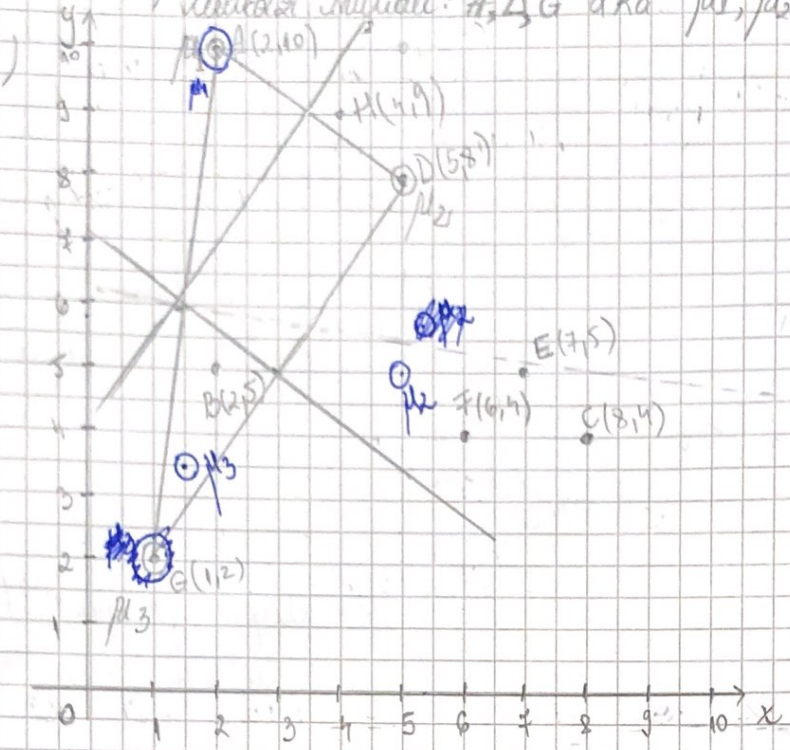
Tema:

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

prima iterație:

centroizi inițiali: A, D, G aka μ_1, μ_2, μ_3

a)



Se observă că punctele E, F, C sunt de aceeași parte a mediatoarei (GD) ca și centroidul μ_2. Iar punctul H este de aceeași parte a mediatoarei (A,D) ca și centroidul μ_2.

$\Rightarrow$ clusterii sunt: $C_1 = \{A(2,10)\}$; $C_2 = \{B(2,5), H(4,9), D(5,8), E(7,5), F(6,4), C(8,4)\}$; $C_3 = \{G(1,2), B(2,5)\}$

recalculăm centroizii:

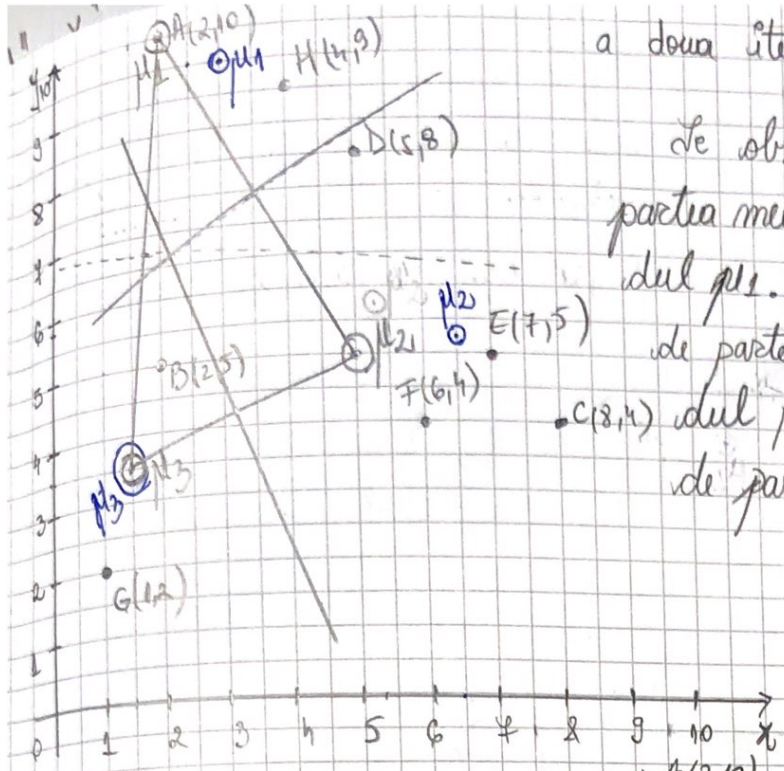
$\mu_1 = A(2,10)$

$\mu_3 = \left(\frac{x_G + x_B}{2}; \frac{y_G + y_B}{2}\right) = \left(\frac{1+2}{2}; \frac{2+5}{2}\right) = \left(\frac{3}{2}; \frac{7}{2}\right) = (1{,}5; 3{,}5)$

μ_2: $x_2 = \frac{x_H + x_D + x_E + x_F + x_C}{6} = \frac{4+5+7+6+8}{6} = \frac{30}{6} = 5$

$y_2 = \frac{y_H + y_D + y_E + y_F + y_C}{6} = \frac{9+8+5+4+4}{6} = \frac{30}{6} = 5$

~~$\Rightarrow \mu_2 = (16/3; 35/6)$~~ $\mu_2 = (5,5)$



a doua iterație:

Se observă că punct H este de partea mediatoarei (μ_1, μ_2) ca și centroidul μ_1. Punctele D, E, F, C sunt de partea mediatoarei ca și centroidul μ_2. Iar punctele B și G sunt de partea centroidului μ_3

$\Rightarrow$ clusterele sunt: $C_1 = \{A(2,10), H(4,9)\}$; $C_2 = \{D(5,8), E(7,5), C(8,4), F(6,4)\}$

$C_3 = \{B(2,5); G(1,2)\}$

Recalculăm centroizii:

$$\mu_1 = x_1 = \frac{x_A + x_H}{2} = \frac{2+4}{2} = 3 \quad y_1 = \frac{y_A + y_H}{2} = \frac{10+9}{2} = \frac{19}{2} = 9,5$$

$$\mu_1 = (3; 9,5)$$

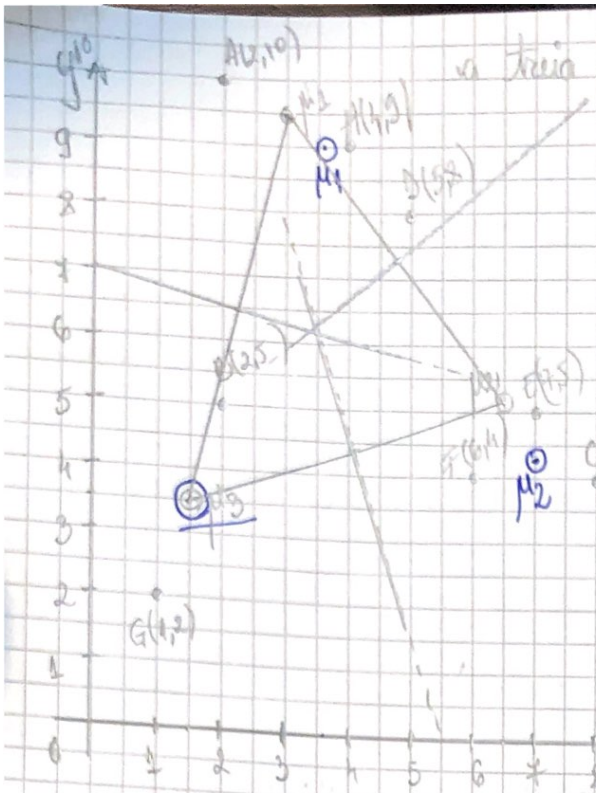
$$\mu_2 = x_2 = \frac{x_D + x_E + x_C + x_F}{4} = \frac{5+7+8+6}{4} = \frac{26}{4} = 6,5$$

$$y_2 = \frac{y_D + y_E + y_C + y_F}{4} = \frac{8+5+4+4}{4} = \frac{21}{4} = 5,25$$

$$\mu_2 = (6,5; 5,25)$$

$$\mu_3 = x_3 = \frac{x_B + x_G}{2} = \frac{2+1}{2} = \frac{3}{2} = 1,5 \; ; \; y_3 = \frac{y_B + y_G}{2} = \frac{5+2}{2} = \frac{7}{5} = 3,5$$

$$\mu_3 = (1,5; 3,5)$$



a treia iterație:

Se observă că punctele A, H, D sunt apropiate de centroidul μ_1.

Punctele F, E, C de centroidul μ_2

Iar punctele B, G de centroidul μ_3

$C_1 = \{A(2,10), H(4,9), D(5,8)\}$

$C_2 = \{E(7,5), C(8,4), F(6,4)\}$

$C_3 = \{B(2,5), G(1,2)\}$

Centroidul $\mu_3 = (1,5; 3,5)$ rămâne același.

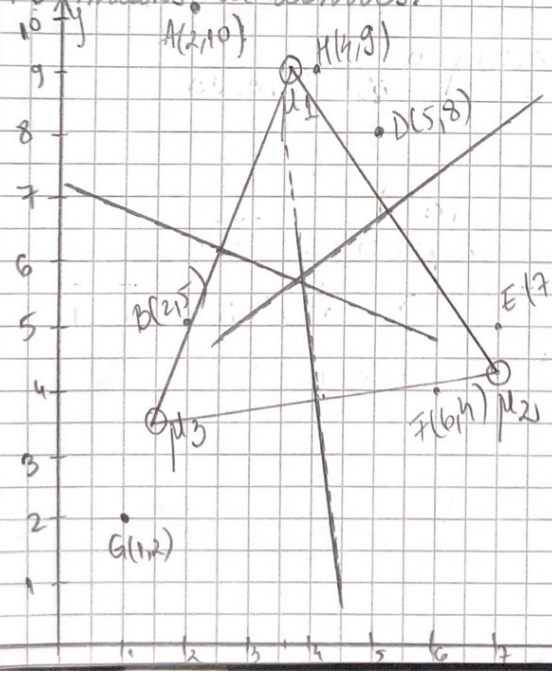
μ_1: $x_1 = \frac{2+4+5}{3} = \frac{11}{3} = 3,67$ $y_1 = \frac{10+9+8}{3} = \frac{27}{3} = 9$

$\mu_1 = (3,67; 9)$

μ_2: $x_2 = \frac{7+8+6}{3} = \frac{21}{3} = 7$ $y_2 = \frac{5+4+4}{3} = \frac{13}{3} = 4,3(3)$

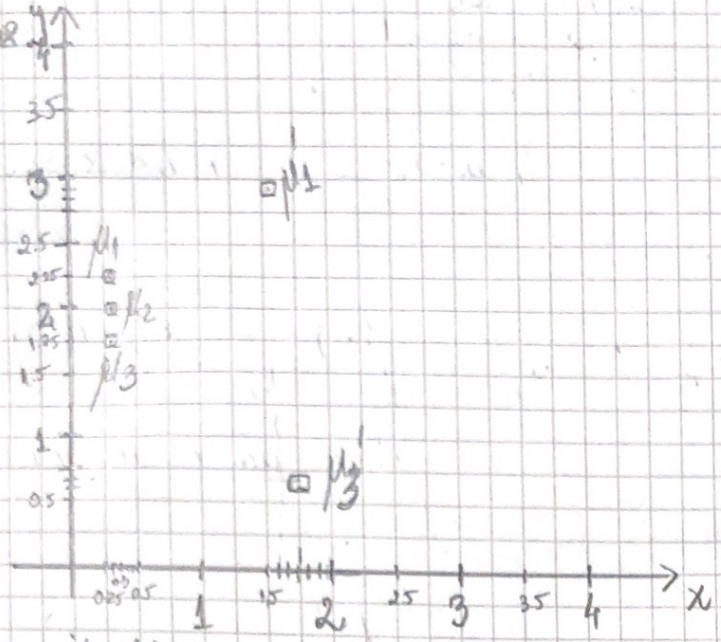
$\mu_2 = (7; 4,(3))$

Se poate observa că după a patra iterație algoritmul K-means a convers.



a patra iterație:

clusterele au rămas aceleași.



iteratia 1:

μ_3': $x_3 = \frac{0.25+0.50+0.75+1+1.4+0.5+0.25+0.75+0.35+0.85+3.25+3.50+3+3.25+3.45+3.75+3.25}{17}$

$= \frac{30,05}{17} = 1,768$ $\Rightarrow \boxed{\mu_3' = (1,8 ; 0,7)}$

$y_3 = \frac{0,5+0.35+0.50+0.35+0.70+0.85+2+2.5+0.85+1+0.85+0.85+0.85}{17}$

$= \frac{12,15}{17} = 0,715$

μ_1': $x_1 = \frac{20,47}{13} = 1,575 = 1,6$ $\Rightarrow \mu_1' = (1,6 ; 2,9)$

$y_1 = \frac{37,45}{13} = 2,88 = 2,9$

$\mu_2' = (0.30, 2.00)$

iteratia 2:

$d(\mu_2', A) = d^{(0,30,2)}(\mu_2', (0.25; 0.50)) = \sqrt{2.2525} < d^{(1,8;0,7)}(\mu_3', A) = \sqrt{2.44}$

$d(\mu_2', D) = d((0,30,2);(0.75;1.00)) = \sqrt{0.45^2+1^2} = \sqrt{1,20} < d(\mu_3', D) = \sqrt{3,94}$

$d(\mu_2', T) = d((0,30,2);(1,25)) = \sqrt{0,74}$ $d(\mu_1', T) = d((1.6;2.9);(1,25)) = \sqrt{0.52}$

Recalculăm centroizii: $\mu_1'' = (1,7 ; 2,9)$

$x_1'' = \frac{0.85+1.0+1.0+1.10+1.20+1.25+1.50+1.20+1.37+3.0+3.25+3.1}{12} = \frac{19.82}{12} = 1,65 = 1,7$

$y_2'' = \frac{3.35+3.0+2.5+3.20+2.40+2.50+2.50+2.75+2.75+3.25+3.0+3.50}{12} = \frac{34,7}{12} = 2,9$

39. $J(C^{(t)}, \mu^{(t)}) = \sum_{i=1}^{n} \left(x_i - \mu^{(t)}_{C^{(t)}(x_i)}\right)^2$

a) set de date: $-9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9$

$\mu_1^{(0)} = -20$

$\mu_2^{(0)} = -10$

dem că $\forall\, t=1 \Rightarrow J(C^{(t)}, \mu^{(t)}) \leq J(C^{(t-1)}, \mu^{(t-1)})$

Inegalitatea de mai sus este dată din două inegalități:

$$J(C^t, \mu^t) \overset{(1)}{\geqslant} J(C^t, \mu^{t+1}) \overset{(2)}{\geqslant} J(C^{t+1}, \mu^{t+1})$$

Inegalitatea 1: $J(C^t, \mu^t) \geqslant J(C^t, \mu^{t+1})$

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

not. $x_{i_1}, x_{i_2}, \ldots, x_{i_l}$ instanțele din comp. clusterului C_j^t, unde $l \overset{not.}{=} |C_j^t|$

$$\Rightarrow J(C_j^t, \mu_j^t) = \sum_{p=1}^{l} \left(x_{i_p} - \mu_j^t\right)^2$$

iar $J(C^t, \mu^t) = \sum_{j=1}^{k} J(C_j^t, \mu_j^t)$

Se consideră C_j^t fixat, iar μ_j^t variabil.

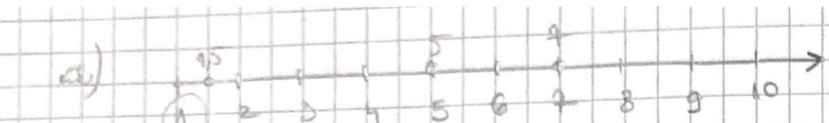
! Este normal ca pentru o viitoare iterație să avem un J mai mic sau egal deoarece distanțele de la centroid la puncte scad sau rămân la fel, ele se înmulțesc / împărțesc.

40. Dacă algoritmul K-means reîncărcează K-partiția respectivă, lucru care este posibil, înseamnă că algoritmul K-means a convers. Deoarece clusterele rămân aceleași și presupunem că și centroizii rămân aceiași.

41. Este soluție optimă pentru alegerea centroizilor, deoarece astfel, vom afla punctul în care putem plasa centroidul iar distanța până la valori să fie minimă. Așa cum se alg centroizii și în cazul algoritmului K-means.

48. Single-linkage va conduce cel mai probabil la formarea unor clustere asemănătoare cu cele obținute de K-means deoarece K-means se bazează pe distanța minimă dintre punct și centroid, iar single-linkage se bazează de asemea pe distanța minimă dintre punctele (x,y), cu $x \in X$ și $y \in Y$.

43 a)



Clusterizarea optimă pentru setul de date este:

$C_1 = \{1, 2\}$ $C_2 = \{5\}$ $C_3 = \{7\}$

Iar valoarea pt. funcția J este

$$J = \sum_{i=1}^{n} \min_{j \in \{1..K\}} \|x_i - \mu_j\|^2$$

$J = 0,5 + 0,5 + 0 + 0 = 1.$

b) $C_1 = \{1, 2\}$ $C_2 = \{5, 7\}$ $C_3 = \{\emptyset\}$

Valoarea din clusterul C_3 înseamnă că centroidul aflat la poziția pp. 10, are nicăieri element apropiat lui.

K-m nu poate îmbunătăți aceste clustere deoarece μ_3 rămâne în aceeași poziție, iar C_1 și C_2 vor converge tot spre centroizii lor.

c) Se poate observa că această proprietate este satisfăcută pe parcursul alg. K-means, deoarece centroizii vor fi plasați în centrul unor valori apropiate, creșcătoare, nu există cazuri în care

$C_1 = \{1, 2, 7\}$

$C_2 = \{5, 8\}$

Deoarece centroidul pt. C_2 va trage aproape de el și valoarea 7; iar pentru val $\{1, 2\}$ va rămâne centroidul pt. C_1.

d) Ideea mea ar fi să aleg din cei K centroizi, unul în limita inf, unul în limita

suprafață, după care luăm lungimea spațiului de de valori și o împărțim cla (K-2) pentru a împărți centroizi pe toată suprafața spațiului. După care fiecare valoare va fi asignată unui centroid, adică celui mai apropiat, iar centroizi vor fi recalculați pentru a fi în mijlocul valorilor.