Exercice K-means :

Appliquer l'algorithme du K-means sur le dataset suivant en utilisant k=2 :

Point	X	Y
A	1	1
B	1	2
C	2	1
D	2	2
E	8	8
F	8	9
G	9	8
H	9	9

Solution :

1ère itération :

K =2 donc on a 2 classes

On choisit les points A et H comme étant les centroïdes de la première et la deuxième classe :

$$g_{C_1} = A = (1,1)$$

$$g_{C_2} = H = (9,9)$$

Au départ, les classes ne contiennent que les points A et B (les centres de gravité) :

$$C_1 = \{ A \}$$

$$C_2 = \{ B \}$$

On calcule les distances de chaque point avec les deux centres de gravité en utilisant la distance euclidienne :

$$D_{(B,g_{C_1})}= \sqrt[2]{\left(x_B - x_{g_{C_1}}\right) + \left(y_B - y_{g_{C_1}}\right)} = \sqrt[2]{(1-1)^2 + (2-1)^2} = \sqrt[2]{0+1} = 1$$

$$D_{(B,g_{C_2})}= \sqrt[2]{\left(x_B - x_{g_{C_2}}\right) + \left(y_B - y_{g_{C_2}}\right)} = \sqrt[2]{(1-9)^2 + (2-9)^2} = \sqrt[2]{64+49} = 10.63$$

⇨ B plus proche de g_{C_1} donc B appartient à C_1

$$D_{(C,g_{C_1})}= \sqrt[2]{\left(x_C - x_{g_{C_1}}\right) + \left(y_C - y_{g_{C_1}}\right)} = \sqrt[2]{(2-1)^2 + (1-1)^2} = \sqrt[2]{1+0} = 1$$

$$D_{(C,g_{C_2})}= \sqrt[2]{\left(x_C - x_{g_{C_2}}\right) + \left(y_C - y_{g_{C_2}}\right)} = \sqrt[2]{(2-9)^2 + (1-9)^2} = \sqrt[2]{49+64} = 10.63$$

⇨ C plus proche de g_{C_1} donc C appartient à C_1

$D_{(D,g_{C_1})}= \sqrt[2]{\left(x_D - x_{g_{C_1}}\right) + \left(y_D - y_{g_{C_1}}\right)} = \sqrt[2]{(2-1)^2 + (2-1)^2} = \sqrt[2]{1+1} = 1.41$

$D_{(D,g_{C_2})}= \sqrt[2]{\left(x_D - x_{g_{C_2}}\right) + \left(y_D - y_{g_{C_2}}\right)} = \sqrt[2]{(2-9)^2 + (2-9)^2} = \sqrt[2]{49+49} = 9.89$

- ⇨ D plus proche de g_{C_1} donc D appartient à C_1

$D_{(E,g_{C_1})}= \sqrt[2]{\left(x_E - x_{g_{C_1}}\right) + \left(y_E - y_{g_{C_1}}\right)} = \sqrt[2]{(8-1)^2 + (8-1)^2} = \sqrt[2]{49+49} = 9.89$

$D_{(E,g_{C_2})}= \sqrt[2]{\left(x_E - x_{g_{C_2}}\right) + \left(y_E - y_{g_{C_2}}\right)} = \sqrt[2]{(8-9) + (8-9)} = \sqrt[2]{1+1} = 1.41$

- ⇨ E plus proche de g_{C_2} donc E appartient à C_2

$D_{(F,g_{C_1})}= \sqrt[2]{\left(x_F - x_{g_{C_1}}\right) + \left(y_F - y_{g_{C_1}}\right)} = \sqrt[2]{(8-1)^2 + (9-1)^2} = \sqrt[2]{49+64} = 10.63$

$D_{(F,g_{C_2})}= \sqrt[2]{\left(x_F - x_{g_{C_2}}\right) + \left(y_F - y_{g_{C_2}}\right)} = \sqrt[2]{(8-9)^2 + (9-9)^2} = \sqrt[2]{1+0} = 1$

- ⇨ F plus proche de g_{C_2} donc F appartient à C_2

$D_{(G,g_{C_1})}= \sqrt[2]{\left(x_G - x_{g_{C_1}}\right) + \left(y_G - y_{g_{C_1}}\right)} = \sqrt[2]{(9-1)^2 + (8-1)^2} = \sqrt[2]{64+49} = 3.87$

$D_{(G,g_{C_2})}= \sqrt[2]{\left(x_G - x_{g_{C_2}}\right) + \left(y_G - y_{g_{C_2}}\right)} = \sqrt[2]{(9-9) + (8-9)} = \sqrt[2]{0+1} = 1$

- ⇨ G plus proche de g_{C_2} donc G appartient à C_2

C1 = {A, B, C, D}

C2 = {E, F, G, H}

On recalcule les centroïdes des deux classes

$$g_{C_1} = \frac{x_A + x_B + x_C + x_D}{4}, \frac{y_A + y_B + y_C + y_D}{4}$$

$$g_{C_1} = \frac{1+1+2+2}{4}, \frac{1+2+1+2}{4}$$

$$\boldsymbol{g_{C_1} = \frac{3}{2}, \frac{3}{2}}$$

$$g_{C_2} = \frac{x_H + x_E + x_G + x_F}{4}, \frac{y_H + y_E + y_G + y_F}{4}$$

$$g_{C_2} = \frac{8+8+9+9}{4}, \frac{8+9+8+9}{4}$$

$$\boldsymbol{g_{C_2} = \frac{17}{2}, \frac{17}{2}}$$