

# Efectuare

## K-Means Cop.7.

40.

(Algoritmul K-means: aplicare în  $\mathbb{R}^2$ )

■ T.U. Dresden, 2006 summer, S. Hölldobler, A. Grossmann, HW3

Folosiți algoritmul K-means și distanța euclidiană pentru a grupa următoarele 8 instanțe din  $\mathbb{R}^2$  în 3 clustere:

$$A(2, 10), B(2, 5), C(8, 4), D(5, 8), E(7, 5), F(6, 4), G(1, 2), H(4, 9).$$

Se vor lua drept centroizi inițiali punctele  $A, D$  și  $G$ .

a. Rulați prima iterație a algoritmului K-means. Pe un grid de valori  $10 \times 10$  veți marca instanțele date, pozițiile centroizilor la începutul primei iterări și compoziția fiecărui cluster la finalul acestei iterări. (Trasați mediatoarele segmentelor determinate de centroizi, ca separatori ai clusterelor.)

b. Câte iterări sunt necesare pentru ca algoritmul K-means să conveargă? Desenați pe câte un grid rezultatul rulării fiecărei iterări.

<sup>896</sup> De exemplu, pentru dendrograma obținută la problema 1 se poate folosi notația simplă ("flat hierarchy"):

$$((x_1, x_2), ((x_3, ((x_4, x_5), x_6)), ((x_7, x_8), (x_9, x_{10}))))$$

sau, o variantă extinsă cu informații despre ordinea de formare a clusterelor și înălțimile nodurilor (interne) corespunzătoare clusterelor în dendrogramă.

$$((x_1, x_2))^{0.2}_{C1}, ((x_3, ((x_4, x_5))^{0.1}_{C1}, x_6))^{0.2}_{C5}{}^{0.3(6)}_{C7}, (((x_7, x_8))^{0.1}_{C2}, (x_9, x_{10}))^{0.1}_{C3}{}^{0.2(3)}_{C6})^{1.1}_{C8}{}^{1.77(3)}_{C9}.$$

916

d)

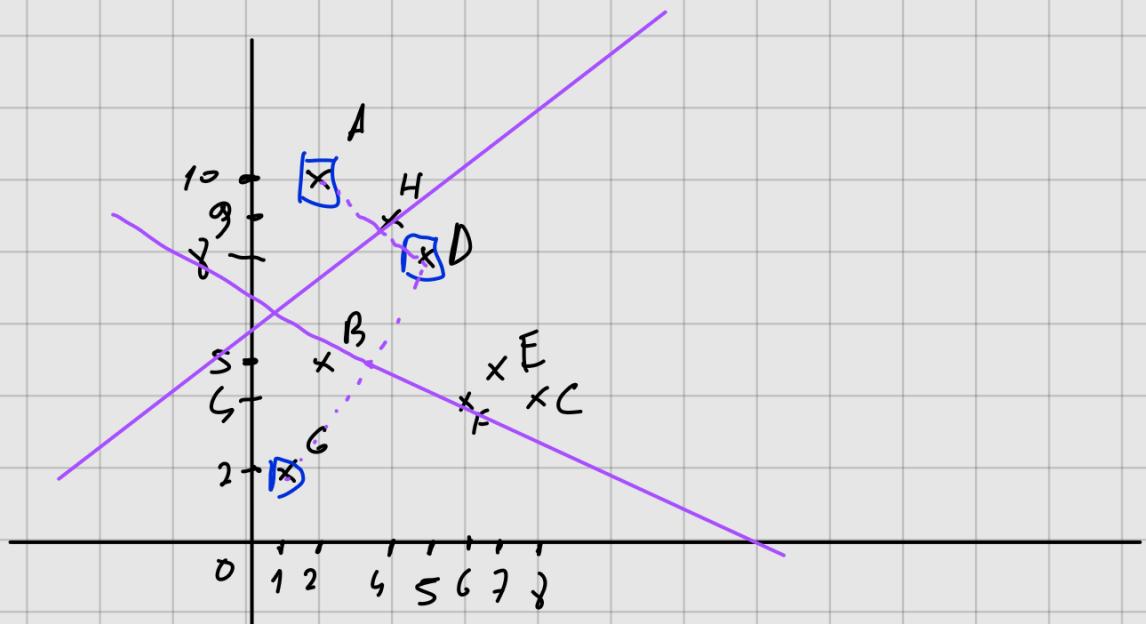
iterație 0: centroizi:

$$\mu_1^0 = \begin{bmatrix} 2 \\ 10 \end{bmatrix}, \mu_2^0 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}, \mu_3^0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$d(\cdot, \cdot)$	A	B	C	D	E	F	G	H
$\mu_1^0$	○							
$\mu_2^0$		○						
$\mu_3^0$				○				

$$\|\mu_2^0 - A\| = \left\| \begin{bmatrix} 5 \\ 8 \end{bmatrix} - \begin{bmatrix} 2 \\ 10 \end{bmatrix} \right\| = \left\| \begin{bmatrix} 3 \\ -2 \end{bmatrix} \right\| = \sqrt{9+4} = \sqrt{13}$$

Folosim met. vizuală:



$$\left\{ \begin{array}{l} \|K_1^o - H\| = \left\| \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 5 \\ 8 \end{bmatrix} \right\| = \left\| \begin{bmatrix} -3 \\ -7 \end{bmatrix} \right\| = \sqrt{5} \\ \|K_2^o - H\| = \left\| \begin{bmatrix} 5 \\ 3 \end{bmatrix} - \begin{bmatrix} 5 \\ 8 \end{bmatrix} \right\| = \left\| \begin{bmatrix} 0 \\ -5 \end{bmatrix} \right\| = \sqrt{2} \end{array} \right\}$$

$$d(K_1^o, H) > d(K_2^o, H) \Rightarrow$$

$\Rightarrow H$  ist kein zentraler Centroidalpunkt;  $K_2^o$ .

Mit. C:

$$\begin{aligned} \|K_3^o - F\| &= \left\| \begin{bmatrix} -5 \\ -2 \end{bmatrix} \right\| = \sqrt{29} \\ \|K_2^o - F\| &= \left\| \begin{bmatrix} 1 \\ 4 \end{bmatrix} \right\| = \sqrt{17} \end{aligned}$$

$$\left\{ \begin{array}{l} K_3^o: \sqrt{16 - 2,25} = \sqrt{13}, \dots \\ K_1^o: \sqrt{7^2 - 2^2} = \sqrt{45} \end{array} \right.$$

= lals. mögliche Clustere C:  $C_1^o = \{A\}$

$$C_2^o = \{B, D, E, F, H\}$$

$$C_3^o = \{C, G\}$$

Iter. 1:

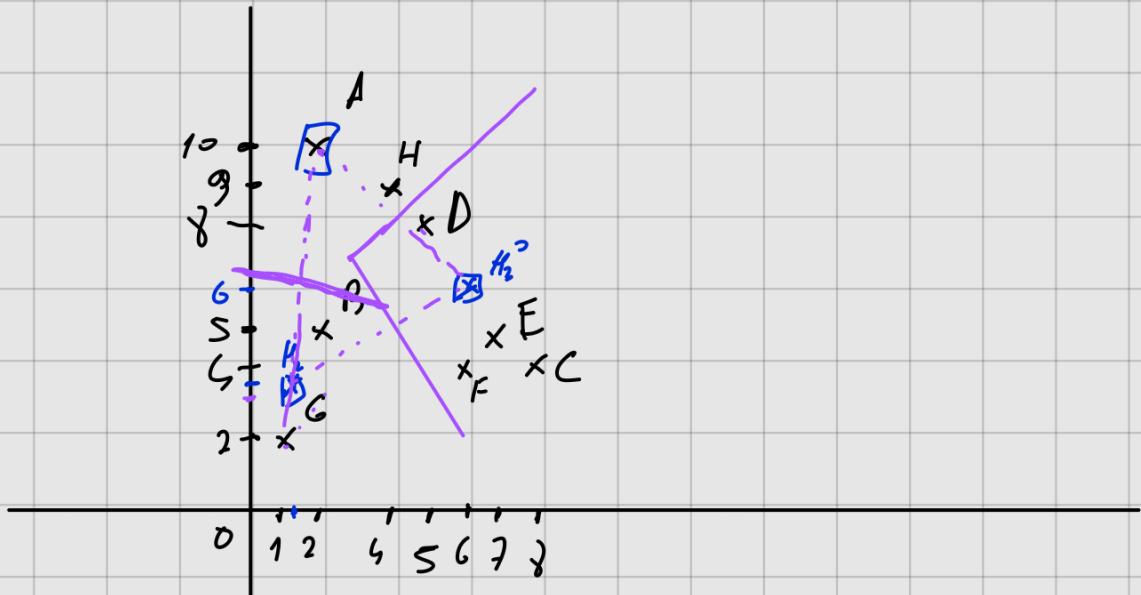
- calc. centroiden:  $K_1^1 = \frac{A}{1} = A$

$$K_2^1 = \frac{C + D + E + F + H}{5} = \underbrace{\left| \begin{bmatrix} 8 \\ 9 \end{bmatrix} \right| + \left| \begin{bmatrix} 5 \\ 8 \end{bmatrix} \right| + \left| \begin{bmatrix} 7 \\ 6 \end{bmatrix} \right| + \left| \begin{bmatrix} 6 \\ 7 \end{bmatrix} \right| + \left| \begin{bmatrix} 5 \\ 8 \end{bmatrix} \right|}_{j}$$

$$= \frac{\begin{vmatrix} 2 \\ 1 \end{vmatrix} + \begin{vmatrix} 1 \\ 1 \end{vmatrix}}{2} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

$$\mu_3 = \frac{B+G}{2} = \frac{\begin{vmatrix} 2 \\ 5 \end{vmatrix} + \begin{vmatrix} 1 \\ 2 \end{vmatrix}}{2} = \frac{\begin{vmatrix} 3 \\ 7 \end{vmatrix}}{2} = \begin{bmatrix} \frac{3}{2} \\ \frac{7}{2} \end{bmatrix}$$

5)



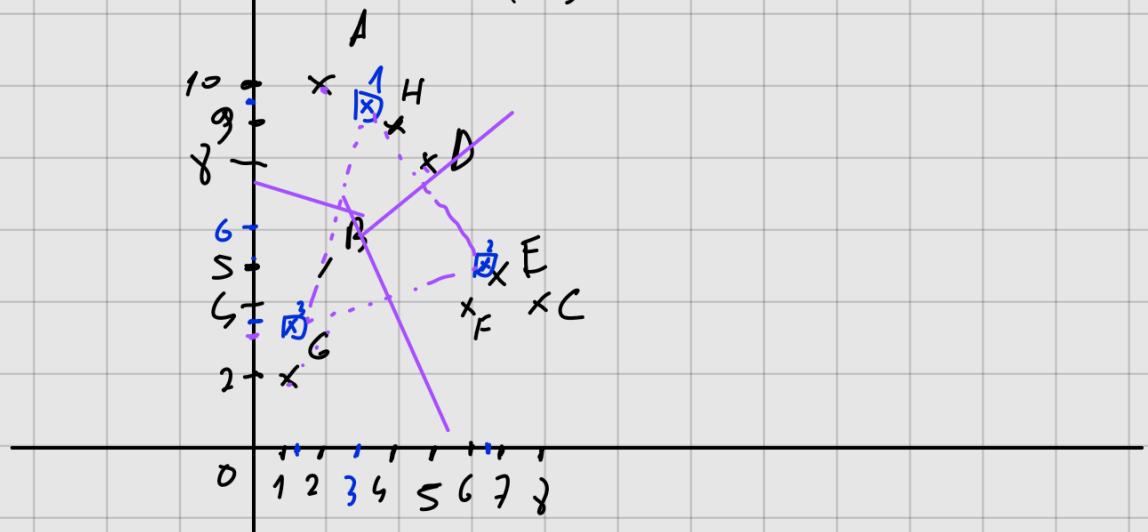
$$C_1' = \{A, H\}, C_2' = \{D, E, F, C\}, C_3' = \{B, G\}$$

if: 2 o

$$\mu_1^2 = \frac{A+H}{2} = \frac{\begin{vmatrix} 2 \\ 10 \end{vmatrix} + \begin{vmatrix} 1 \\ 9 \end{vmatrix}}{2} = \begin{bmatrix} \frac{6}{2} \\ \frac{9}{2} \end{bmatrix} = \begin{bmatrix} 3 \\ 9.5 \end{bmatrix}$$

$$\mu_2^2 = \frac{\begin{vmatrix} 2 \\ 8 \end{vmatrix} + \begin{vmatrix} 7 \\ 5 \end{vmatrix} + \begin{vmatrix} 6 \\ 4 \end{vmatrix} + \begin{vmatrix} 8 \\ 3 \end{vmatrix}}{4} = \frac{\begin{vmatrix} 26 \end{vmatrix}}{4} = \begin{bmatrix} 6.5 \\ 5.2 \end{bmatrix}$$

$$\mu_3^2 = \text{distanz} = \begin{bmatrix} 1.5 \\ 3.5 \end{bmatrix}$$



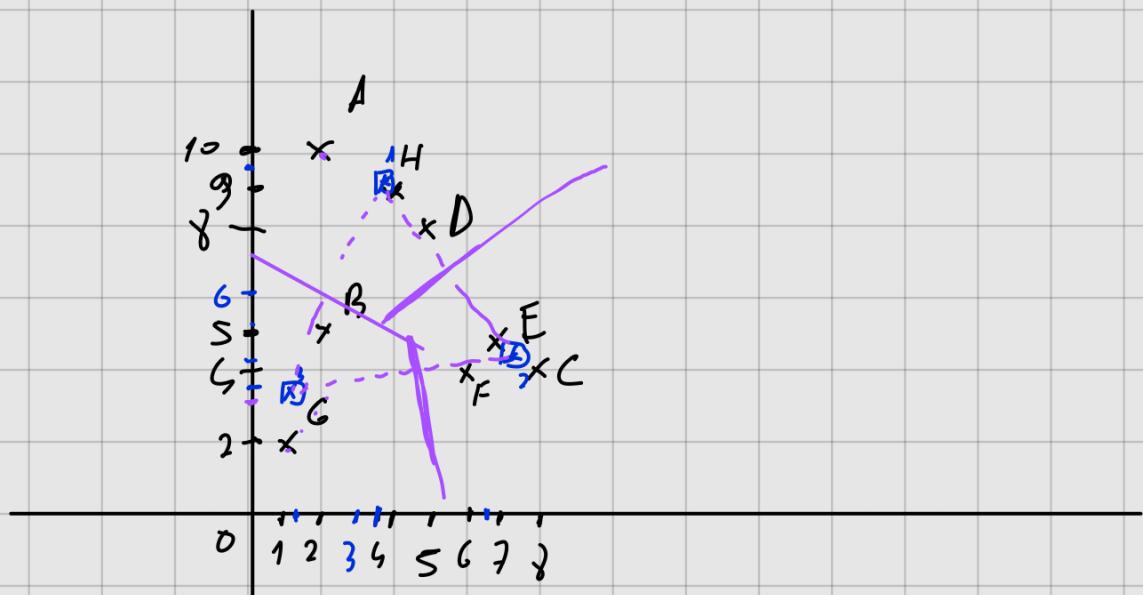
- ad. Clusterde:  $C_1^2 = \{A, H, D\}$ ,  $C_2^2 = \{E, F, C\}$ ,  $C_3^2 = \{B, G\}$

itas. 3:

$$\mu_1^3 = \frac{|1_9| + |5_8|}{3} = \frac{|11|}{3} = \left| \begin{smallmatrix} 3.6 \\ 9 \end{smallmatrix} \right|$$

$$\mu_2^3 = \frac{|2_1|}{3} = \left| \begin{smallmatrix} 7 \\ 5.3 \end{smallmatrix} \right|$$

$$\mu_3^3 = \mu_3^2$$



$C_1^3 = \{A, H, D\}$ ,  $C_2^3 = \{E, F, C\}$ ,  $C_3^3 = \{B, G\}$

$$\begin{cases} C_1^2 = C_1^3 \\ C_2^2 = C_2^3 \\ C_3^2 = C_3^3 \end{cases} \Rightarrow \text{Ne optim.}$$

Sunt mereu de 3 ferestre cu alg. să convergă.

41.

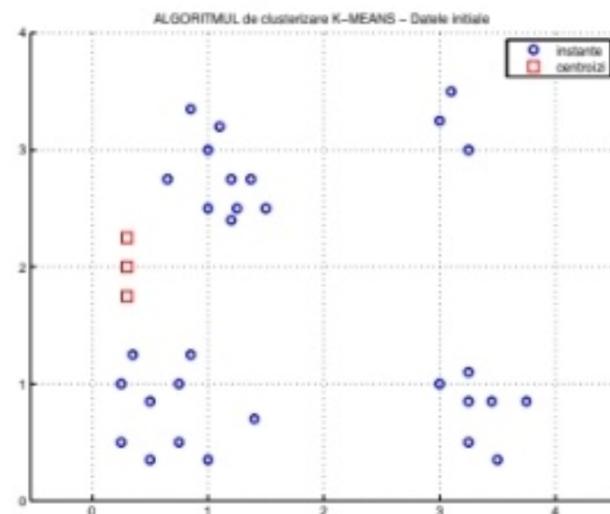
(Algoritmul  $K$ -means: aplicare pe date din  $\mathbb{R}^2$ )

\* CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 5

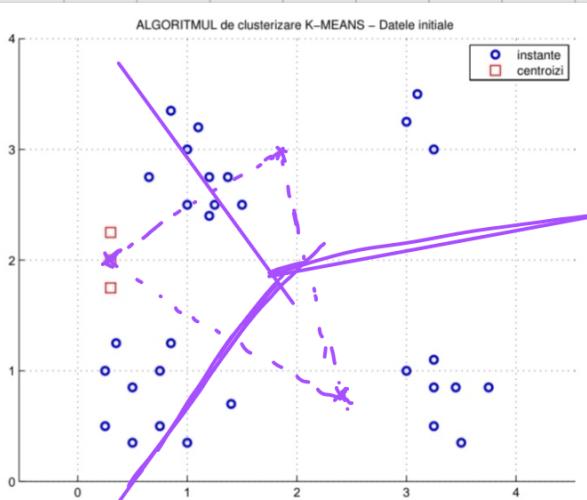
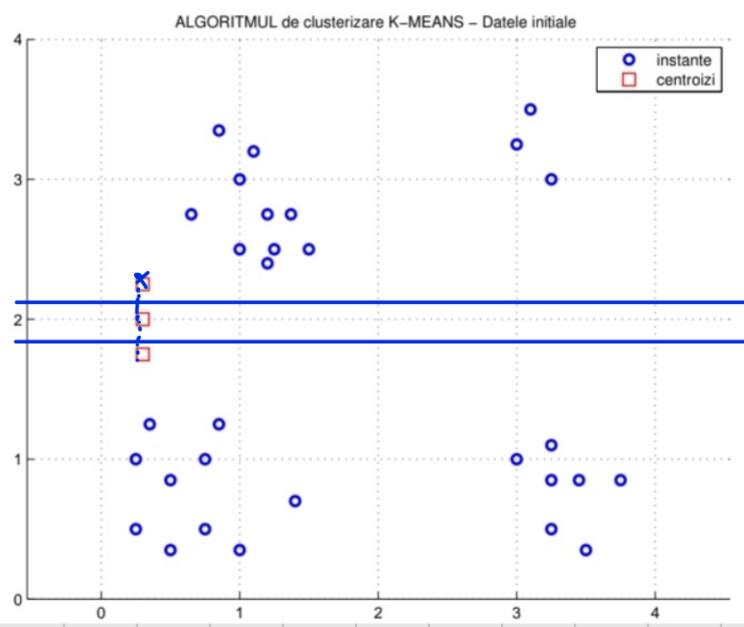
Aplicați algoritmul  $K$ -means pe setul de date din imaginea următoare.

Cerculețele reprezintă instanțele de clusterizat, iar pătrățelele sunt centroizii inițiali ai clusterelor. Pentru fiecare **iterație** a algoritmului desenați **centroizii și separatorii** care definesc fiecare cluster. Folosiți oricără imagini aveți nevoie până ajungeți la convergență.

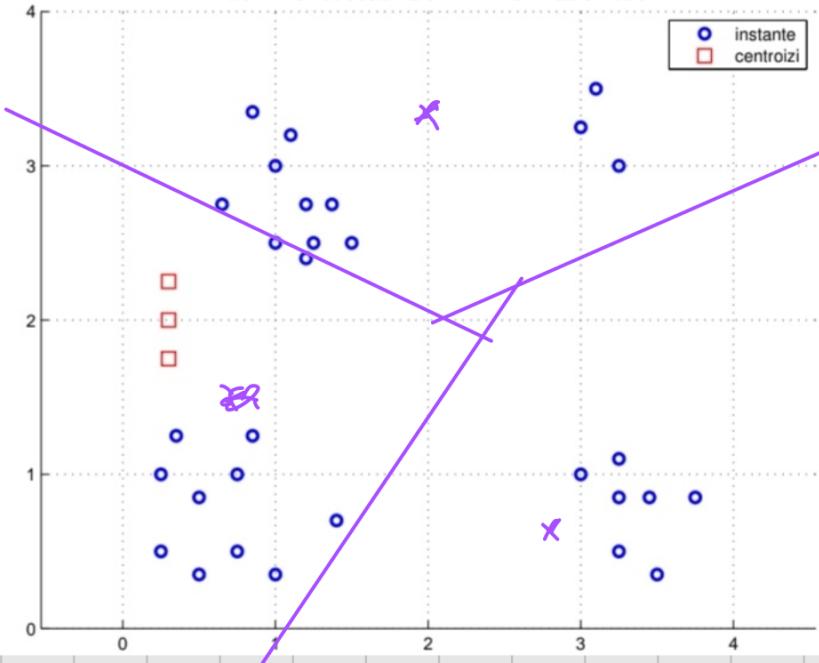
Coordonatele acestor instanțe, precum și cele ale centroizilor vă sunt puse la dispoziție în fișiere depuse pe site-ul acestei cărți.<sup>897</sup>



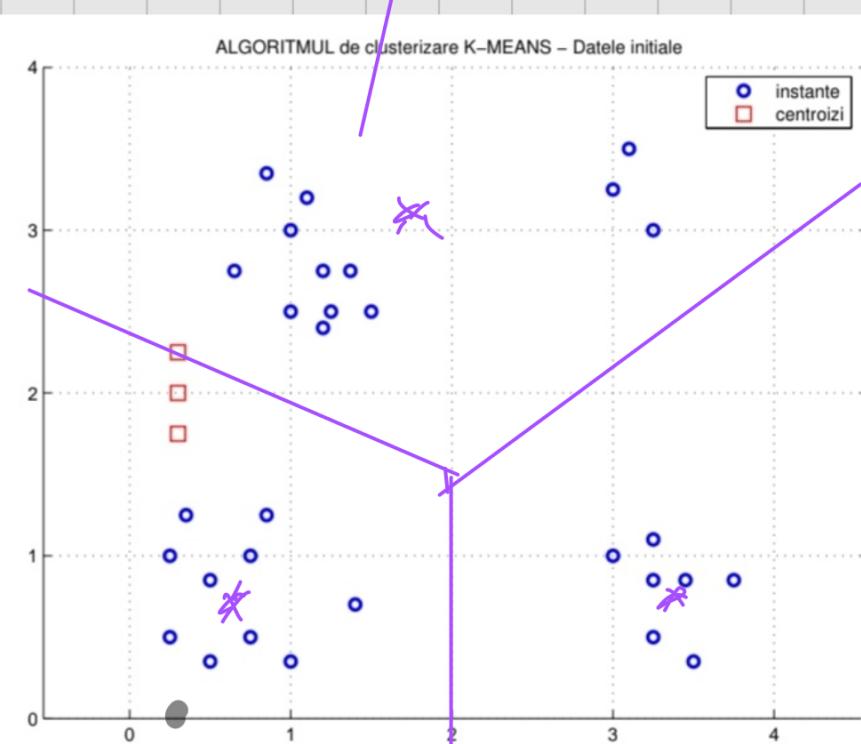
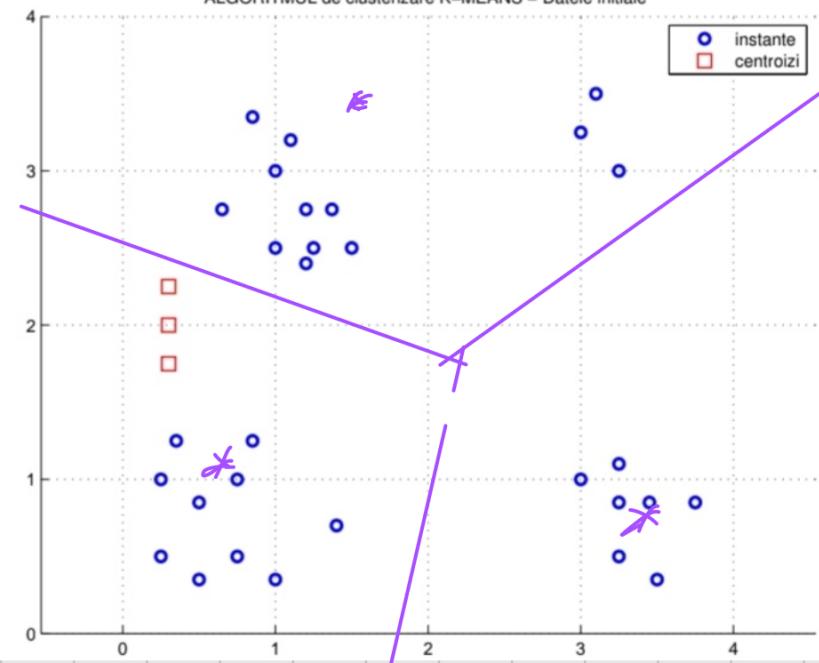
**Observație:** La execuția algoritmului se consideră că în cazul în care un centroid nu are puncte asignate lui, atunci el rămâne pe loc în iterată respectivă.



ALGORITMUL de clusterizare K-MEANS – Datele initiale



ALGORITMUL de clusterizare K-MEANS – Datele initiale



*Se termină execuție.*

42.

(Algoritmul K-means: aplicare în  $\mathbb{R}$  și  $\mathbb{R}^2$ ;  
verificarea monotoniei „criteriului” coeziunii intra-clustere ( $J$ ))  
 ■ (pt. punctul a) □ \* Liviu Ciortuz, 2020

În acest exercițiu veți folosi algoritmul K-means în varianta dată în enunțul exercițiului 12.<sup>898</sup>

Vă readucem aminte definiția aşa-numitului criteriu  $J$ , care este o măsură a coeziunii intra-clustere:

$$J(C^{(t)}, \mu^{(t)}) = \sum_{i=1}^n (x_i - \mu_{C^{(t)}(x_i)}^{(t)})^2,$$

unde  $C^{(t)}$  este ansamblul clusterelor la momentul / iterația  $t$ , apoi  $\mu^{(t)}$  desemnează ansamblul centroizilor la iterația  $t$  și, în sfârșit,  $\mu_{C^{(t)}(x_i)}^{(t)}$  este centroidul clusterului la care este asignată instanța  $x_i$  la iterația  $t$ .

a. Fie următorul set de date din  $\mathbb{R}$ :  $-9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9$ . Considerăm  $\mu_1^{(0)} = -20$  și  $\mu_2^{(0)} = -10$ .<sup>899</sup> Demonstrați în manieră analitică (NU numeric!) că pentru  $t = 1$  avem

$$J(C^{(t)}, \mu^{(t)}) \leq J(C^{(t-1)}, \mu^{(t-1)}).$$

a)  $\mu_1^{(0)} = -20, \mu_2^{(0)} = -10$ .

Din  $\forall t \geq 1$  avem  $J(C^{(t)}, \mu^{(t)}) \leq J(C^{(t-1)}, \mu^{(t-1)}) \Rightarrow$

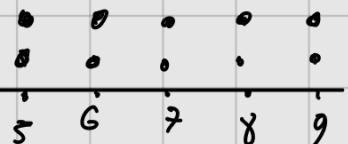
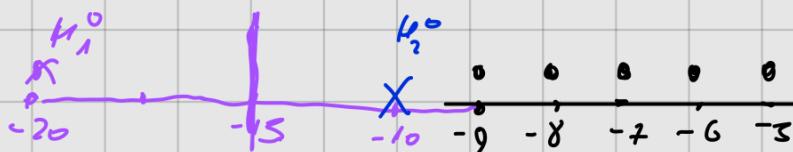
$$\Rightarrow J(C^0, \mu^0) \geq J(C^1, \mu^1) \text{ pt } t = 1$$

$\forall t = 0$ :

$$\begin{cases} \mu_1^0 = -20 \\ \mu_2^0 = -10 \end{cases}$$

calc. mediana:

$$\frac{-20 - 10}{2} = -15$$

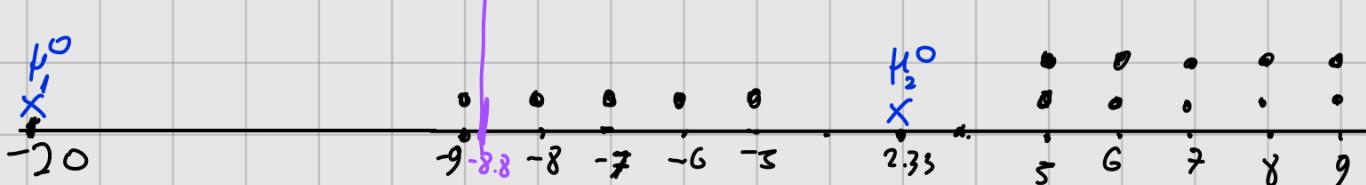


$$C_1^0 = \{\varnothing\}, C_2^0 = \{-9, -8, \dots\}$$

$\forall t = 1$ :

$$\mu_1^1 = -20$$

$$\mu_2^1 = \frac{-8 - 7 - 6 - 5 + 5 + 6 + 7 + 8 + 9 + 8 + 7 + 6 + 5 + 9}{15} = \frac{11 + 24}{15} = \frac{35}{15} = \frac{7}{3} = 2.3$$



calc. mediana:  $\frac{-20 + 2.3}{2} = \frac{-17.7}{2} = -8.8 \dots$

$$C_1^1 = \{-9\}$$

$$C_2^1 = \{-8, \dots, 9\}$$

Ca răbdare:  $\hat{J}(C^0, \mu^0) \geq \hat{J}(C^1, \mu^1) \rightarrow \hat{J}(C^0, \mu^0) \stackrel{(1)}{\geq} \hat{J}(C^1, \mu^1) \stackrel{(2)}{\geq} J(C^1, \mu^1)$

Dem. îneg.(1)

$$\hat{J}(C^0, \mu^0) = 0 + \| -9 - (-10) \|^2 + \| -8 - (-10) \|^2 + \dots + \| 9 - (-10) \|^2$$

$$\hat{J}(C^0, \mu^1) = 0 + \| -9 - \frac{7}{3} \|^2 + \dots + \| 9 - \frac{7}{3} \|^2$$

$$\hat{J}(C^0, \mu^0) = p(-10)$$

$$\hat{J}(C^0, \mu^1) = p\left(\frac{7}{3}\right)$$

Fie  $f(x) = (-9-x)^2 + (-8-x)^2 + \dots + (9-x)^2$

$$= 81 + \underbrace{18x}_{b_1} + x^2 + 64 + \underbrace{16x}_{b_2} + x^2 + \dots + 81 - \underbrace{18x}_{b_{15}} + x^2 =$$

$$= \underbrace{(18 + 16 + 16 + 16 + 16 - 16 - 10 - 12 - 14 - 16 - 16 - 18 - 18)}_b X + \underbrace{15x^2}_a =$$

$$= (-12 - 30 - 18)X + 15x^2 = -70X + 15x^2$$

$\Leftrightarrow a = 15 > 0 \Rightarrow$  mult. do min =)

$$\Rightarrow -\frac{b}{2a} = \frac{70}{30} = \frac{7}{3} \rightarrow$$
 mult. do min =)  $\forall X \neq \frac{7}{3} \Rightarrow X > \frac{7}{3}$

Deci,  $\hat{J}(C^0, \mu^0) \geq \hat{J}(C^1, \mu^1)$

Dem. îneg.(2).

$$\hat{J}(C^0, \mu^1) = 0 + \| -9 - \frac{7}{3} \|^2 + \dots + \| 9 - \frac{7}{3} \|^2$$

$$\hat{J}(C^1, \mu^1) = \| -9 - (-2) \|^2 + \| -8 - \frac{7}{3} \|^2 + \dots + \| 9 - \frac{7}{3} \|^2$$

Bin comp. termen cu termen  $\Rightarrow \hat{J}(C^0, \mu^1) \geq \hat{J}(C^1, \mu^1)$

Deci, din dem(1) și(2) am arătat că mereu  $\hat{J}(C^0, \mu^0) \geq \hat{J}(C^1, \mu^1)$

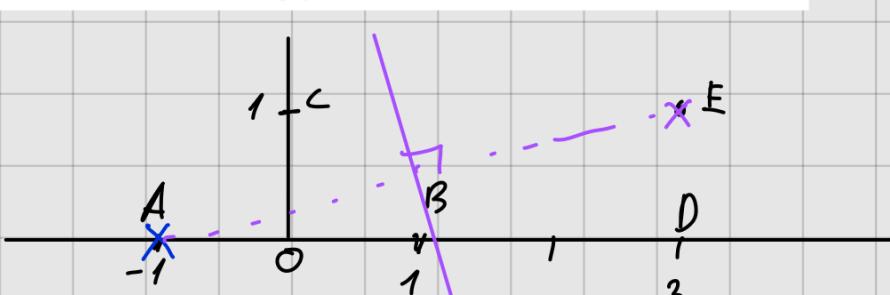
b. Cerințele de la acest punct sunt similare cu cele de la punctul precedent, însă de data aceasta veți lucra pe următorul set de date din  $\mathbb{R}^2$ :  $A(-1, 0)$ ,

<sup>897</sup> <http://prof.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2004f.TM+AM.HW3.pr5.cl.dat>,

<sup>898</sup> <http://prof.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2004f.TM+AM.HW3.pr5.init.dat>.

<sup>899</sup> Atenție! În eventualitatea că la o iterație oarecare a algoritmului K-means un cluster este vid, centroidul său rămâne pe loc la iterarea respectivă.

<sup>900</sup> Acestea sunt datele de la exercițiul 17.



$B(1, 0), C(0, 1), D(3, 0), E(3, 1)$ ; dreptul centroizii inițiali vezi consideră  $\mu_1^{(0)} = A$  și  $\mu_2^{(0)} = E$ .<sup>900</sup>

$$\begin{array}{c|cc}
 & x_1 & x_2 \\
 \hline
 A & -1 & 0 \\
 B & 1 & 0 \\
 C & 0 & 1 \\
 D & 3 & 0 \\
 E & 3 & 1
 \end{array}
 \quad \text{d.e. o: } \left\{ \begin{array}{l} \mu_1^0 = \left| \begin{smallmatrix} -1 \\ 0 \end{smallmatrix} \right| \quad C_1^0 = \{A, B, C\} \\ \mu_2^0 = \left| \begin{smallmatrix} 3 \\ 0 \end{smallmatrix} \right| \quad C_2^0 = \{D, E\} \end{array} \right.$$

$$\text{it 1: } \left\{ \begin{array}{l} \mu_1' = \frac{A+B+C}{3} = \frac{\left| \begin{smallmatrix} -1 \\ 0 \end{smallmatrix} \right| + \left| \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right| + \left| \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right|}{3} = \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \\ \mu_2' = \frac{D+E}{2} = \frac{\left| \begin{smallmatrix} 3 \\ 0 \end{smallmatrix} \right| + \left| \begin{smallmatrix} 3 \\ 1 \end{smallmatrix} \right|}{2} = \left| \begin{smallmatrix} 3 \\ \frac{1}{2} \end{smallmatrix} \right| \end{array} \right.$$

$$C_1' = \{A, B, C\}, \quad C_2' = \{D, E\}$$

$$\text{Dem } j(C^0, \mu^0) \stackrel{(1)}{\geq} j(C^0, \mu') \stackrel{(2)}{\geq} j(C', \mu')$$

Dem. imed. (1):

$$\begin{aligned}
 j(C^0, \mu^0) &= \left\| \left| \begin{smallmatrix} -1 \\ 0 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 3 \\ 0 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 3 \\ 0.5 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 3 \\ 1 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 3 \\ 0.5 \end{smallmatrix} \right| \right\|^2 \\
 j(C^0, \mu') &= \left\| \left| \begin{smallmatrix} -1 \\ 0 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 3 \\ 0 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 3 \\ 0.5 \end{smallmatrix} \right| \right\|^2 + \left\| \left| \begin{smallmatrix} 3 \\ 1 \end{smallmatrix} \right| - \left| \begin{smallmatrix} 3 \\ 0.5 \end{smallmatrix} \right| \right\|^2
 \end{aligned}$$

$$j(C^0, \mu^0) = f(\left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right|) + g(\left| \begin{smallmatrix} 3 \\ 0.5 \end{smallmatrix} \right|)$$

$$j(C^0, \mu') = f(\left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right|) + g(\left| \begin{smallmatrix} 3 \\ 0.5 \end{smallmatrix} \right|)$$

$$\text{fie } f(x) = (\left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| - x)^2 + (\left| \begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right| - x)^2 + (\left| \begin{smallmatrix} 0 \\ 1 \end{smallmatrix} \right| - x)^2$$

$$\left| \begin{smallmatrix} 0 \\ 0.33 \end{smallmatrix} \right| \leq$$

$$\begin{aligned}
 & (-1 - x_1)^2 + (0 - x_2)^2 + (1 - x_1)^2 + (0 - x_2)^2 + (0 - x_1)^2 + (1 - x_2)^2 \\
 & 1 + 2x_1 + x_1^2 + x_2^2 + 1 - 2x_1 + x_1^2 + x_2^2 + x_1^2 + 1 - 2x_2 + x_2^2
 \end{aligned}$$

$$\underbrace{f_1(x_1)}_{(2+0+3x_1^2)} + \underbrace{f_2(x_2)}_{(1-2x_2+3x_2^2)}$$

$$d_1 = 3 > 0 \rightarrow \text{rel. min} \Rightarrow x_1 = 0$$

$$d_2 = 3 > 0 \rightarrow \text{rel. min} \Rightarrow x_2 = \frac{2}{6} = \frac{1}{3}$$

$$g(y) = (| \frac{3}{0} - y |)^2 + (| \frac{3}{1} - y |)^2$$

$$= (3-y_1)^2 + (0-y_2)^2 + (3-y_1)^2 + (1-y_2)^2$$

$$= 9 - 6y_1 + y_1^2 + y_2^2 + 9 - 6y_2 + y_2^2 + 1 - 2y_2 + y_2^2$$

$$= \underbrace{(18 - 12y_1 + 2y_1^2)}_{g_1(y_1)} + \underbrace{(1 - 2y_2 + 2y_2^2)}_{g_2(y_2)}$$

$$d_1 = 2 > 0 \rightarrow \text{rel. dom. min} \Rightarrow y_1 = \frac{12}{9} = \frac{6}{3} = 3$$

$$d_2 = 2 > 0 \rightarrow \text{rel. dom. min} \Rightarrow y_2 = \frac{2}{4} = \frac{1}{2}$$

$$\hat{j}(c^0, u^0) = f_1(-1) + f_2(0) + g_1(3) + g_2(1)$$

$$\hat{j}(c^0, u^1) = f_1(0) + f_2(\frac{1}{3}) + g_1(3) + g_2(\frac{1}{2})$$

Bem. ineq. (2):

$$\hat{j}(c^0, u^1) = f\left(\left|\frac{0}{3}\right|\right) + g\left(\left|\frac{3}{1}\right|\right)$$

$$\text{Auch } c_1^0 = c_1^1 \text{ ?? } c_2^0 = c_2^1 //$$

$$\hat{j}(c^1, u^1) = f\left(\left|\frac{0}{1}\right|\right) + g\left(\left|\frac{3}{2}\right|\right)$$

$$\text{Drei; dim } (1), (2) \Rightarrow \hat{j}(c^0, u^0) \geq \hat{j}(c^1, u^1)$$

46.

(Algoritmul  $K$ -means: aplicare pe date din  $\mathbb{R}^3$   
calcularea variației / coeziunii intra- și inter-clustere)

prelucrare de Liviu Ciortuz, 2021, după

□ • Andreas Wickert, Luis Sa-Couto,

Machine Learning – A Journey to Deep Learning, 2021, pag. 370-385

Fie următoarele instanțe de antrenament, neetichetate:

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, x_2 = \begin{pmatrix} 8 \\ 8 \\ 4 \end{pmatrix}, x_3 = \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix}, x_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, x_5 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, x_6 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}.$$

a. Aplicați algoritmul  $K$ -means cu  $K = 2$  pe acest set de date, până ajungeți la convergență. Pentru inițializarea centroizilor, veți lua primele  $K$  instanțe din setul de date furnizat.

b. Aplicați algoritmul  $K$ -means cu  $K = 3$  pe același set de date, până ajungeți la convergență. Pentru inițializarea centroizilor, veți lua primele  $K$  instanțe din setul de date furnizat.

c. Ce valoare a lui  $K \in \{2, 3\}$  produce o clusterizare mai bună dacă se folosește drept criteriu de evaluare variația / coeziunea intra-clustere (adică, suma pătratelor distanțelor de la fiecare instanță la centroidul cel mai apropiat)?

d. Ce valoare a lui  $K \in \{2, 3\}$  produce o clusterizare mai bună dacă se folosește drept criteriu de evaluare variația / coeziunea inter-clustere? (Vedeti definiția de la problema 45.)

a)  $\mu_1^0 = x_1, \mu_2^0 = x_2$

$$\left\| \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \\ x_3' \end{pmatrix} \right\| = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_3 - x_3')^2}$$

$\hat{t}=0$ .

$$C_1 = \{x_1, x_4, x_5, x_3, x_6\}$$

$$\left\| \mu_1^0 - x_3 \right\| = \sqrt{5^2 + 5^2 + 5^2} = \sqrt{64} \quad X$$

$$C_2 = \{x_2\}$$

$$\left\| \mu_1^0 - x_3 \right\| = \sqrt{(-2)^2 + (-2)^2 + 0^2} = \sqrt{8} \quad \checkmark$$

$\hat{t}=1$ :

$$\mu_1^1 = \frac{x_1 + x_3 + x_5 + x_3 + x_6}{5} = \frac{\begin{pmatrix} 7 \\ 6 \\ 2 \end{pmatrix}}{5} = \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix}$$

$$\mu_2^1 = \mu_2^0 = x_2$$

$$C_1 = \{x_1, x_3, x_5, x_3, x_6\}$$

$$C_2 = \{x_2\}$$

$$\begin{cases} C_1^0 = C_1^1 \\ C_2^0 = C_2^1 \end{cases} \Rightarrow \text{ne optim.}$$

b)

$$\text{if } \sigma. \quad \mu_1^{\sigma} = x_1, \quad \mu_2^{\sigma} = x_2, \quad \mu_3^{\sigma} = x_3$$

$$C_1 = \{x_1, x_3, x_5\}, \quad C_2 = \{x_2\}, \quad C_3 = \{x_3, x_6\}$$

$x_6$  este mai aproape de  $\mu_1^{\sigma}$  sau  $\mu_3^{\sigma}$ ?

$$\begin{aligned} \|\mu_1^{\sigma} - x_6\| &= \sqrt[3]{9+9+1} = \sqrt[3]{9} \\ \|\mu_3^{\sigma} - x_6\| &= \sqrt[3]{1+1} = \sqrt[3]{2} \end{aligned} \quad \Rightarrow d(x_6, \mu_3^{\sigma}) < d(x_6, \mu_1^{\sigma})$$

$$\text{if } \tau: \quad \mu_1' = \frac{x_1 + x_3 + x_5}{3} = \begin{pmatrix} \frac{1}{3} \\ \frac{3}{3} \\ \frac{5}{3} \end{pmatrix} \quad \mu_2' = \mu_2^{\sigma} = x_2 = \begin{pmatrix} 8 \\ 8 \\ 1 \end{pmatrix}$$

$$\mu_3' = \frac{x_3 + x_6}{2} = \begin{pmatrix} \frac{3}{2} \\ \frac{5}{2} \\ \frac{1}{2} \end{pmatrix}$$

$$C_1' = \{x_1, x_3, x_5\}, \quad C_2 = \{x_2\}, \quad C_3 = \{x_3, x_6\}$$

$\square \quad C_1' = C_1^{\sigma}, \quad C_2' = C_2^{\sigma}, \quad C_3' = C_3^{\sigma} \Rightarrow \text{no change}$

c) calc. coeeficienții wt. output:

$$\text{wt. } K=2: \quad J(C', \mu') = \underbrace{\left\| \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{2}{3} \\ \frac{3}{3} \\ \frac{5}{3} \end{pmatrix} \right\|^2}_{C_1} + \left\| \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \end{pmatrix} \right\|^2 +$$

$$+ \left\| \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \\ \frac{1}{5} \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{8}{5} \\ \frac{9}{5} \\ \frac{-2}{5} \\ \frac{1}{5} \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} 3 \\ 2 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{7}{5} \\ \frac{6}{5} \\ \frac{2}{5} \\ \frac{1}{5} \end{pmatrix} \right\|^2 +$$

$$\begin{aligned} \text{C}_2 &= \left( \sqrt{\left(-\frac{2}{5}\right)^2 + \left(-\frac{6}{5}\right)^2 + \left(\frac{-2}{5}\right)^2} \right)^2 + \left( \sqrt{\left(\frac{8}{5}\right)^2 + \left(\frac{9}{5}\right)^2 + \left(-\frac{2}{5}\right)^2} \right)^2 + \\ &\quad + \left( \sqrt{\left(-\frac{7}{5}\right)^2 + \left(-\frac{6}{5}\right)^2 + \left(\frac{3}{5}\right)^2} \right)^2 + \left( \sqrt{\left(-\frac{7}{5}\right)^2 + \left(-\frac{1}{5}\right)^2 + \left(\frac{-2}{5}\right)^2} \right)^2 + \\ &\quad + \left( \sqrt{\left(\frac{8}{5}\right)^2 + \left(\frac{4}{5}\right)^2 + \left(\frac{3}{5}\right)^2} \right)^2 = \end{aligned}$$

$$= 1,76 + 5,96 + 3,76 + 2,16 + 3,56 = 17,2$$

$$\text{pt } k=3 : \bar{J}(c^1, k^1) = \left\| \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|^2 +$$

$$+ 0 + \left\| \begin{pmatrix} 3 \\ 2 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} 3 \\ 2 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \right\|^2 =$$

$$\begin{aligned} &= \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \\ &+ \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + 0 + \left(\frac{12}{5}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \\ &+ \left(\frac{12}{5}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \end{aligned}$$

$$= 0,67 + 0,67 + 0,67 + 0 + 6,26 + 6,26 = 14,53$$

$\Rightarrow$  vol pt  $k=2 >$  vol pt  $k=3$

$\Rightarrow$  closest zone

more pt  $k=3$

(more bren)

$$B(X) = \sum_{j=1}^K \left( \frac{\sum_{i=1}^n \gamma_{ij}}{n} \right) \|\boldsymbol{\mu}_j - \bar{\mathbf{x}}\|^2.$$

$$C_1' = \{x_1, x_3, x_5, x_6\} \quad C_2' = \{x_2\}$$

$$\mu_1 = \begin{pmatrix} 1.5 \\ 1.2 \\ 0.4 \end{pmatrix} \quad |x_2 = \begin{pmatrix} 8 \\ 8 \end{pmatrix}$$

$$B_X = \sum_{j=1}^{K=2} \left( \frac{\sum_{i=1}^{m=6} (\gamma_{ij})}{m=6} \right) \| \mu_j - \bar{x} \|^2$$

↑ cardinal.

$$\gamma_{ij} \leftarrow \begin{cases} 1, & \text{dacă } \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \in C, \\ 0, & \text{în caz contrar.} \end{cases}$$

$$\frac{x_1 + x_2 + \dots + x_6}{6}$$

$$\bar{x} = \dots = \begin{pmatrix} 2.5 \\ 2.33 \\ 1 \end{pmatrix}$$

$$\begin{aligned}
 &= \frac{5}{6} \cdot \left\| \begin{pmatrix} 1.5 \\ 1.2 \\ 0.4 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2.33 \\ 1 \end{pmatrix} \right\| + \frac{1}{6} \left\| \begin{pmatrix} 8 \\ 8 \\ 1 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2.33 \\ 1 \end{pmatrix} \right\| = \\
 &= \frac{5}{6} \left\| \begin{pmatrix} -0.9 \\ -1.13 \\ 0.6 \end{pmatrix} \right\| + \frac{1}{6} \left\| \begin{pmatrix} 5.5 \\ 5.67 \\ 3 \end{pmatrix} \right\| = \frac{5}{6} \cdot 2.27 + \frac{1}{6} \cdot 3.506 = ...
 \end{aligned}$$

45.

(Algoritmul  $K$ -means ca algoritm de optimizare:  
maximizarea aproximativă a distanțelor dintre centroizii clusterelor)

■ • ○ CMU, 2010 fall, Aarti Singh, HW3, pr. 5.2

În această problemă vom lucra cu o versiune a algoritmului  $K$ -means ușor modificată față de cea dată în enunțul problemei 12.

Fie  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  o mulțime de instanțe, iar  $K$  numărul de clustere cu care vom lucra. De data aceasta vom specifica asignările instanțelor la clustere folosind o *matrice-indicator*  $\gamma \in \{0,1\}^{n \times K}$ , cu  $\gamma_{ij} = 1$  dacă și numai dacă  $\mathbf{x}_i$  aparține clusterului  $j$ . Vom impune ca fiecare instanță să aparțină căte unui singur cluster, deci  $\sum_{j=1}^K \gamma_{ij} = 1$ .

După cum s-a arătat la problema 12, algoritmul  $K$ -means „estimează“ matricea  $\gamma$  făcând minimizarea criteriului (sau, a „măsurii de distorsiune“)  $J$ , pe care, folosind matricea  $\gamma$ , îl rescriem sub forma

$$J(\gamma, \mu_1, \mu_2, \dots, \mu_K) = \sum_{i=1}^n \sum_{j=1}^K \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2,$$

<sup>900</sup> Acestea sunt datele de la exercițiul 7.

<sup>901</sup> Pentru definiția riguroasă a acestui criteriu vedeți problema 12.

918

unde  $\|\cdot\|$  desemnează norma vectorială  $L_2$ . Concret, algoritmul  $K$ -means alternează „estimarea“ matricii  $\gamma$  cu re-calcularea centroizilor  $\mu_j$ .

Acestea fiind spuse, putem da acum noua versiune a algoritmului  $K$ -means:<sup>902</sup>

- Se inițializează în mod arbitrar centroizii  $\mu_1, \mu_2, \dots, \mu_K$  și se ia  $C = \{1, \dots, K\}$ .
- Atât timp cât valoarea lui  $J$  descrește în mod strict,<sup>903</sup> repetă:

Pasul 1: Calculează  $\gamma$  astfel:

$$\gamma_{ij} \leftarrow \begin{cases} 1, & \text{dacă } \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \in C, \\ 0, & \text{în caz contrar.} \end{cases}$$

În caz de egalitate, alege în mod arbitrar cărui cluster (dintre cele eligibile) să-i aparțină  $\mathbf{x}_i$ .

Pasul 2: Recalculează  $\mu_j$  folosind matricea  $\gamma$  actualizată:

Pentru fiecare  $j \in C$ , dacă  $\sum_{i=1}^n \gamma_{ij} > 0$ , asignează

$$\mu_j \leftarrow \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}.$$

Altfel, menține neschimbăt centroidul  $\mu_j$ .

Vom nota cu  $\bar{\mathbf{x}}$  media instanțelor date și vom considera următoarele trei cantități:<sup>904</sup>

$$\text{Variația totală: } T(X) = \frac{\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}{n}$$

$$\text{Variația intra-clustere: } W_j(X) = \frac{\sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2}{\sum_{i=1}^n \gamma_{ij}} \text{ pentru } j = 1, \dots, K$$

$$\text{Variația inter-clustere: } B(X) = \sum_{j=1}^K \left( \frac{\sum_{i=1}^n \gamma_{ij}}{n} \right) \|\mu_j - \bar{\mathbf{x}}\|^2.$$

a. Care este relația dintre aceste trei cantități?

*Observație:* Veți ține cont că această relație poate să conțină un termen suplimentar care nu este menționat mai sus.

b. Folosind relația stabilită la punctul a, arătați că putem interpreta algoritmul  $K$ -means ca tinzând să minimizeze (și anume descrescând, dar nu neapărat strict monoton) o medie ponderată a variației intra-clustere în timp ce el tinde să maximizeze (crescând) însă doar în mod *aproximativ* variația inter-clustere.

numărul, și dist. doar wt. instanțele  $\mathbf{x}_i \in C_j$ .

cardinalul fiecărui cluster

49.

(Algoritmul  $K$ -means: aplicare, folosind distanța euclidiană, respectiv distanța Manhattan / norma  $L_1$   
robustetea algoritmului  $K$ -means la prezența outlier-elor)

prelucrare de Liviu Ciortuz, după

• o CMU, 2010 fall, Aarti Singh, HW3, pr. 5.3

CMU, 2014 spring, B. Poczos, A. Singh, HW3, pr. 1.4

CMU, 2014 spring, Seyoung Kim, HW3, pr. 1.1

Fie setul date alăturat ( $X$ ), fiecare rând / linie reprezentând o instanță.

#### A. $K$ -means folosind distanța euclidiană

Aplicați algoritmul  $K$ -means pe acest set de date, folosind  $K = 3$  și distanța euclidiană. La pasul de inițializare nu veți seta pozițiile centroizilor ci, în schimb, veți inițializa clusterele, după cum urmează:

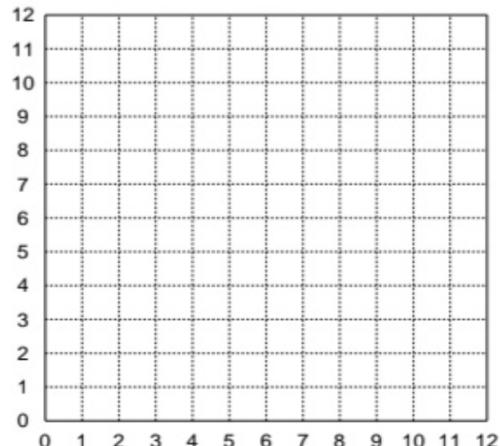
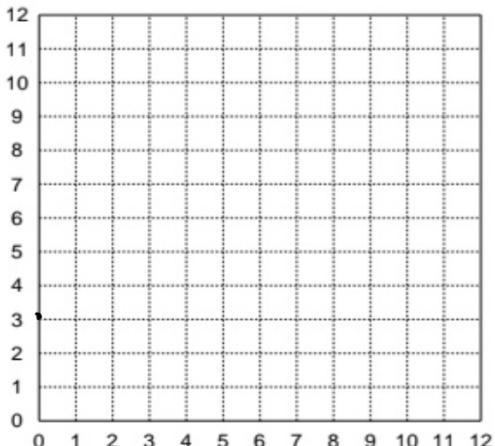
$C1 : \{A, B, F\}$ ,  $C2 : \{C, H, I\}$ ,  $C3 : \{D, E, G\}$ .

În consecință, veți inversa pașii 1 și 2 din ciclul iterativ al algoritmului  $K$ -means.

Folosiți gridurile de mai jos; puteți adăuga și altele decă veți considera că este necesar.

Puteți proceda fie în maniera analitică (calculând distanțele de la instanțe la centroizi), fie trasând mediatoarele segmentelor care unesc centroizii.

A	(1, 1)
B	(3, 3)
C	(6, 6)
D	(6, 12)
E	(9, 9)
F	(11, 11)
G	(0, 3)
H	(3, 0)
I	(9, 3)



#### B. $K$ -means folosind distanța Manhattan

În această parte a exercițiului vom folosi distanța Manhattan, desemnată prin  $\|\cdot\|_1$ , ceea ce înseamnă că vom defini [ca nouă „măsură de distorsiune“] funcția

$$J_1(\gamma, \mu_1, \mu_2, \dots, \mu_K) = \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} \|\mathbf{x}_i - \mu_k\|_1.$$

Vă reamintim faptul că distanța Manhattan ( $d_M$ ) dintre un vector  $x = (x_1, \dots, x_p)$  și un alt vector  $y = (y_1, \dots, y_p)$ , ambele fiind din  $\mathbb{R}^p$ , este definită astfel:

$$d_M = \sum_{i=1}^p |x_i - y_i|.$$

$$\text{If: } \mu_1^0 = \frac{A + B + F}{3} = \begin{bmatrix} 5 \\ 5 \end{bmatrix} \quad \mu_2^0 = \frac{D + E + G}{3} = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

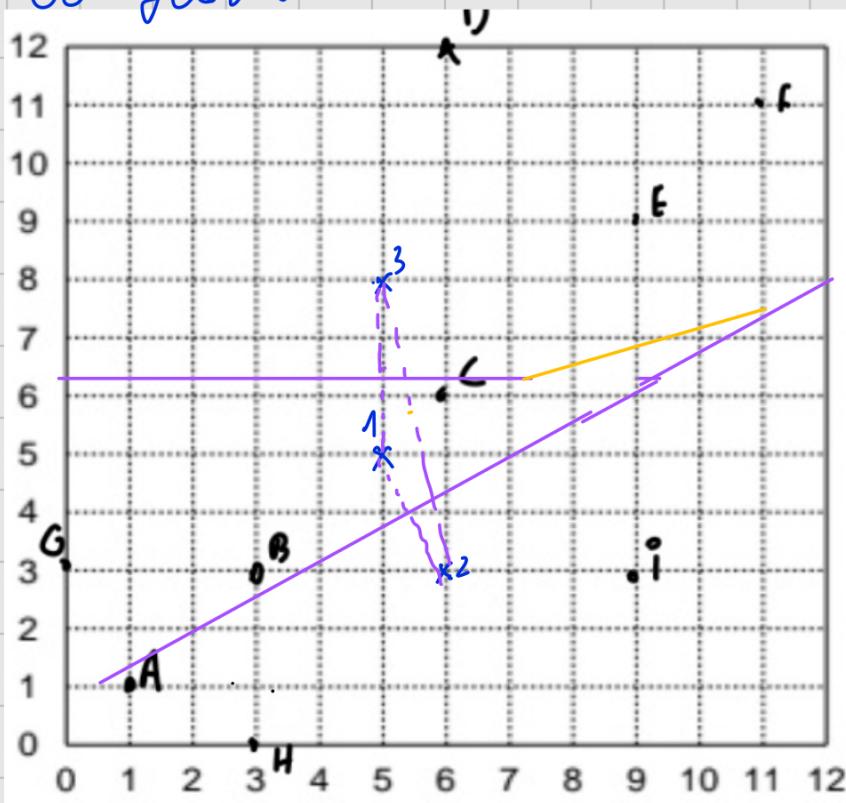
$$\mu_3^0 = \frac{C + H + I}{3} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$$

Var. analitici

$d(-, -)$	A	B	C	D	E	F	G	H	I
$\mu_1^0$									
$\mu_2^0$									
$\mu_3^0$									

$$||A - \mu_1^0|| = ||\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 5 \end{pmatrix}|| = \sqrt{32}$$

Loc. glom.



$$d_2(\mu_2^0, A) = || \begin{pmatrix} 6 \\ 3 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} || = \sqrt{29}$$

$$d_2(\mu_1^0, A) = \sqrt{32}$$

if 1

$$C_1' = \{B, C, G\}$$

$$C_2' = \{I, A, H\}$$

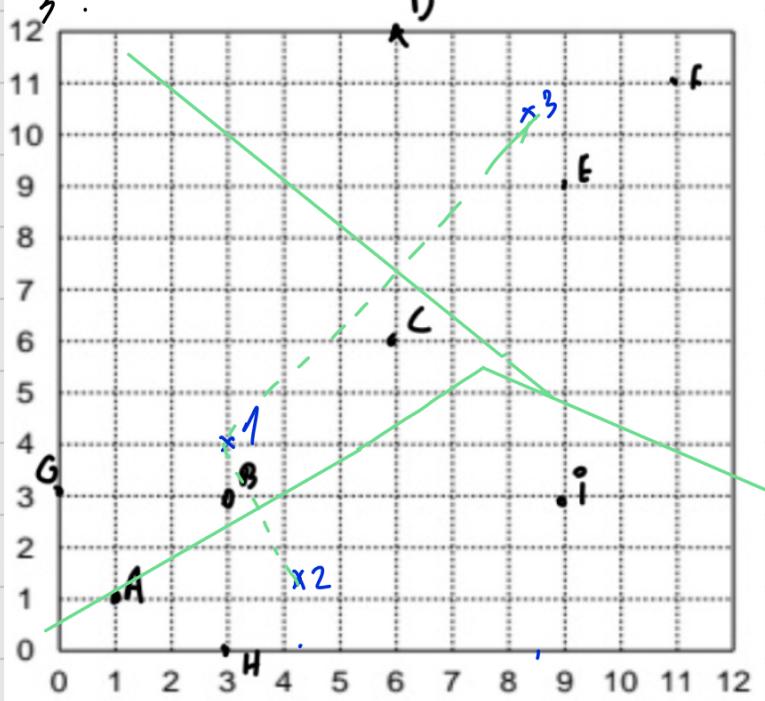
$$C_3' = \{E, F, D\}$$

$$\mu_1^1 = \frac{B+C+G}{3} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

$$\mu_2^1 = \frac{A+I+H}{3} = \begin{pmatrix} \frac{13}{3} \\ \frac{4}{3} \end{pmatrix}$$

$$\mu_3^1 = \frac{E+F+D}{3} = \begin{pmatrix} \frac{26}{3} \\ \frac{32}{3} \end{pmatrix}$$

if<sub>3</sub>:



$$C_1 = \{B, C, G\}$$

$$C_2 = \{A, H, i\}$$

$$C_3 = \{D, E, F\}$$

No option

dist. manhattan

B)

$$C_1^0 = \left( \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 11 \\ 11 \end{bmatrix} \right) \quad C_2^0 = \left( \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 9 \\ 3 \end{bmatrix} \right)$$

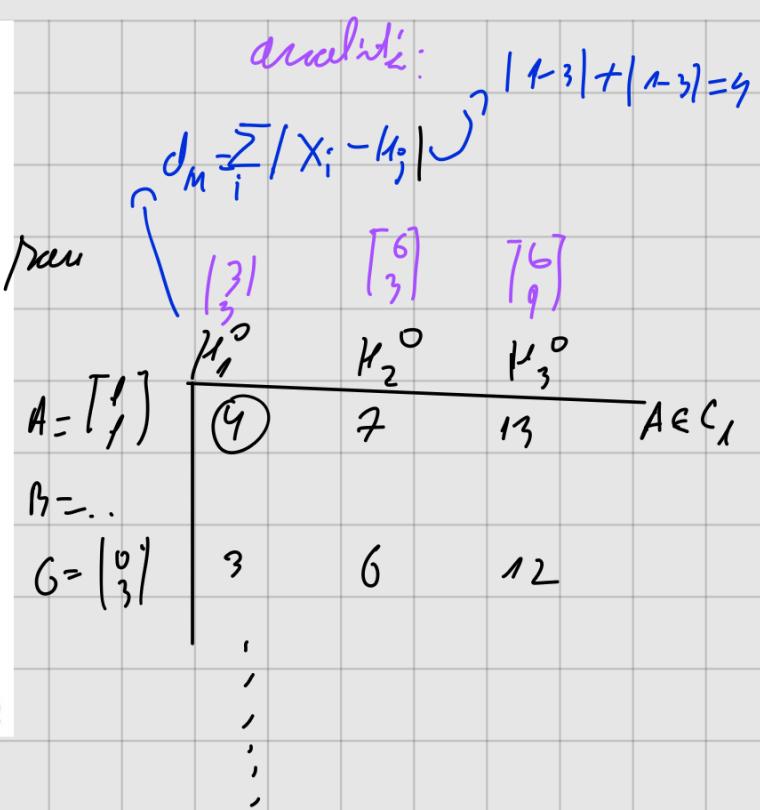
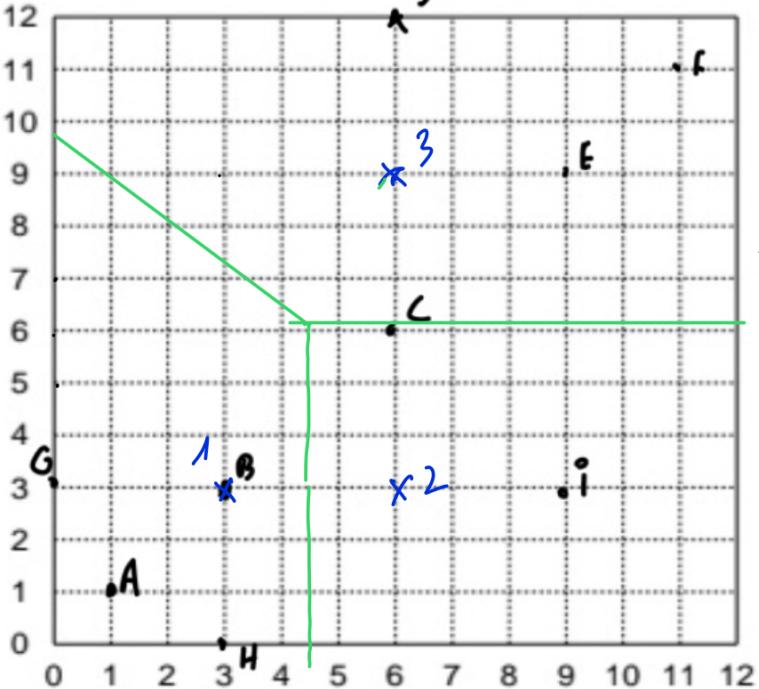
if: o :

$$C_3^0 = \left\{ \begin{bmatrix} 6 \\ 12 \end{bmatrix}, \begin{bmatrix} 9 \\ 9 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \end{bmatrix} \right\}$$

$$\mu_1^0 = \left\{ \begin{array}{l} \text{median}_x(1, 3, 11) = \boxed{3} \\ \text{median}_x(1, 3, 11) = \boxed{3} \end{array} \right.$$

$$\mu_2^0 = \left\{ \begin{array}{l} \text{median}_x(3, 6, 9) = \boxed{6} \\ \text{median}_x(0, 3, 9) = \boxed{3} \end{array} \right.$$

$$\mu_3^0 = \left\{ \begin{array}{l} \text{median}_x(0, 6, 9) = \boxed{6} \\ \text{median}_x(3, 0, 12) = \boxed{9} \end{array} \right.$$



if: 1:  $C_1^1 = \{A, B, 4, 6\}$

$$C_2^1 = \{9, C\}$$

$$C_3^1 = \{D, E, F\}$$

$\vdots$

