agrupar algún grupo

Ejemplo

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$



Ejemplo: inicialización

- Considere las observaciones y centros dados en la figura
  - $X = \left\{ \begin{matrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{matrix} \right\}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Input:

$n-$ the problem's dimension.

$m-$ the number of the given samples.

$X = \{x_i\}$, $1 \le i \le m-$ the m samples in R^n.

$Y = \{y_i\}$, $Z = \{z_i\}$, $1 \le i \le c-$ two identical sequences which contain the initial cluster centers.

$k-$ the desired number of clusters.

m_0- minimum allowed size of a cluster.

σ_0- standard deviation threshold (for splitting).

$\lambda-$ splitting fraction: $0 < \lambda \le 1$.

d_0- lumping threshold.

$l-$ maximum number of pairs of clusters which may be lumped simultaneously.

$\varepsilon-$ a given tolerance.

$N-$ maximum number of iterations allowed.

$S, L-$ vectors of size N. Initially

$$S(i) = L(i) = 2,\ 1 \le i \le N$$

After the $i-$th iteration, set $S(i) = 0$ or $L(i) = 0$ if splitting or lumping starts respectively. If splitting or lumping is completed successfully, set $S(i) = 1$ or $L(i) = 1$ respectively.

$NC-$ indicates a change in the set of cluster centers during the classification: Step 2 - Step 4.

Ejemplo: Paso 1

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 1. Set $it = 0$; $S(i) = L(i) = 2$, $1 \le i \le N$.

Ejemplo: Paso 2

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 2. Set $c' = c$, $z_j = y_j$ $1 \le j \le c$ and $NC = 1$.

Use the existing cluster centers and the minimum-distance principle to classify the samples, i.e.

$$x \in C_j \;\; iff \; \|x - y_j\| \le \|x - y_i\|, \;\; 1 \le i \le c, \;\; i \ne j \qquad (3.4.1)$$

for all $x \in X$, where C_j is the cluster centered at y_j with m_j samples $\{x_{l_{ij}}\}_{i=1}^{m_j}$.

C_1	C_2	C_3	C_4	C_5
$(0,0)^T$	$(4,1)^T$	$(7,1)^T$	$(2,3)^T$	$(2,5)^T$
$(1,0)^T$	$(5,1)^T$	$(7,2)^T$	$(1,4)^T$	$(3,5)^T$
$(1,1)^T$	$(5,2)^T$	$(6,3)^T$		
$(0,1)^T$				

Ejemplo: Paso 3

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 3. Each cluster center with fewer than m_0 samples is discarded. Its elements are distributed among the remaining clusters and we set $c \leftarrow c-1$.

Ejemplo: Paso 3

- Considere las observaciones y centros dados en la figura
    - $X = \left\{ \begin{matrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{matrix} \right\}$
    - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
    - $c = 5$
    - $k = 3$
    - $m_0 = 2$
    - $\sigma_0 = 1.5$
    - $\lambda = 0.5$
    - $d_0 = 2.5$
    - $l = 2$
    - $N = 10$

Step 3. Each cluster center with fewer than m_0 samples is discarded. Its elements are distributed among the remaining clusters and we set $c \leftarrow c-1$.

C_1	C_2	C_3	C_4	C_5
$(0,0)^T$	$(4,1)^T$	$(7,1)^T$	$(2,3)^T$	$(2,5)^T$
$(1,0)^T$	$(5,1)^T$	$(7,2)^T$	$(1,4)^T$	$(3,5)^T$
$(1,1)^T$	$(5,2)^T$	$(6,3)^T$		
$(0,1)^T$				

Ejemplo: Paso 4

- Considere las observaciones y centros dados en la figura
  - $X = \left\{ \begin{matrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{matrix} \right\}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 4. For $1 \leq j \leq c$ update the existing cluster centers by

$$\boldsymbol{y}_j \leftarrow \frac{1}{m_j} \sum_{i=1}^{m_j} \boldsymbol{x}_{l_{ij}} \tag{3.4.2}$$

If $c = c'$ and $\sum_{i=1}^{c} \left\| \boldsymbol{y}_j - \boldsymbol{z}_j \right\| < \varepsilon$ set $NC = 0$.

Ejemplo: Paso 4

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 4. For $1 \le j \le c$ update the existing cluster centers by

$$\boldsymbol{y}_j \leftarrow \frac{1}{m_j} \sum_{i=1}^{m_j} \boldsymbol{x}_{l_{ij}} \qquad (3.4.2)$$

If $c = c'$ and $\sum_{i=1}^{c} \left\| \boldsymbol{y}_j - \boldsymbol{z}_j \right\| < \varepsilon$ set $NC = 0$.

$$Y = \{(0.5,0.5)^T, (4.667,1.333)^T, (6.667,2)^T, (1.5,3.5)^T, (2.5,5)^T\}$$

Ejemplo: Paso 5

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 5. For $1 \le j \le c$ calculate the average distance of $\boldsymbol{x}_{I_{ij}}$, $1 \le i \le m_j$ from $\boldsymbol{y}_j$:

$$\bar{d}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \left\| \boldsymbol{x}_{I_{ij}} - \boldsymbol{y}_j \right\| \qquad (3.4.3)$$

Ejemplo: Paso 5

- Considere las observaciones y centros dados en la figura
  - $X = \left\{\begin{matrix}(0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T\end{matrix}\right\}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 5. For $1 \leq j \leq c$ calculate the average distance of $\boldsymbol{x}_{l_{ij}}$, $1 \leq i \leq m_j$ from $\boldsymbol{y}_j$:

$$\bar{d}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \left\| \boldsymbol{x}_{l_{ij}} - \boldsymbol{y}_j \right\| \tag{3.4.3}$$

$$\bar{d}_1 = 0.707, \bar{d}_2 = 0.654, \bar{d}_3 = 0.863, \bar{d}_4 = 0.707, \bar{d}_5 = 0.500$$

Ejemplo: Paso 6

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 6. Calculate the global average distance $\bar{d}$ of all the m samples from their respective cluster centers, i.e.

$$\bar{d} = \frac{1}{m}\sum_{j=1}^{c} m_j \bar{d}_j \qquad (3.4.4)$$

This is the end of an iteration. Set $it \leftarrow it + 1$.

Ejemplo: Paso 6

- Considere las observaciones y centros dados en la figura
  - $X = \left\{\begin{matrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{matrix}\right\}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 6. Calculate the global average distance $\bar{d}$ of all the m samples from their respective cluster centers, i.e.

$$\bar{d} = \frac{1}{m}\sum_{j=1}^{c} m_j \bar{d}_j \qquad (3.4.4)$$

This is the end of an iteration. Set $it \leftarrow it + 1$.

$$\bar{d}_1 = 0.707, \bar{d}_2 = 0.654, \bar{d}_3 = 0.863, \bar{d}_4 = 0.707, \bar{d}_5 = 0.500$$

$$\bar{d} = \left(4\bar{d}_1 + 3\bar{d}_2 + 3\bar{d}_3 + 2\bar{d}_4 + 2\bar{d}_5\right)/14 = 0.700$$

Ejemplo: Paso 7

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 7. If $it = N$ go to Step 13. Otherwise

(a) If $c \le \left[\frac{k+1}{2}\right]$ go to Step 8 (splitting a cluster).

(b) If $\left[\frac{k+1}{2}\right] < c < 2k$ and it is odd, go to Step 8.

(c) If $c \ge 2k$ go to Step 10 (lumping clusters).

(d) If $\left[\frac{k+1}{2}\right] < c < 2k$ and it is even, go to Step 10.

Ejemplo: Paso 8

- Considere las observaciones y centros dados en la figura
    - $X = \left\{ \begin{matrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{matrix} \right\}$
    - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
    - $c = 5$
    - $k = 3$
    - $m_0 = 2$
    - $\sigma_0 = 1.5$
    - $\lambda = 0.5$
    - $d_0 = 2.5$
    - $l = 2$
    - $N = 10$

Step 8. Trying to split. Set $S(it) = 0$. For every cluster denote the cluster center and the cluster samples by

$$\mathbf{y}_j = \left(y_j^{(1)}, y_j^{(2)}, \ldots, y_j^{(n)}\right)^T, \ 1 \le j \le c \tag{3.4.5}$$

$$\mathbf{x}_{l_{kj}} = \left(x_{l_{kj}}^{(1)}, x_{l_{kj}}^{(2)}, \ldots, x_{l_{kj}}^{(n)}\right)^T, \ 1 \le j \le c, \ 1 \le k \le m_j \tag{3.4.6}$$

respectively. Calculate the standard deviation vectors

$$\sigma_j = \left(\sigma_j^{(1)}, \sigma_j^{(2)}, \ldots, \sigma_j^{(n)}\right)^T, \ 1 \le j \le c \tag{3.4.7}$$

where

$$\sigma_j^{(i)} = \left(\frac{\sum_{k=1}^{m_j} \left(x_{l_{kj}}^{(i)} - y_j^{(i)}\right)^2}{m_j} \right)^{1/2}, \ 1 \le j \le c, \ 1 \le i \le n \tag{3.4.8}$$

Each $\sigma_j^{(i)}$ is the standard deviation of the $j-$th cluster population along the $i-$th coordinate. Denote $\sigma_j^{(i_0)} = \max \sigma_j^{(i)}$, $1 \le i \le n$ (clearly i_0 depends on j).

Ejemplo: Paso 8

- Considere las observaciones y centros dados en la figura
  - $X = \left\{\begin{matrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{matrix}\right\}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

C_1	C_2	C_3	C_4	C_5
$(0,0)^T$	$(4,1)^T$	$(7,1)^T$	$(2,3)^T$	$(2,5)^T$
$(1,0)^T$	$(5,1)^T$	$(7,2)^T$	$(1,4)^T$	$(3,5)^T$
$(1,1)^T$	$(5,2)^T$	$(6,3)^T$		
$(0,1)^T$				

$$\boldsymbol{\sigma}_1 = \begin{pmatrix} 0.500 \\ 0.500 \end{pmatrix}, \boldsymbol{\sigma}_2 = \begin{pmatrix} 0.471 \\ 0.471 \end{pmatrix}, \boldsymbol{\sigma}_3 = \begin{pmatrix} 0.471 \\ 0.816 \end{pmatrix}, \boldsymbol{\sigma}_4 = \begin{pmatrix} 0.500 \\ 0.500 \end{pmatrix}, \boldsymbol{\sigma}_5 = \begin{pmatrix} 0.500 \\ 0.000 \end{pmatrix}$$

Step 8. Trying to split. Set $S(it) = 0$. For every cluster denote the cluster center and the cluster samples by

$$\boldsymbol{y}_j = \left(y_j^{(1)}, y_j^{(2)}, \ldots, y_j^{(n)}\right)^T, \ 1 \le j \le c \qquad (3.4.5)$$

$$\boldsymbol{x}_{l_{kj}} = \left(x_{l_{kj}}^{(1)}, x_{l_{kj}}^{(2)}, \ldots, x_{l_{kj}}^{(n)}\right)^T, \ 1 \le j \le c, \ 1 \le k \le m_j \qquad (3.4.6)$$

respectively. Calculate the standard deviation vectors

$$\sigma_j = \left(\sigma_j^{(1)}, \sigma_j^{(2)}, \ldots, \sigma_j^{(n)}\right)^T, \ 1 \le j \le c \qquad (3.4.7)$$

where

$$\sigma_j^{(i)} = \left(\frac{\sum_{k=1}^{m_j}\left(x_{l_{kj}}^{(i)} - y_j^{(i)}\right)^2}{m_j}\right)^{1/2}, \ 1 \le j \le c, \ 1 \le i \le n \qquad (3.4.8)$$

Each $\sigma_j^{(i)}$ is the standard deviation of the $j-$th cluster population along the $i-$th coordinate. Denote $\sigma_j^{(i_0)} = \max \sigma_j^{(i)}$, $1 \le i \le n$ (clearly i_0 depends on j).

Ejemplo: Paso 9

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $\bar{d}_1 = 0.707, \bar{d}_2 = 0.654, \bar{d}_3 = 0.863, \bar{d}_4 = 0.707, \bar{d}_5 = 0.500$
  - $N = 10$

$$\boldsymbol{\sigma}_1 = \begin{pmatrix} 0.500 \\ 0.500 \end{pmatrix}, \boldsymbol{\sigma}_2 = \begin{pmatrix} 0.471 \\ 0.471 \end{pmatrix}, \boldsymbol{\sigma}_3 = \begin{pmatrix} 0.471 \\ 0.816 \end{pmatrix}, \boldsymbol{\sigma}_4 = \begin{pmatrix} 0.500 \\ 0.500 \end{pmatrix}, \boldsymbol{\sigma}_5 = \begin{pmatrix} 0.500 \\ 0.000 \end{pmatrix}$$

Step 9. For $j : 1 \le j \le c$ if $\sigma_j^{(i_0)} \le \sigma_0$ do not split the $j-$th cluster; otherwise split it, provided that *at least* one of the relations

$$c \le \left[\frac{k+1}{2}\right] \tag{3.4.9}$$

$$\bar{d}_j > \bar{d} \quad \text{and} \quad m_j \ge 2m_0 \tag{3.4.10}$$

holds. Splitting the $j-$th cluster is done as follows. The cluster center $\boldsymbol{y}_j$ is deleted while two new cluster centers $\boldsymbol{y}_{j+}, \boldsymbol{y}_{j-}$ defined as

$$\boldsymbol{y}_{j+} = \left(y_j^{(1)}, \ldots, y_j^{(i_0-1)}, y_j^{(i_0)} + \lambda\sigma_j^{(i_0)}, y_j^{(i_0+1)}, \ldots, y_j^{(n)}\right) \tag{3.4.11}$$

$$\boldsymbol{y}_{j-} = \left(y_j^{(1)}, \ldots, y_j^{(i_0-1)}, y_j^{(i_0)} - \lambda\sigma_j^{(i_0)}, y_j^{(i_0+1)}, \ldots, y_j^{(n)}\right) \tag{3.4.12}$$

are created, and we set $c \leftarrow c+1$. Thus, $\boldsymbol{y}_j$ is splitted along the i_0-th coordinate. The splitting is controlled by the parameter λ which ensures a noticeable but not dramatic change in the cluster centers arrangement. If splitting occurred, set $S(it) = 1$ and go to Step 2. Otherwise:

1. If $it > 1$, $L(it-1) = 0$ and $NC = 0$ go to Step 12.
2. If $it > 1$, $L(it-1) = 0$ and $NC = 1$ go to Step 2.
3. If $it > 1$, $L(it-1) \ne 0$ continue.
4. If $it = 1$ continue.

Ejemplo: Paso 10

- Considere las observaciones y centros dados en la figura
  - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
  - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
  - $c = 5$
  - $k = 3$
  - $m_0 = 2$
  - $\sigma_0 = 1.5$
  - $\lambda = 0.5$
  - $d_0 = 2.5$
  - $l = 2$
  - $N = 10$

Step 10. Lumping. Set $L(it) = 0$. If $c < 2$, $S(it) = 0$ and $NC = 0$, go to Step 12. If $c < 2$ $S(it) = 0$ and $NC = 1$, go to Step 2. If

$c < 2$ and $S(it) = 2$, go to Step 2; otherwise calculate all the distances between arbitrary two cluster centers, i.e.

$$d_{ij} = \|y_i - y_j\|, 1 \le i \le c-1, i+1 \le j \le c \qquad (3.4.13)$$

Rearrange $\{d_{ij}\}$ as a monotonic increasing sequence and denote by l' the number of $d_{ij}'s$ which do not exceed d_0. Consider now the first $l^* = \min(l, l')$ numbers of this sequence which satisfy

$$d_{i_1 j_1} \le d_{i_2 j_2} \le \ldots \le d_{i_{l^*} j_{l^*}} \le d_0 \qquad (3.4.14)$$

If $l^* = 0$ no lumping occurs: if $S(it) = 2$ go to Step 2 and if $S(it) = 0$ go to Step 12. If $l^* \neq 0$ set $L(it) = 1$ and continue.

Ejemplo: Paso 11-13

- Considere las observaciones y centros dados en la figura
    - $X = \begin{Bmatrix} (0,0)^T, (1,0)^T, (1,1)^T, (0,1)^T, \\ (4,1)^T, (5,1)^T, (5,2)^T, (7,1)^T, \\ (7,2)^T, (6,3)^T, (2,3)^T, (1,4)^T, \\ (2,5)^T, (3,5)^T \end{Bmatrix}$
    - $Y = \{(0,0)^T, (4,1)^T, (7,2)^T, (2,3)^T, (3,5)^T\}$
- Los parámetros
    - $c = 5$
    - $k = 3$
    - $m_0 = 2$
    - $\sigma_0 = 1.5$
    - $\lambda = 0.5$
    - $d_0 = 2.5$
    - $l = 2$
    - $N = 10$

Step 11. The lumping starts with the pair of cluster centers (i_1, j_1) and terminates with (i_{l*}, j_{l*}). Each two cluster centers are lumped together and if a given pair (i_r, j_r) is such that either the i_r − th or the j_r − th cluster center had already been lumped, this pair is ignored. The lumping is done by replacing the i_r − th and the j_r − th cluster centers by

$$y_{(i_r, j_r)} = \frac{m_{i_r} y_{i_r} + m_{j_r} y_{j_r}}{m_{i_r} + m_{j_r}} \tag{3.4.15}$$

i.e. by their center of gravity based on their current populations. Since y_{i_r} and y_{j_r} are deleted we also set $c \leftarrow c - 1$. When the lumping is completed go to Step 2.

Step 12. Output $\{y_j\}$, $1 \le j \le c$; *it* and stop.

Step 13. Output y_j, $1 \le j \le c$; 'number of iterations exceeded' and stop.

Comentarios

- Para una implementación exitosa de ISODATA es necesario experimentar con σ_0, λ, d_0 para encontrar valores apropiados
- Valores inapropiados pueden llevar a oscilaciones (secuencias infinitas de divisiones y agrupamientos)
- Esto puede suceder cuando $2\lambda\sigma_0 < d_0$

Ejercicio

- Revisar el algoritmo de Isodata
- Implementar el algoritmo en Python
- Ejecute el algoritmo de Isodata con los datos del ejemplo y grafique la solución encontrada
- De su opinión del algoritmo y contesté las 4 preguntas típicas de análisis y diseño de algoritmos

Bibliografía

- Introduction to Pattern Recognition: Statistical, Structural, Neural, and Fuzzy Logic Approaches. Libro de Abraham Kandel y Menachem Friedman

Para la otra vez…

- Agrupamiento y reconocimiento de patrones