TP : Clustering

Exercice I :

$X_1 = 1$; $X_2 = 2$; $X_3 = 9$; $X_4 = 12$; $X_5 = 20$

1) a- $K = 2$ $g_1 = 1$ et $g_2 = 20$ (K-means)

1ère itération :

	X_1 1	X_2 2	X_3 9	X_4 12	X_5 20
g_1 1 $(X_1 = 1)$	0	1	8	11	19
g_2 20 $(X_5 = 20)$	19	18	11	8	0

$C_1 = \{X_1, X_2, X_3\} \longrightarrow g_1 = \frac{X_1 + X_2 + X_3}{3} = \frac{1+2+9}{3} = 4$

$C_2 = \{X_4, X_5\} \longrightarrow g_2 = 16$

2ème Itération :

	X_1 1	X_2 2	X_3 9	X_4 12	X_5 20
$g_1 = 4$	3	2	5	8	16
$g_2 = 16$	15	14	7	4	4

$C_1 = \{X_1, X_2, X_3\}$

$C_2 = \{X_4, X_5\}$

- les classes n'ont pas changer, alors on arrête l'algorithme.

- Centre de gravité $g = \frac{1+2+9+12+20}{5} = 8,8$

- Inertie totale : $\frac{1}{5} \sum_{i=1}^{5} d^2(x_i, g)$

$$= \frac{1}{5}\left[(\underbrace{1}_{x_1} - \underbrace{8,8}_{g})^2 + (\underbrace{2}_{x_2} - \underbrace{8,8}_{g})^2 + (\underbrace{9}_{x_3} - \underbrace{8,8}_{g})^2 + (12-8,8)^2 + (20-8,8)^2\right]$$

$= 48,56$

on a 2 classes

- Inertie Inter classe : $\frac{1}{5} \sum_{k=1}^{2} n_k \cdot d^2(g_k, g)$

$$= \frac{1}{5}\left[\underbrace{3}_{\text{eff du classe } C_1} \times (\underbrace{4}_{g_1} - \underbrace{8,8}_{g})^2 + \underbrace{2}_{\text{eff class } C_2} \times (\underbrace{16}_{g_2} - \underbrace{8,8}_{g})^2\right] = 34,56$$

- Inertie Intra-classe = Inertie totale − Inertie inter-classe

$= 48,56 - 34,56 = 14.$

- pourcentage d'inertie expliqué par les classes :

$$R^2 = \frac{\text{Inertie Inter-classe}}{I_{\text{total}}} = \frac{34,56}{48,56} = 0.71$$

b) 1ère Itération :

	X_1	X_2	X_3	X_4	X_5
	1	2	9	12	20
$g_1 = 2$	1	0	7	10	18
$g_2 = 9$	8	7	0	3	11

$C_1 = \{X_1, X_2\} \rightarrow g_1 = \frac{1+2}{2} = 1.5$

$C_2 = \{X_3, X_4, X_5\} \rightarrow g_2 = \frac{9+12+20}{3} = 13,66$

2ème Itération :

	X_1	X_2	X_3	X_4	X_5
	1	2	9	12	20
$g_1 = 1,5$	0.5	0.5	7.5	10.5	18.5
$g_2 = 13,66$	12,66	11,66	4,66	1,66	6,34

- $C_1 = \{X_1, X_2\}$
- $C_2 = \{X_3, X_4, X_5\}$; Les classes n'ont pas changé, alors on arrête l'algorithme.

- Centre de gravité : $g = 8.8$.
- Inertie inter_classes : $I_{inter} = \frac{1}{5}\left[2 \times (1,5 - 8,8)^2 + 3 \times (13,66 - 8,8)^2\right]$
$= 35,48$

- Inertie total $= 48,56$

$$R^2 = \frac{35,48}{48,56} = 0,73$$

c) $K=3$; $g_1 = 1$; $g_2 = 9$; $g_3 = 12$

1er itération :

	X_1	X_2	X_3	X_4	X_5
	1	2	9	12	20
$g_1=1$	0	1	8	11	19
$g_2=9$	8	7	0	3	11
$g_3=12$	11	10	3	0	8

$C1 = \{X_1, X_2\} \rightarrow g_1 = \frac{1+2}{2} = 1.5$

$C_2 = \{X_3\} \rightarrow g_2 = 9$

$C_3 = \{X_4, X_5\} \rightarrow g_3 = \frac{12+20}{2} = 16$

2e itération :

	X_1	X_2	X_3	X_4	X_5
	1	2	9	12	20
$g_1=1.5$	0.5	0.5	7.5	10.5	18.5
$g_2=9$	8	7	0	3	11
$g_3=16$	15	14	7	4	4

$C_1 = \{X_1, X_2\}$ $\rightarrow g_1 = 1{,}5$

$C_2 = \{X_3, X_4\}$ $\rightarrow g_2 = 10.5$

$C_3 = \{X_5\}$ $\rightarrow g_3 = 20$

3ème Itération :

	X_1 1	X_2 2	X_3 9	X_4 12	X_5 20
			7.5	10.5	18.5
$g_1 = 1.5$	0.5	0.5			
			1.5	1.5	9.5
$g_2 = 10.5$	8.5	8.5			
$g_3 = 20$	19	18	11	8	0

$C_1 = \{X_1, X_2\}$

$C_2 = \{X_3, X_4\}$

$C_3 \{ X5 \}$; les classes ne pas changé, donc on arrête l'algorithme.

- $I_{inter} = 1/5 \left[2 \times (1{,}5 - 8{,}8)^2 + 2(10{,}5 - 8{,}8)^2 + 1.(20 - 8{,}8)^2 \right]$
  $= 47{,}56.$

$$R^2 = \frac{I_{inter}}{I_{total}} = \frac{47{,}56}{48{,}56} = 0.98$$

d) On peut utiliser l'inertie inter-classes pour déterminer le meilleur parmi deux partitions qui ont un nombre de classes. plus ce inertie est grande plus le poucentage d'inertie expliqué par la partition obtenue est grande. Ici, le meilleur partition entre "a" et "b" est celle du dernière partie. (R^2 plus élevée que 1°).

⚠ Le partition du "c" a le R^2 le plus élevée, mais c'est mieux d'avoir un petit nb de classes.

2) Méthode CAH. en utilisant distance minimale :

. 1ère itération : On considère chaque point X_i comme un classe et on calcule distance entre chaque deux points (classes)

$C_1 = \{X_1\}$

$C_2 = \{X_2\}$

$C_3 = \{X_3\}$

$C_4 = \{X_4\}$

$C_5 = \{X_4\}$

	C_1 $\{1\}$ X_1	C_2 $\{2\}$ X_2	C_3 $\{9\}$ X_3	C_4 $\{12\}$ X_4	C_5 $\{20\}$ X_5
C_1 $\{1\}$	0	1	8	11	19
C_2 $\{2\}$	1	0	7	10	18
C_3 $\{9\}$	8	7	0	3	11
C_4 $\{12\}$	11	10	3	0	8
C_5 $\{20\}$	19	18	11	8	0

Les classes les plus proches sont (C_1) et (C_2), Ont les regroupe alors.

$C_1 = \{X_1, X_2\}$

$C_2 = \{X_3\}$

$C_3 = \{X_4\}$

$C_4 = \{X_5\}$

2ème itération : On regroupe les classes les plus proches.

	C_1 $\{1,2\}$	C_2 $\{9\}$	C_3 $\{12\}$	C_4 $\{20\}$
C_1 $\{1,2\}$*	0	7	10	18
C_2 $\{9\}$	7	0	3	11
C_3 $\{12\}$	10	3	0	8
C_4 $\{20\}$	18	11	8	0

(*) on prend $(9-2)=7$ et pas $(9-1)=8$ car on veut la distance minimale

Les classes les plus proches sont : C_2 et C_3 ; ont les groupent

on regroupe X_3 avec X_4

donc $C_1 = \{X_1, X_2\}$

$C_2 = \{X_4, X_3\}$

$C_3 = \{X_5\}$

3ème itération

	C_1 $\{1,2\}$	C_2 $\{9,12\}$*	C_3 $\{20\}$
C_1 $\{1,2\}$	0	7	18
C_2 $\{9,12\}$	7	0	8
C_3 $\{20\}$	18	8	0

pour calculer distance entre $\{9,12\}$ et $\{1,2\}$ et prend la combinaison qui va nous donner le plus petit distance.
exp: $9-2=7$

Les classes les plus proches sont : C_2 avec C_1 ont les regroupent ; $C_1 = \{X_1, X_2, X_3, X_4\}$

$C_2 = \{X_5\}$

. 4ème Itération:

C1 {1,2,9,12}

C2 {20}

	C1 {1,2,9,12}	C2 {20}
	0	8
	8	0

On groupe à la fin tous les classes ensemble.

. Dendogramme: