Institut Supérieur d'Informatique et de Multimédia de Sfax

Auditoire: 2ème année LSI-ADBD

Chapitre 4:

Méthodes de regroupement : clustering

Souhir BOUAZIZ AFFES
souhir.bouaziz@isims.usf.tn

Amal ABBES
amal.abbes@isims.usf.tn

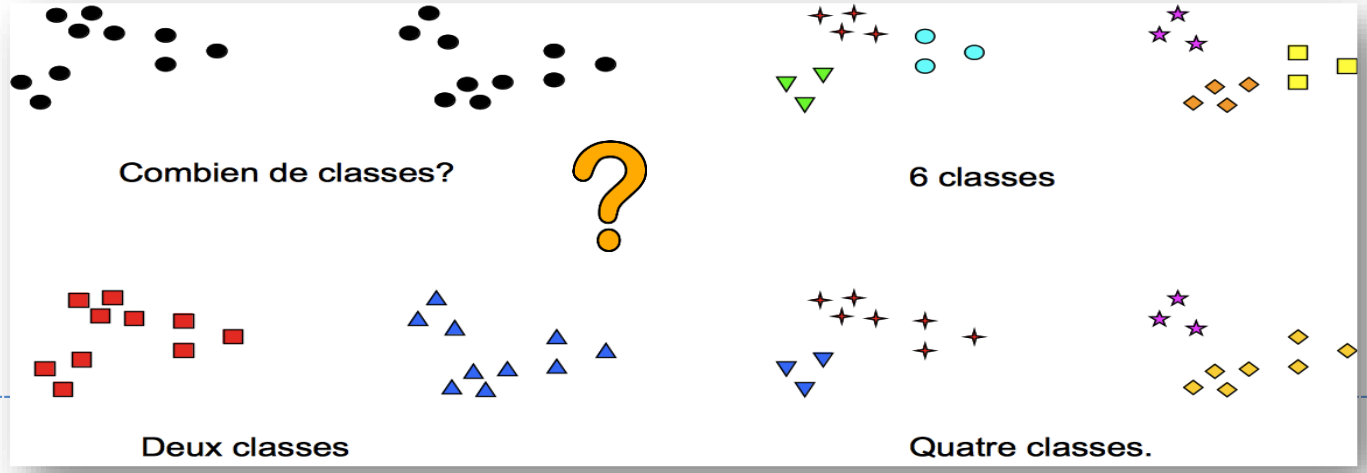
Plan

- Introduction
- Notion de ressemblance
- Normalisation des données
- Méthodes de regroupement
- Regroupement par ressemblance
  - Méthode des K-means
- Classification hiérarchique
  - Méthode de Classification Ascendante Hiérarchique : CAH
- Exercice: CHA & K-means

Introduction

- **Regroupement (Clustering):** construire une collection d'individus ou d'objets
  - Similaires au sein d'un même groupe
  - Dissimilaires quand ils appartiennent à des groupes différents

- Le regroupement est de la **classification non supervisée** :
  - Elle vise à identifier des ensembles d'éléments qui partagent certaines similarités (**ressemblance**)
  - Elle ne se base pas sur des classes prédéfinies: ***classification automatique***

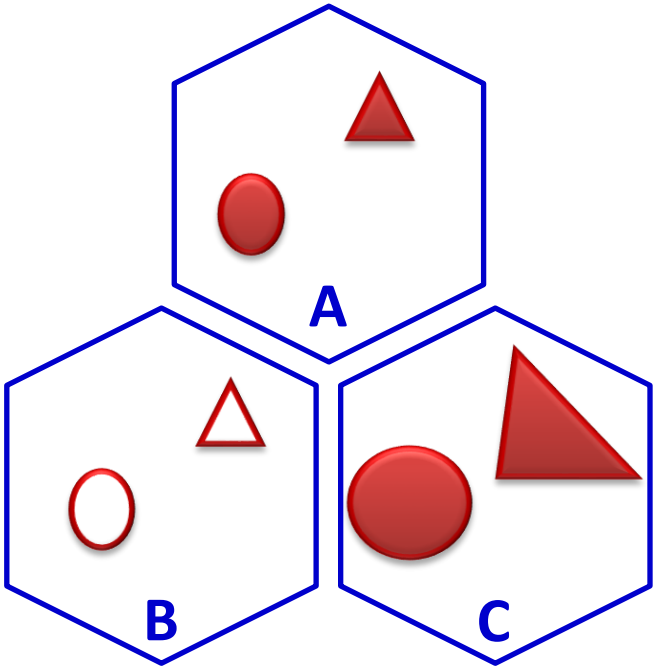


Introduction

- **Exemples d'application de la classification automatique:**
  - **Gestion (*Marketing*) :** découper le marché en sous-ensembles dont les éléments réagissent de façon similaire aux variations des variables du marché
  - **CRM (*Customer Relationship Management*) :** identifier des groupes d'individus ayant un comportement homogène vis-à-vis de:
    - la consommation de différents produits,
    - la consommation de différentes marques ou variétés
    - l'attitude par rapport à un produit,
    - …
  - **Réseaux sociaux:** Identifier les communautés (ensemble de nœuds entre lesquels les interactions sont fréquentes) dans un réseau
    - Amis, collègues, …
    - Personnes avec des intérêts similaires,
    - Pages web avec un même contenu, etc.

Notion de ressemblance

- Chercher les ressemblances entre ces trois ensembles



- L'ensemble A est-il plus proche à l'ensemble B ou C?

Notion de ressemblance

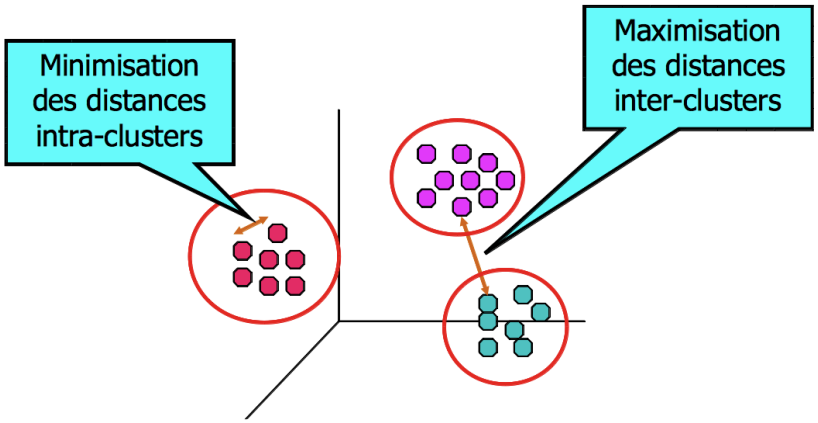
- La notion de ressemblance entre objets ne peut pas être évaluée objectivement car elle dépend du:
  - contexte de l'application considérée
  - but à atteindre
- La ressemblance entre deux objets est l'ensemble des propriétés communes
- Pour notre exemple:
  - Chaque forme géométrique peut être représentée par:
    - Forme
    - Couleur
    - Taille
    - Position

Chaque objet peut être donc représentée sous forme d'un vecteur

  - Les caractéristiques les plus pertinentes peuvent avoir **des poids** plus importants dans le calcul de distance

Notion de ressemblance

- Une bonne méthode de regroupement permet de garantir :
  - Une grande similarité intra-groupe
  - Une faible similarité inter-groupe



- Pour définir l'homogénéité d'un groupe d'observations, il est nécessaire de mesurer la ressemblance entre deux observations.

Notion de ressemblance

- Les mesures de distances sont utilisées pour mesurer la ressemblance entre objets
- Soit **d** une fonction: **d** est une distance ***ssi*** elle respecte les propriétés suivantes:
  - **La propriété de positivité:** $d(x,y) \geq 0$
  - **La propriété de séparation:** $d(x,y) = 0 \text{ si } x = y$
  - **La propriété de symétrie:** $d(x,y) = d(y,x)$
  - **L'inégalité triangulaire:** $d(x,z) \leq d(x,y) + d(y,z)$

Notion de ressemblance

- Les distances les plus connues sont celles de **Minkowski**:
  - soit deux objets $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ et $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$,
  - soit q est un entier t.q. $q \geq 0$

$$d(x_i, x_j) = \sqrt[q]{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q}$$

Matrice de données

n: nombre d'individus de BD

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

Matrice de dissimilarité

$$\begin{bmatrix} 0 & & & & \\ d(x_2, x_1) & 0 & & & \\ d(x_3, x_1) & d(x_3, x_2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(x_n, x_1) & d(x_n, x_2) & \dots\dots\dots\dots & d(x_n, x_{n-1}) & 0 \end{bmatrix}$$

Notion de ressemblance

- Les distances les plus utilisées:
  - $q = 1$: **Distance de Manhattan**

$$d(x_i, x_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

  - $q = 2$: **Distance Euclidienne**

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

  - $q \to \infty$: **Distance du maximum (ou distance de Tchebychev)**

$$d(x_i, x_j) = \max_{1 \le k \le p} |x_{ik} - x_{jk}|$$

  - **Distance de Canberra**

$$d(x_i, x_j) = \sum_{i=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik} + x_{jk}|}$$

Notion de ressemblance

- Afin de pouvoir calculer ces distances, il faut que les différents attributs soient de **nature numérique**

Personne	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

Que faire avec d'autres types des données : Ex. Binaires?

➔ une mesure de distance spécifique doit être définie

Normalisation des données

- **Variables quantitatives:**
  - La normalisation est généralement nécessaire lorsque les données sont réparties sur des échelles différentes
  - Normalisation des données: consiste à transformer les caractéristiques pour qu'elles soient sur une échelle similaire.

Personne	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

  - Il existe différentes techniques pour normaliser les données :
    - Normalisation Min-Max
    - Mise à l'échelle décimale
    - Normalisation du Z-score

Normalisation des données

- **Variables quantitatives …**
  - **Normalisation Min-Max :**

$$z_{ij} = \frac{x_{ij} - \min(X_{(j)})}{\max(X_{(j)}) - \min(X_{(j)})}\left(new\_\max(X_{(j)}) - new\_min(X_{(j)})\right) + new\_min(X_{(j)})$$

- $\boldsymbol{min(X_{(j)}), max(X_{(j)})}$ **:** les valeurs absolues minimale et maximale de $X_{(j)}$ respectivement
- $\boldsymbol{new\_max(X_{(j)}), new\_min(X_{(j)})}$: les valeurs max et min de champs de travail respectivement
- **Exemple:**

Notes
8
10
15
20

Notes	Notes après normalisation Min-Max
8	0
10	0.17
15	0.58
20	1

Normalisation des données

- **Variables quantitatives …**
  - **Mise à l'échelle décimale :**

$$z_{ij} = \frac{x_{ij}}{10^k}$$

  - $\boldsymbol{k}$: le plus petit nombre entier tel que $\max(|z_{ij}|) < 1$
  - **Exemples:**

MPC	Formule	MPC après mise à l'echelle décimale
2	2/10	0.2
3	3/10	0.3

Prime de rendement	Formule	Prime après mise à l'echelle décimale
400	400 / 1000	0.4
310	310 / 1000	0.31

Salaire	Formule	Salaire après mise à l'echelle décimale
40 000	40 000 / 100 000	0.4
31 000	31 000 / 100 000	0.31

Normalisation des données

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

- **Variables quantitatives …**
  - **Normalisation du z-score (ou *standardisation*):**

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

- z_{ij}: la valeur normalisée de x_{ij}: c-à-d sa distance à la moyenne exprimée en écart-type
- μ_j: la moyenne de $X_{(j)}$: $\mu_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$
- σ_j: l'écart type **absolu** de $X_{(j)}$: $\sigma_j = \frac{1}{n}\sum_{k=1}^{n}\left|x_{kj} - \mu_j\right|$

L'utilisation de l'écart absolu est plus robuste que celle de l'écart type

- **Exemple:**

Notes
8
10
15
20

Notes	Notes après normalisation du z-score
8	-1.24
10	-0.76
15	0.41
20	1.59

Normalisation des données

- **Variables quantitatives : Exemple:**

Personne	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

Calcul distance de Manhattan

d(P1,P2)=120

d(P1,P3)=132

=> P1 ressemble plus à P2 qu'à P3 ☹

Normalisation Min-Max :

$$\min(Age) = 50,\ \max(Age) = 70,\ \text{new\_min}(Age) = 0,\ \text{new\_max}(Age) = 1$$

$$\min(Salaire) = 11000,\ \max(Salaire) = 11122,\ \text{new\_min}(Salaire) = 0,\ \text{new\_max}(Salaire) = 1$$

Mise à l'échelle décimale :

$$k_{Age} = 2\ (10^k = 100)$$

$$k_{Salaire} = 5\ (10^k = 100000)$$

Normalisation du Z-Score :

$$\mu_{Age} = 60\ \ \sigma_{Age} = 5$$

$$\mu_{Salaire} = 11074\ \ \sigma_{Salaire} = 37$$

Normalisation des données

- **Variables quantitatives : Exemple:**

Personne	Age après normalisation Min-Max	Salaire après normalisation Min-Max
P1	0	0
P2	1	0,82
P3	0,5	1
P4	0,5	0,61

Personne	Age après mise à l'échelle décimale	Salaire après mise à l'échelle décimale
P1	0,5	0,11
P2	0,7	0,111
P3	0,6	0,11122
P4	0,6	0,11074

Personne	Age après normalisation du Z-score	Salaire après normalisation du Z-score
P1	-2	-2
P2	2	0,703
P3	0	1,297
P4	0	0

Calcul distance de Manhattan

d(P1,P2)=1,82

d(P1,P3)=1,5

d(P1,P2)=0,201

d(P1,P3)=0,10122

d(P1,P2)=6,703

d(P1,P3)=5,297

=> P1 ressemble plus à P3 qu'à P2 ☺

Normalisation des données

- **Variables binaires:** Il faut tout d'abord tracer la table de contingence (table de dissimilarité) de ces données

x_i \ x_j	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

- a = nombre de positions où x_i et x_j sont à 1
- d = nombre de positions où x_i et x_j sont à 0
- c = nombre de positions où x_i est à 0 et x_j est à 1
- b = nombre de positions où x_i est à 1 et x_j est à 0

Normalisation des données

- **Variables binaires …**
  - **Distances utilisées:**
    - **Le coefficient de correspondance simple:** dans le cas des attributs binaires **symétriques** (leurs deux états ont la même importance (poids))

$$d_{cs}(x_i, x_j) = \frac{b + c}{a + b + c + d}$$

    - **Le coefficient de Jaccard:** dans le cas des attributs binaires **asymétriques** (leurs deux états n'ont pas la même importance (fréquence))

$$d_{jc}(x_i, x_j) = \frac{b + c}{a + b + c}$$

  - **Exemple:** x_1 =(1,1,0,1,0) et x_2 =(1,0,0,0,1)

x_1	1	1	0	1	0
x_2	1	0	0	0	1

  - ➔ $a = 1, b = 2, c = 1, d = 1$
  - ➔ $d_{cs}(x_1, x_2) = {}^3/_5$ et $d_{jc}(x_1, x_2) = {}^3/_4$

Normalisation des données

- **Variables binaires …**
  - **Exemple: Quels sont les deux patients qui atteints de la même maladie?**

Nom	Sexe	Fièvre	Tousse	Test-1	Test-2	Test-3	Test-4
Jacques	M	O	N	P	N	N	N
Marie	F	O	N	P	N	P	N
Jean	M	O	P	N	N	N	N

- Sexe est un attribut symétrique et les autres attributs sont asymétriques
- O et P ≡ 1, N ≡ 0, la distance n'est mesurée que sur les asymétriques

$$d(Jacques, Marie) = \frac{0+1}{2+0+1} = 0{,}33$$

$$d(Jacques, Jean) = \frac{1+1}{1+1+1} = 0{,}67$$

$$d(Jean, Marie) = \frac{1+2}{1+1+2} = 0{,}75$$

➔ Jacques et Marie sont atteints de la même maladie

Normalisation des données

- **Variables de types mixtes: le cas le plus probable**
  - **But:** essayer de normaliser toutes les valeurs entre 0 et 1
    - Les **variables binaires** restent intactes
    - Pour les **variables quantitatives**: Normalisation Min-Max :

$$z_{ij} = \frac{x_{ij} - \min(X_{(j)})}{\max(X_{(j)}) - \min(X_{(j)})}\left(new\_\max(X_{(j)}) - new\_\min(X_{(j)})\right) + new\_\min(X_{(j)})$$

Avec:

- $new\_\max(X_{(j)}) = 1$
- $new\_\min(X_{(j)}) = 0$

Normalisation des données

- **Variables de types mixtes …**
  - **Exemple:** **Quels sont les voisins les plus proches?**

Personne	Age	Maison	Salaire
P1	30	1	1000
P2	40	0	2200
P3	45	1	4000

Calcul distance Euclidienne

$$d(\mathrm{P}_1, \mathrm{P}_2) = \sqrt{\left(\frac{10}{15}\right)^2 + 1^2 + \left(\frac{1200}{3000}\right)^2} = 1{,}27$$

$$d(\mathrm{P}_1, \mathrm{P}_3) = \sqrt{\left(\frac{15}{15}\right)^2 + 0^2 + \left(\frac{3000}{3000}\right)^2} = 1{,}41$$

$$d(\mathrm{P}_2, \mathrm{P}_3) = \sqrt{\left(\frac{5}{15}\right)^2 + 1^2 + \left(\frac{1800}{3000}\right)^2} = 1{,}21$$

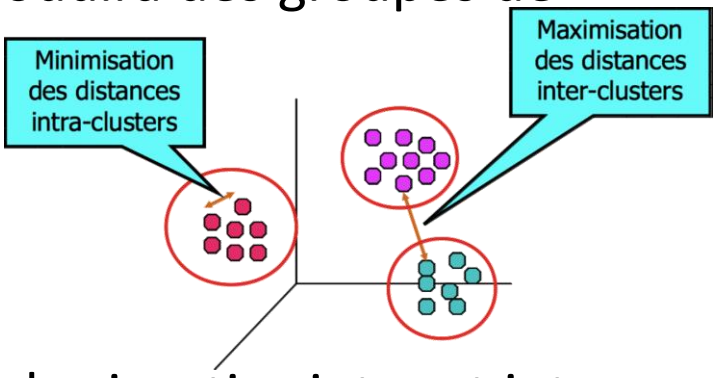
→ P_2 et P_3 sont les plus proches

Méthodes de regroupement

- **Méthodes de partitionnement (ou regroupement par ressemblance)**
  - deux classes sont toujours disjointes
  - **Principe:** partitionnement des objets et évaluation des partitions
  - Ex. méthode K-means
- **Méthodes hiérarchiques**
  - deux classes sont disjointes ou l'une contient l'autre
  - **Principe:** décomposition hiérarchique d'ensembles d'objets
  - Ex. classification hiérarchique
- **méthodes basées sur la densité**
  - **Principe:** se base sur une fonction de densité ou de connectivité
- **Méthodes basées sur la grille**
  - **Principe:** se base sur une structure de granularité à plusieurs niveaux

Regroupement par ressemblance

- **Principe:** Utiliser une mesure de distance pour trouver un partitionnement de la base de données D contenant n objets en k groupes
- Une bonne méthode de regroupement produira des groupes de **bonne qualité** avec :
  - Une grande similarité intra-cluster
  - Une faible similarité inter-cluster



- Ces similarités peuvent être mesurées par les inerties intra et inter cluster
- Parmi les critères d'évaluation d'une méthode de regroupement:
  - Stabilisation des **centres** des clusters
  - Stabilisation de l'**inertie totale** de la population

Regroupement par ressemblance

- **Inertie totale de la population:** somme de l'inertie intra- cluster $\mathbf{I_{intra}}$ et de l'inertie inter-cluster $\mathbf{I_{inter}}$: *Théorème d'Huygens*:

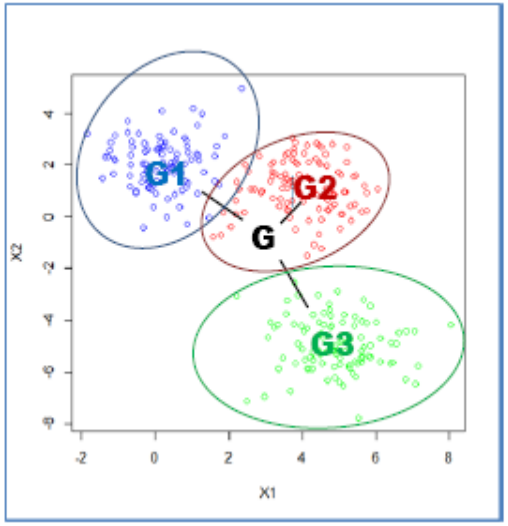
$$\boldsymbol{I_{total} = I_{intra} + I_{inter}}$$

  - **Inertie intra-cluster $\mathbf{I_{intra}}$:** dispersion à l'intérieur de chaque groupe: ***Indicateur de compacité des clusters***

$$\boldsymbol{I_{intra} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_j, G_i)}$$

Avec: n_i: nombre de points du cluster C_i

G_i: centre de gravité du cluster C_i



  - **Inertie inter-cluster $\mathbf{I_{inter}}$:** dispersion des barycentres conditionnels autour du barycentre global: ***Indicateur de séparabilité des clusters***

$$\boldsymbol{I_{inter} = \frac{1}{k} \sum_{i=1}^{k} d^2(G_i, G)}$$ Avec: G: centre de gravité global

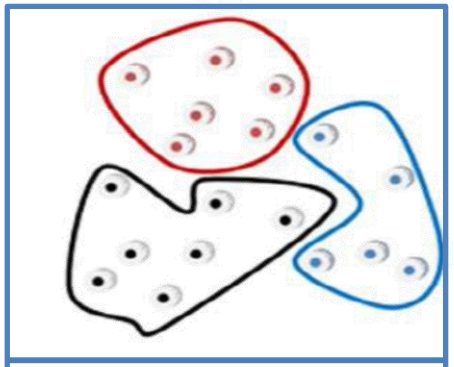
- Comparaison de deux partitions en $\boldsymbol{k}$ clusters : La meilleure est celle qui a l'inertie $\mathbf{I_{intra}}$ la plus faible (ou l'inertie $\mathbf{I_{inter}}$ la plus forte).

Regroupement par ressemblance

- **Méthodes:**
  - Etant donné k, trouvé une partition en k clusters qui optimisent le critère de partitionnement
    - **Optimum global :** traiter toutes les partitions exhaustivement
    - **Heuristique :** k-means ou k-médoïdes
      - **k-means** (MacQueen'67)**:** chaque cluster est représenté par son centre
      - **k-médoïdes ou PAM (partition around medoids**) (Kaufman & Rousseeuw'87)**:** chaque cluster est représenté par un des objets du cluster

Méthode des k-means

- **But:** rechercher une partition des données (uni- ou multidimensionnelles) en k clusters ou groupes
  - k représente le nombre de clusters que l'algorithme doit former à partir des propriétés des échantillons.
  - k peut être supposé fixe (donné par l'utilisateur) ou fixé par la nature du problème à traiter.
    - **Exemple:** si l'on s'intéresse à classer des images de chiffres manuscrits (nbre de classes = 10 : 0, …, 9)

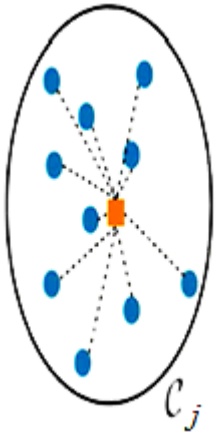


Clustering sur un Ens. de Points 2D, en 3 Clusters

- **Principe:** optimiser l'inertie intra-cluster (compacité de chaque cluster)

Méthode des k-means

- **Algorithme des k-means**
  - **Entrée:**
    - **k:** nombre des groupes à générer
    - **Les données à grouper**
  - **Sortie:**
    - **Les données réparties en k groupes**
  - **Notation:**
    - **$X=\{x_1, ..., x_n\}$:** l'ensemble des objets à grouper
    - **C_j avec j entre 1 et k:** le groupe numéro j
    - **G_j avec j entre 1 et k:** le barycentre du groupe numéro j



$$n_j = |C_j|$$

$$G_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i$$

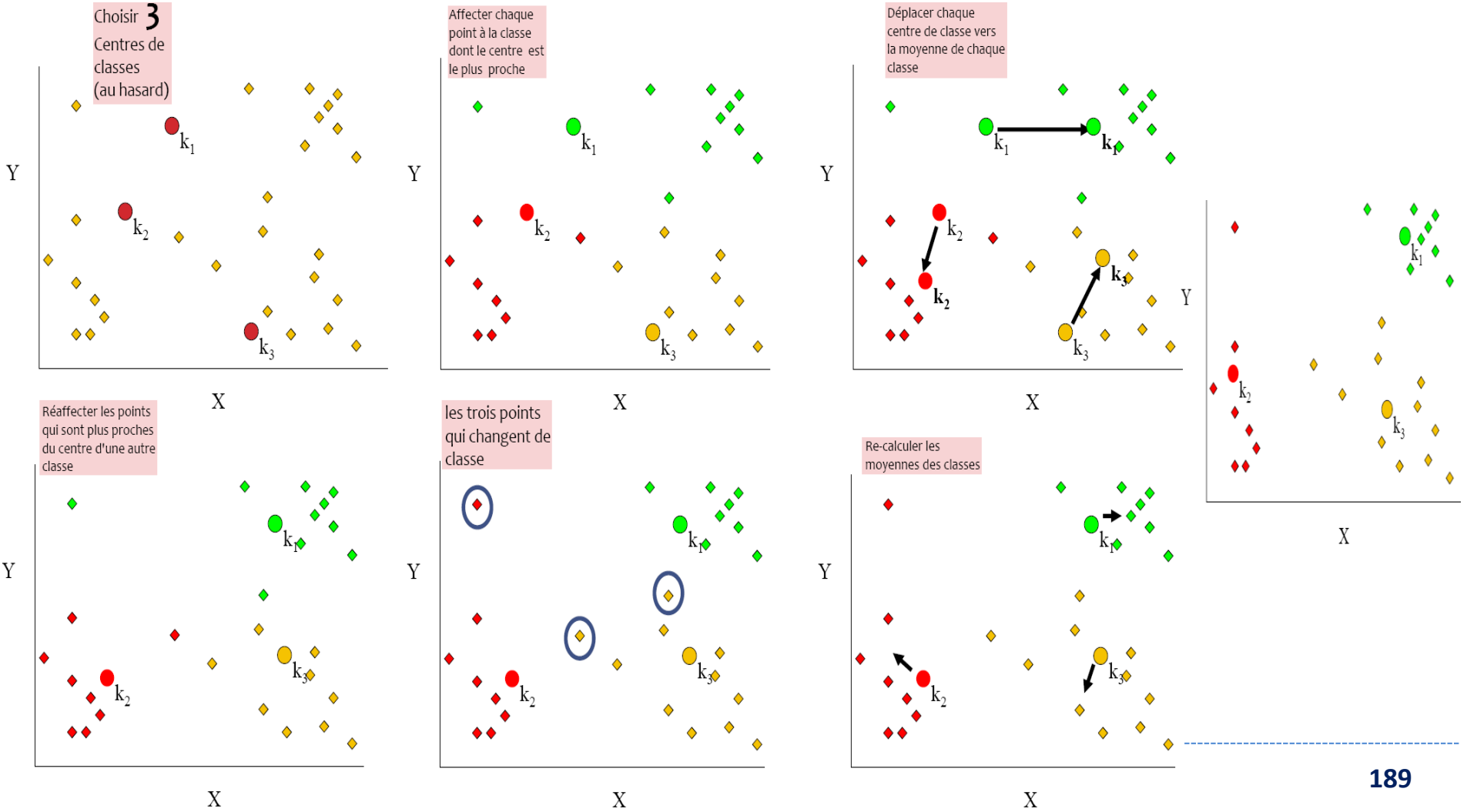
Méthode des k-means

- **Algorithme des k-means …**

1- choisir *k* centres initiaux G_1, G_2, …, G_k (au hasard par exemple)
2- **Répéter**
3- **Pour** chaque x_i de X **faire**
4- **Pour** chaque groupe C_j **faire**
5- calculer $d(G_j, x_i)$
6- **Fin Pour**
7- **(Ré)affecter** x_i au cluster C_f de centre G_f tel que $d(G_f, x_i)$ est minimale
8- **Fin Pour**
9- Recalculer le centre de chaque cluster
10- **Jusqu'à** (stabilité des ***centres*** : pas d'affectations d'objets à faire)
OU (nombre d'itérations = ***t***)
OU (stabilisation de l'***inertie totale*** de la population)

Méthode des k-means

- **Exemple 1:**



Méthode des k-means

- **Exemple 2:**
  - A={1,2,3,6,7,8,13,15,17} . Créer 3 clusters à partir de A
  - On prend 3 objets au hasard. Supposons que c'est 1, 2 et 3.
    Ça donne C_1={1}, G_1=1, C_2={2}, G_2=2, C_3={3} et G_3=3
  - Chaque objet x est affecté au cluster au milieu duquel, x est le plus proche.
    - 6 est affecté à C_3 car $d(G_3,6)=3<d(G_2,6)=4$ et $d(G_3,6)=3<d(G_1,6)=5$
    - 7 est affecté à C_3 car $d(G_3,7)=4<d(G_2,7)=5$ et $d(G_3,7)=4<d(G_1,7)=6$
    - 8 est affecté à C_3 car $d(G_3,8)=5<d(G_2,8)=6$ et $d(G_3,8)=5<d(G_1,8)=7$
    - 13 est affecté à C_3 car $d(G_3,13)=10<d(G_2,13)=11$ et $d(G_3,13)=10<d(G_1,13)=12$
    - 15 est affecté à C_3 car $d(G_3,15)=12<d(G_2,15)=13$ et $d(G_3,15)=12<d(G_1,15)=14$
    - 17 est affecté à C_3 car $d(G_3,17)=14<d(G_2,17)=15$ et $d(G_3,17)=14<d(G_1,17)=16$

    On a C_1={1}, G_1=1,
    C_2={2}, G_2=2,
    C_3={3, 6,7,8,13,15,17}, G_3=69/7=9.86

Méthode des k-means

- **Exemple 2 …**
  - $d(3,G_2)=1<d(3,G_3)=6.86$ → 3 passe dans C_2. Tous les autres objets ne bougent pas.
    $C_1=\{1\}$, $G_1=1$, $C_2=\{2,3\}$, $G_2=2.5$, $C_3=\{6,7,8,13,15,17\}$ et $G_3= 66/6=11$
  - $d(6,G_2)=3.5<d(6,G_3)=5$ → 6 passe dans C_2. Tous les autres objets ne bougent pas.
    $C_1=\{1\}$, $G_1=1$, $C_2=\{2,3,6\}$, $G_2=11/3=3.67$, $C_3=\{7,8,13,15,17\}$, $G_3= 12$
  - $d(2,G_1)=1<d(2,G_2)=1.67$ → 2 passe en C_1. $d(7,G_2)=3,33<d(7,G_3)=5$ → 7 passe en C_2. Les autres ne bougent pas.
    $C_1=\{1,2\}$, $G_1=1.5$, $C_2=\{3,6,7\}$, $G_2=5.34$, $C_3= \{8,13,15,17\}$, $G_3=13.25$
  - $d(3,G_1)=1.5<d(3,G_2)=2.34$ → 3 passe en 1. $d(8,G_2)=2.66<d(8,G_3)=5.25$ → 8 passe en C_2.
    $C_1=\{1,2,3\}$, $G_1=2$, $C_2=\{6,7,8\}$, $G_2=7$, $C_3=\{13,15,17\}$, $G_3=15$

Plus rien ne bouge

Méthode des k-means

- **Exemple 3:**
  - 8 points A, …, H de l 'espace euclidéen 2D. k=2 (2 groupes)
  - Tire aléatoirement 2 centres : B et D choisis.

d(A,B)=d(A,D)=$\sqrt{2} = 1{,}41$

d(C,B)=d(C,D)=1

d(E,B)=2; d(E,D)=$\sqrt{8} = 2{,}83$

d(F,B)=3; d(F,D)=$\sqrt{13} = 3{,}61$

d(G,B)=4; d(G,D)=$\sqrt{20} = 4{,}47$

d(H,B)=d(H,D)= $\sqrt{26} = 5{,}1$

points	Centre D(2,4), B(2,2)	Centre D(2,4), I(27/7,17/7)	Centre J(5/3,10/3), K(24/5,11/5)
A(1,3)	B	D	J
B(2,2)	B	I	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Méthode des k-means

- **Exemple 4:**

	X	Y	Cluster
I1	5	0	1
I2	5	2	2
I3	3	1	1
I4	0	4	2
I5	2	1	1
I6	4	2	2
I7	2	2	1
I8	2	3	2
I9	1	3	1
I10	5	4	2

- Soit X et Y deux variables indépendantes caractérisant un objet donné.
- L'objectif est d'appliquer le K-means pour classifier les différents objets en une des deux classes.
- Soit une initialisation aléatoire de l'affectation des différents objets de la base à une classe donnée.

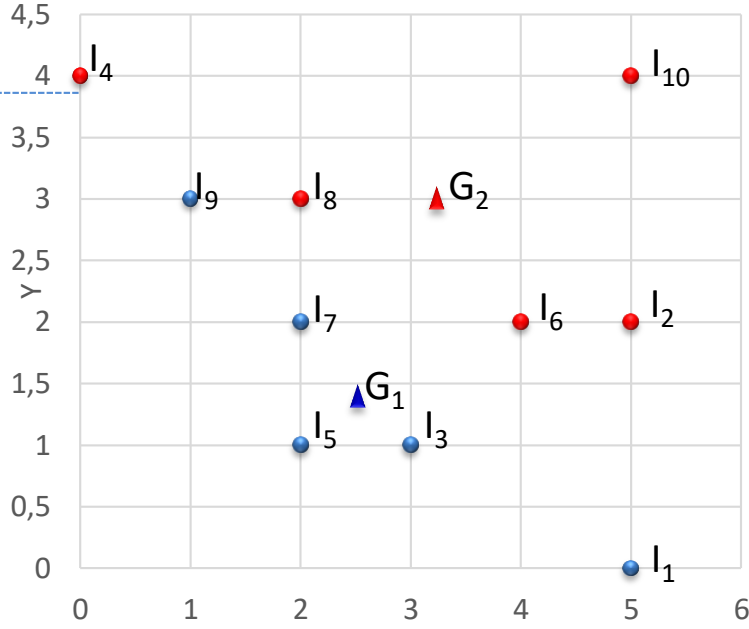
Méthode des k-means

Exemple 4 …

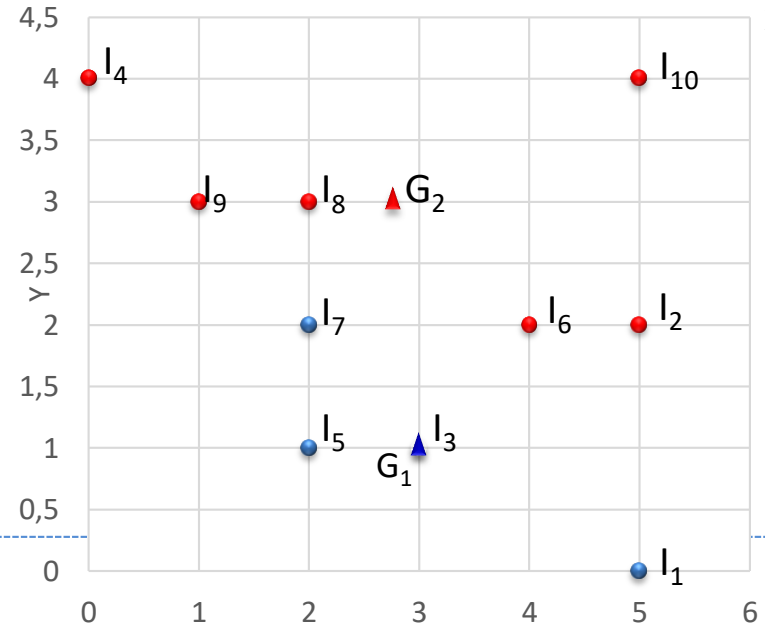
C_1={I1,I3,I5,I7,I9}, G_1=(2.6, 1.4)
C_2={I2,I4,I6,I8,I10}, G_2=(3.2, 3)

Distance de Manhattan

d(I1,G_1)= **3.8** d(I1,G_2)= 4.8	d(I2,G_1)= 3 d(I2,G_2)= **2.8**
d(I3,G_1)= **0.8** d(I3,G_2)= 2.2	d(I4,G_1)= 5.2 d(I4,G_2)= **4.2**
d(I5,G_1)= **1** d(I5,G_2)= 3.2	d(I6,G_1)= 2 d(I6,G_2)= **1.8**
d(I7,G_1)= **0.8** d(I7,G_2)= 2.2	d(I8,G_1)= 2.2 d(I8,G_2)= **1.2**
d(I9,G_1)= 3.2 d(I9,G_2)= **2.2**	d(I10,G_1)= 5 d(I10,G_2)= **2.8**



C_1={I1,I3,I5,I7}, G_1=I3=(3, 1)
C_2={I2,I4,I6,I8,**I9**,I10}, G_2=(2.83, 3)

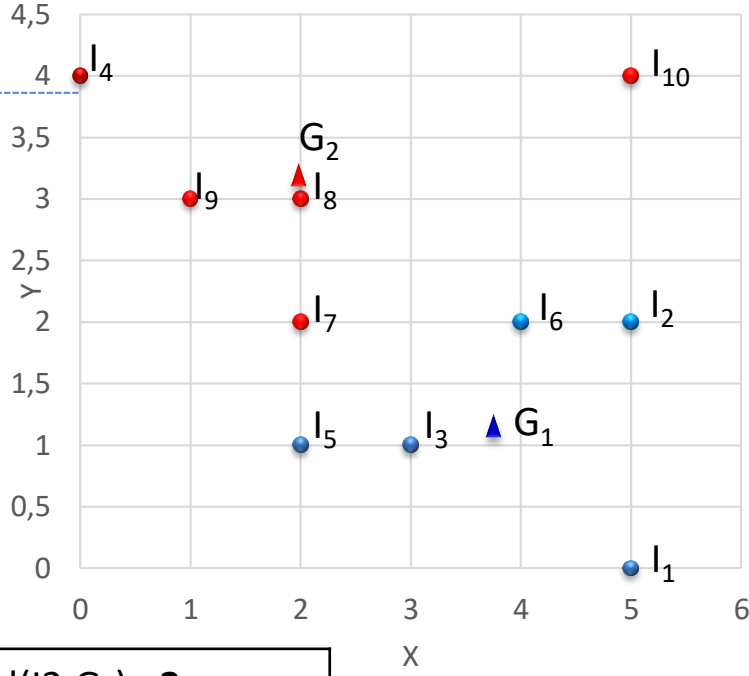


Méthode des k-means

- **Exemple 4 …**

d(I1,G_1)= **3** d(I1,G_2)= 5.17	d(I2,G_1)= **3** d(I2,G_2)= 3.17
d(I3,G_1)= **0** d(I3,G_2)= 2.17	d(I4,G_1)= 9 d(I4,G_2)= **2.17**
d(I5,G_1)= **1** d(I5,G_2)= 2.83	d(I6,G_1)= **2** d(I6,G_2)= 2.17
d(I7,G_1)= 2 d(I7,G_2)= **1.83**	d(I8,G_1)= 3 d(I8,G_2)= **0.83**
d(I9,G_1)= 4 d(I9,G_2)= **1.83**	d(I10,G_1)= 5 d(I10,G_2)= **3.17**

C_1={I1,**I2**,I3,I5,**I6**}, G_1=(3.8, 1.2)
C_2={I4,**I7**,I8,I9,I10}, G_2=(2, 3.2)



d(I1,G_1)= **2.4** d(I1,G_2)= 6.2	d(I2,G_1)= **3** d(I2,G_2)= 4.2
d(I3,G_1)= **1** d(I3,G_2)= 3.2	d(I4,G_1)= 6.6 d(I4,G_2)= **2.8**
d(I5,G_1)= **2** d(I5,G_2)= 2.2	d(I6,G_1)= **1** d(I6,G_2)= 3.2
d(I7,G_1)= 2.6 d(I7,G_2)= **1.2**	d(I8,G_1)= 3.6 d(I8,G_2)= **0.2**
d(I9,G_1)= 4.6 d(I9,G_2)= **1.2**	d(I10,G_1)= 4 d(I10,G_2)= **3.8**

Plus rien ne bouge

Méthode des k-means

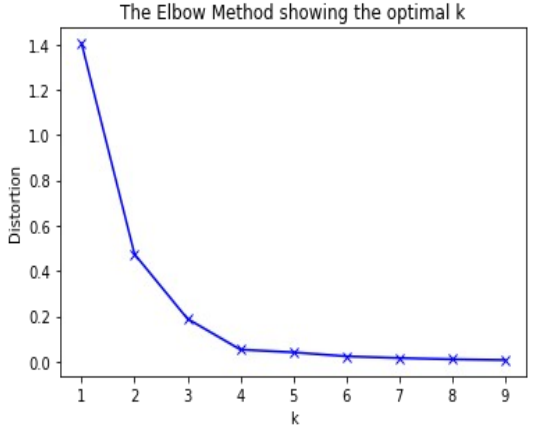
- **Application:**
  - **Marketing :** segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats
  - **Environnement :** identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observations de la terre
  - **Assurance :** identification de groupes d'assurés distincts associés à un nombre important de déclarations
  - **Planification de villes :** identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique, ...
  - **Médecine :** Localisation de tumeurs dans le cerveau → Nuage de points du cerveau fournis par le neurologue → Identification des points définissant une tumeur

Méthode des k-means

- **Avantages de la méthode des k-means:**
  - **efficace :** complexité de $\mathbf{O}(\boldsymbol{tkn})$,
    - avec $\boldsymbol{n}$ le nombre d'objets,
    - $\boldsymbol{t}$ le nombre d'itérations et en général $\boldsymbol{t}$ et $\boldsymbol{k} << \boldsymbol{n}$
  - **Très populaire:** très facile à comprendre et à mettre en œuvre
  - La méthode résolve une tâche **non supervisée**, donc elle ne nécessite aucune information sur les données
- **Inconvénients de la méthode des k-means:**
  - Besoin de spécifier $\boldsymbol{k}$ à l'avance
  - Incapable de traiter des données bruitées ou manquantes
  - Le résultat peut varier considérablement en fonction du choix initial des centres de cluster

Méthode des k-means

- **Comment trouver le bon k?**
  - Essayer plusieurs k
  - Calculer à chaque fois la distance moyenne avec le barycentre de chaque cluster: I_{intra}



- **Si les classes se chevauchent ?**
  - x_i appartient à C_j avec un poids w_{ij}
  - Probabilité que x_i appartient à C_j: $P(C_j|x_i)$
  - w_{ij} est normalisée pour tous les points x_i

$$\sum_{j=1}^{k} w_{ij} = 1$$

  - Il s'agit de la variante: **k-means floues**
  - Autre amélioration possible:
    - Essayer diverses métriques pour calculer la ressemblance entre les données

k-means & ACP

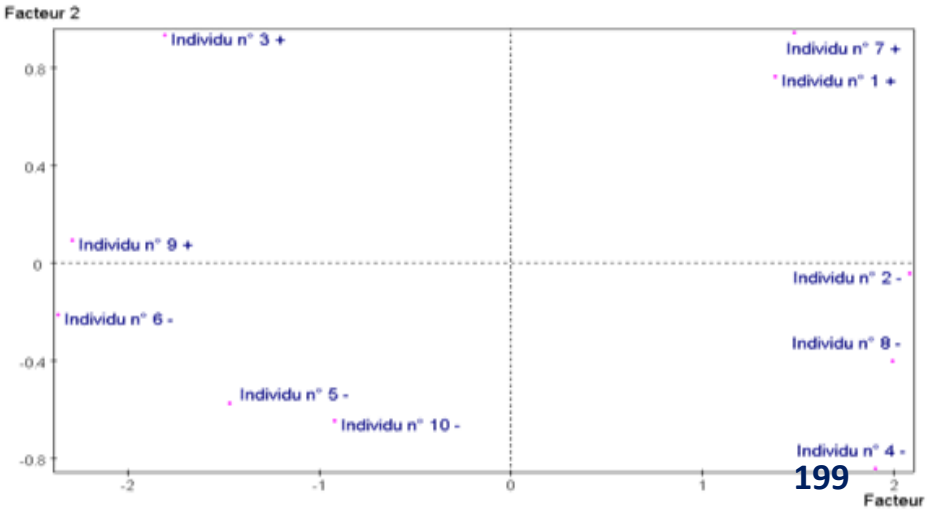
ACP

- **Reprenons l'exercice D145**

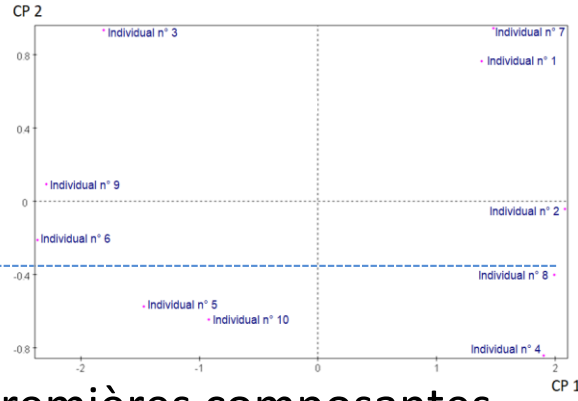
	Stat.	Math	Cpta	G° Fi
Individu n° 1	19	14	8	18
Individu n° 2	20	12	4	4
Individu n° 3	10	10	32	38
Individu n° 4	13	17	4	4
Individu n° 5	6	8	26	24
Individu n° 6	6	3	28	32
Individu n° 7	19	16	8	20
Individu n° 8	15	18	6	6
Individu n° 9	9	2	32	30
Individu n° 10	8	7	20	20

	CP_1	CP_2	CP_3	CP_4
Individu n° 1	1.38	0.77	0.20	0.14
Individu n° 2	2.08	-0.04	0.97	-0.15
Individu n° 3	-1.81	0.93	-0.71	-0.17
Individu n° 4	1.90	-0.85	-0.43	0.04
Individu n° 5	-1.47	-0.58	-0.36	-0.07
Individu n° 6	-2.37	-0.22	0.21	0.28
Individu n° 7	1.48	0.94	-0.16	0.16
Individu n° 8	2.00	-0.40	-0.48	-0.15
Individu n° 9	-2.29	0.09	0.63	-0.21
Individu n° 10	-0.92	-0.65	0.13	0.13

	Valeur propre	Pourcentage	Pourcentage cumulé
1	3.3189	82.9700	82.9700
2	0.4035	10.0900	93.0600
3	0.2508	6.2700	99.3300
4	0.0268	0.6700	100.0000



k-means & ACP



- **Reprenons l'exercice D145 …**
  - L'algorithme des K-means sera appliqué sur les deux premières composantes principales au lieu des 4 variables initiales
  - Pour K=2, G_1=I2 et G_2=I6, avec la distance Euclidienne

d(I1,G_1)= **1,07** d(I1,G_2)= 3,88	d(I2,G_1)= **0** d(I2,G_2)= 4,45
d(I3,G_1)= 4,01 d(I3,G_2)= **1,28**	d(I4,G_1)= **0,83** d(I4,G_2)= 4,32
d(I5,G_1)= 3,59 d(I5,G_2)= **0,97**	d(I6,G_1)= 4,45 d(I6,G_2)= **0**
d(I7,G_1)= **1,15** d(I7,G_2)= 4,02	d(I8,G_1)= **0,37** d(I8,G_2)= 4,37
d(I9,G_1)= 4,37 d(I9,G_2)= **0,32**	d(I10,G_1)= 3,06 d(I10,G_2)= **1,51**

d(I1,G_1)= **0,79** d(I1,G_2)= 3,27	d(I2,G_1)= **0,34** d(I2,G_2)= 3,85
d(I3,G_1)= 3,68 d(I3,G_2)= **1,02**	d(I4,G_1)= **0,94** d(I4,G_2)= 3,75
d(I5,G_1)= 3,31 d(I5,G_2)= **0,58**	d(I6,G_1)= 4,15 d(I6,G_2)= **0,61**
d(I7,G_1)= **0,90** d(I7,G_2)= 3,41	d(I8,G_1)= **0,54** d(I8,G_2)= 3,79
d(I9,G_1)= 4,06 d(I9,G_2)= **0,55**	d(I10,G_1)= 2,79 d(I10,G_2)= **1,02**

Rien ne bouge

C_1={I1,I2,I4,I7,I8}, G_1=(1.77, 0.08)
C_2={I3,I5,I6,I9,I10}, G_2=(-1.77, -0.09)

$$I_{intra} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_j, G_i) = 1,163$$

k-means & ACP

- **Reprenons l'exercice D145 …**

$$On\ a\ K = 2, n_1 = 5, n_2 = 5$$

$$I_{intra} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_j, G_i)$$

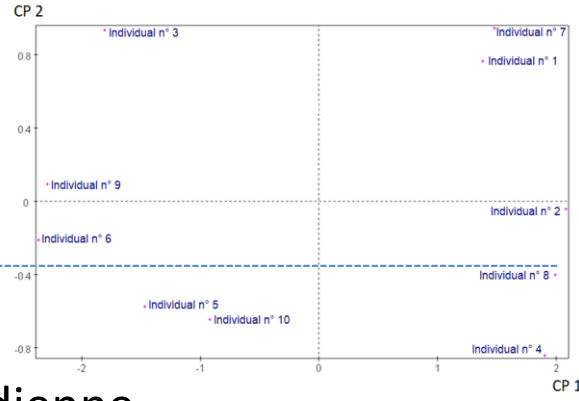
$$= \frac{1}{5}\left(d^2(I1, G_1) + d^2(I2, G_1) + d^2(I4, G_1) + d^2(I7, G_1) + d^2(I8, G_1)\right)$$

$$+ \frac{1}{5}\left(d^2(I3, G_2) + d^2(I5, G_2) + d^2(I6, G_2) + d^2(I9, G_2) + d^2(I10, G_2)\right)$$

$$= \frac{1}{5}\left(0{,}79^2 + 0{,}34^2 + 0{,}94^2 + 0{,}9^2 + 0{,}54^2\right) + \frac{1}{5}\left(1{,}02^2 + 0{,}58^2 + 0{,}61^2 + 0{,}55^2 + 1{,}02^2\right)$$

$$= 1{,}163$$

k-means & ACP



- **Reprenons l'exercice D145 …**
  - Pour K=3, G_1=I1, G_2=I2 et G_3=I6, avec la distance Euclidienne

d(I1,G_1)= **0** d(I1,G_2)= 1,07 d(I1,G_3)= 3,88	d(I2,G_1)= 1,07 d(I2,G_2)= **0** d(I2,G_3)= 4,45	d(I1,G_1)= **0,10** d(I1,G_2)= 1,35 d(I1,G_3)= 3,27	d(I2,G_1)= 1,45 d(I2,G_2)= **0,24** d(I2,G_3)= 3,85
d(I3,G_1)= 3,19 d(I3,G_2)= 4,01 d(I3,G_3)= **1,28**	d(I4,G_1)= 1,70 d(I4,G_2)= **0,83** d(I4,G_3)= 4,32	d(I3,G_1)= 3,24 d(I3,G_2)= 4,04 d(I3,G_3)= **1,02**	d(I4,G_1)= 1,77 d(I4,G_2)= **0,43** d(I4,G_3)= 3,75
d(I5,G_1)= 3,15 d(I5,G_2)= 3,59 d(I5,G_3)= **0,97**	d(I6,G_1)= 3,88 d(I6,G_2)= 4,45 d(I6,G_3)= **0**	d(I5,G_1)= 3,24 d(I5,G_2)= 3,47 d(I5,G_3)= **0,58**	d(I6,G_1)= 3,95 d(I6,G_2)= 4,37 d(I6,G_3)= **0,61**
d(I7,G_1)= **0,20** d(I7,G_2)= 1,15 d(I7,G_3)= 4,02	d(I8,G_1)= 1,32 d(I8,G_2)= **0,37** d(I8,G_3)= 4,37	d(I7,G_1)= **0,10** d(I7,G_2)= 1,46 d(I7,G_3)= 3,41	d(I8,G_1)= 1,38 d(I8,G_2)= **0,03** d(I8,G_3)= 3,79
d(I9,G_1)= 3,73 d(I9,G_2)= 4,37 d(I9,G_3)= **0,32**	d(I10,G_1)= 2,70 d(I10,G_2)= 3,06 d(I10,G_3)= **1,51**	d(I9,G_1)= 3,80 d(I9,G_2)= 4,31 d(I9,G_3)= **0,55**	d(I10,G_1)= 2,79 d(I10,G_2)= 2,92 d(I10,G_3)= **1,02**

Rien ne bouge

C_1={I1, I7}, G_1=(1.43, 0.86)
C_2={I2,I4, I8}, G_2=(1.99, -0.43)
C_3={I3,I5,I6,I9,I10}, G_3=(-1.77, -0.09)

$$I_{intra} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_j, G_i) = 0,709$$

k-means & ACP

- **Reprenons l'exercice D145 …**

$$On\ a\ K = 3, n_1 = 2, n_2 = 3, n_3 = 5$$

$$I_{intra} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} d^2(x_j, G_i)$$

$$= \frac{1}{2}\left(d^2(I1, G_1) + d^2(I7, G_1)\right) + \frac{1}{3}\left(d^2(I2, G_2) + d^2(I4, G_2) + d^2(I8, G_2)\right)$$
$$+ \frac{1}{5}\left(d^2(I3, G_3) + d^2(I5, G_3) + d^2(I6, G_3) + d^2(I9, G_3) + d^2(I10, G_3)\right)$$

$$= \frac{1}{2}\left(0{,}1^2 + 0{,}1^2\right) + \frac{1}{3}\left(0{,}24^2 + 0{,}43^2 + 0{,}03^2\right) + \frac{1}{5}\left(1{,}02^2 + 0{,}58^2 + 0{,}61^2 + 0{,}55^2 + 1{,}02^2\right)$$

$$= 0{,}709$$

Classification hiérarchique

- **Principe:** la classification hiérarchique est une autre méthode de regroupement qui fonctionne par groupement successif de groupes « proches »
  - Au départ chaque point est considéré comme un groupe
    - Si on dispose de $\mathbf{N}$ données donc nous constituons $\mathbf{N}$ groupes: $\mathbf{C = N}$
  - Détecter, par la suite, les 2 groupes les plus proches par le calcul du **distance entre groupe**
  - Les **agréger** pour n'en former qu'un seul, nous disposons maintenant de $(\mathbf{N - 1})$ groupes: $\mathbf{C = N - 1}$
  - Répéter les deux étapes précédentes jusqu'à tous les individus forment un **seul groupe:** $\mathbf{C = 1}$
- Une hiérarchie des données est alors construire appelée: **dendrogramme**

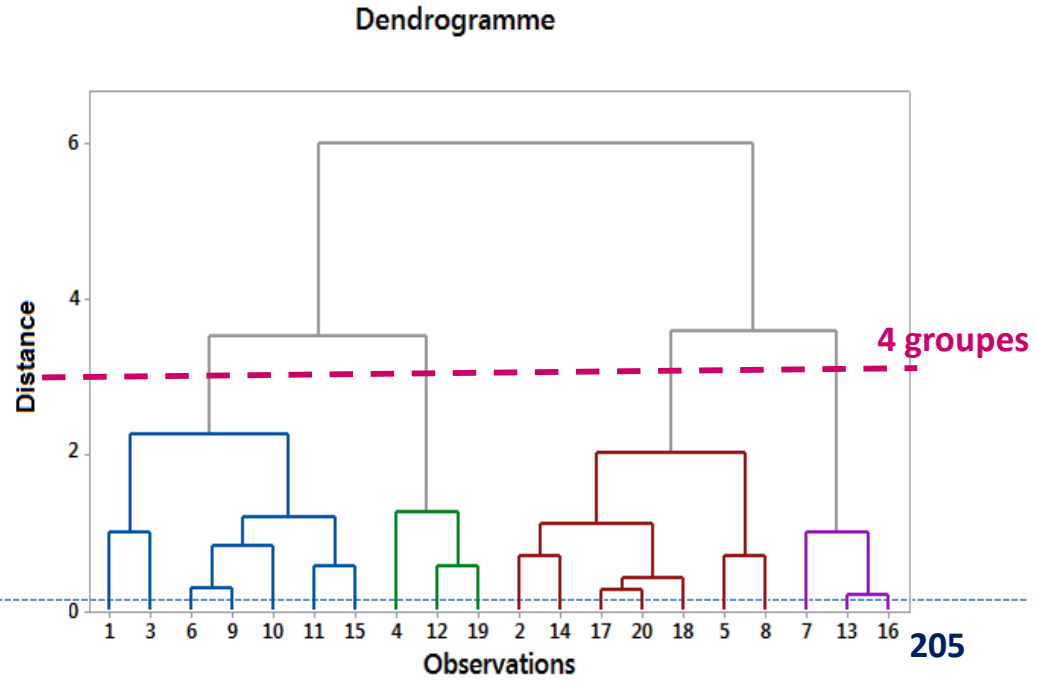
Classification hiérarchique

- Le dendrogramme est un diagramme en forme d'**arbre** qui montre comment les groupes sont fusionnés hiérarchiquement
  - La racine représente l'ensemble du jeu de données
  - Une feuille représente un seul objet
  - Un nœud interne représente l'union de tous les objets du sous-arbre
  - La hauteur d'un nœud interne représente la distance entre ses 2 nœuds enfants

Au départ d'une CH → $I_{intra} = 0$ et $I_{inter} = I_{tot}$

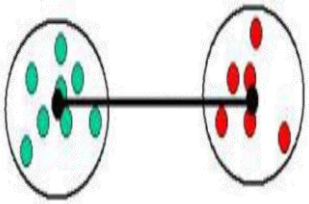
À la fin d'une CH → $I_{intra} = I_{tot}$ et $I_{inter} = 0$

→Une partition des observations en k clusters est obtenue en coupant le dendrogramme à un niveau souhaité



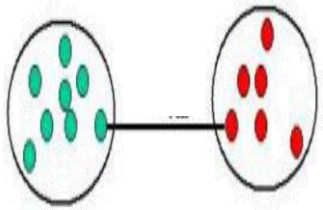
Classification hiérarchique

- **Comparaison des groupes (distance entre groupes)**
  - **Méthode du centroïde:** distance entre les centroïdes (centres de gravité) des clusters: $d(C_1, C_2) = d(G_1, G_2)$



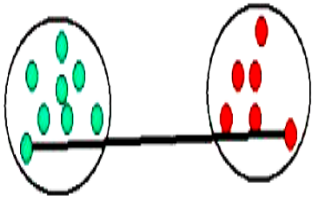
  - **Méthode du lien simple (*single linkage*) ou saut minimal:** plus petite distance entre toutes les paires d'éléments de 2 clusters C_1 et C_2

$$d(C_1, C_2) = \min_{a \in C_1, b \in C_2} d(a, b)$$



  - **Méthode du lien complet (*complete linkage*) ou saut maximal:** plus grande distance entre toutes les paires d'éléments de 2 clusters C_1 et C_2

$$d(C_1, C_2) = \max_{a \in C_1, b \in C_2} d(a, b)$$



  - **Méthode du lien moyen (*average linkage*):** moyenne des distances entre les paires d'éléments entre 2 clusters C_1 et C_2

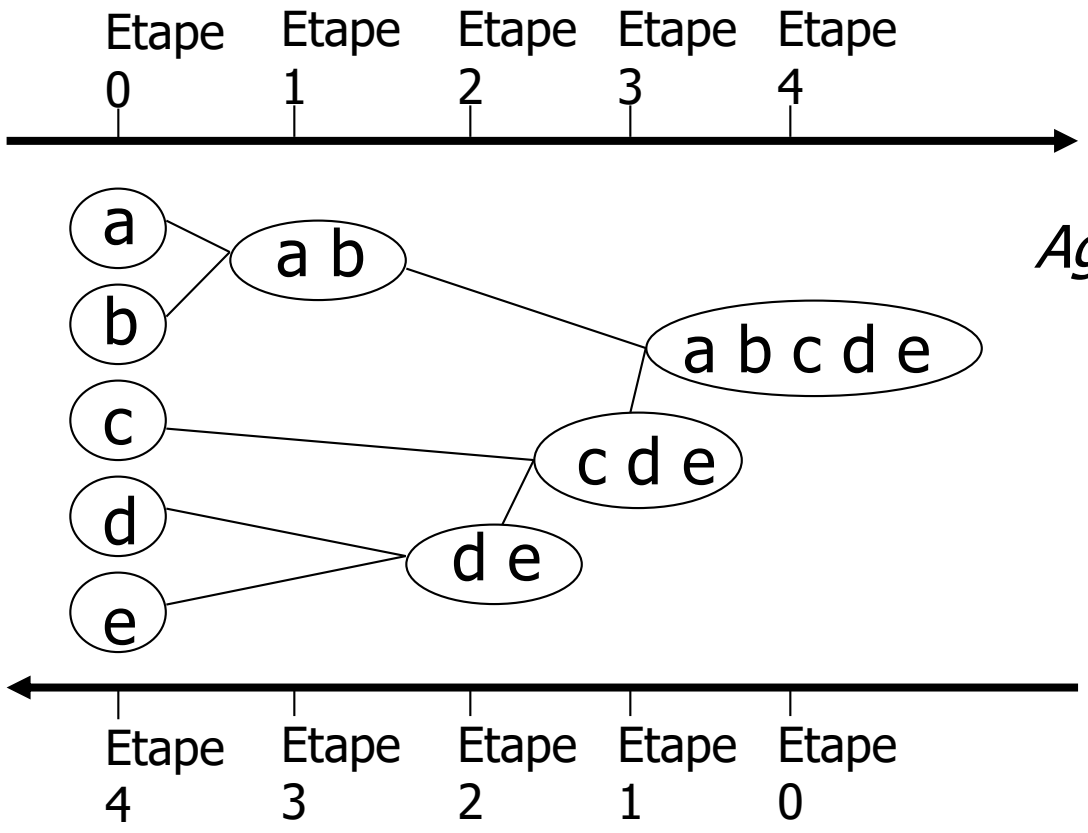
$$d(C_1, C_2) = \underset{a \in C_1, b \in C_2}{moyenne}(d(a, b)) = \frac{\sum_{a \in C_1} \sum_{b \in C_2} d(a, b)}{n_1 n_2}$$

Classification hiérarchique

- **Deux grandes approches**
  - **Classification ascendante hiérarchique : CAH (Agglomération)**
    - La méthode la plus communément utilisée
    - Commencer avec les points en tant que clusters individuels.
    - A chaque étape, **grouper** les clusters les plus proches jusqu'à obtenir 1 seul ou k clusters.
  - **Classification descendante hiérarchique (Division)**
    - Commencer avec 1 seul cluster comprenant tous les points.
    - A chaque étape, **diviser** un cluster jusqu'à obtenir des clusters ne contenant qu'un point ou jusqu'à obtenir k clusters.

Classification hiérarchique

- **Deux grandes approches …**



agglomération

(AGNES:

Agglomerative Nesting)

division

(DIANA:

Divisive Analysis)

Classification Ascendante Hiérarchique

- **Algorithme de classification ascendante hiérarchique:**

1- C = N
2- **Pour** chaque **i** de 1 à N **faire**
3- $C_i \leftarrow \{ x_i \}$
4- **Pour** chaque **j** de 1 à (i - 1) **faire**
5- calculer **$d(C_i, C_j)$**
6- **Fin Pour**
7- **Fin Pour**
8- Construire la matrice de dissimilarité **M**
9- **Répéter**
10- Sélection dans **M** des deux clusters les plus proches C_i et C_j
11- Fusion de C_i et C_j pour former un cluster C_k
12- Mise à jour de **M** en calculant la distance entre C_k et tous les autres clusters
13- **C - -**
14- **Jusqu'à** (C ==1)

Classification Ascendante Hiérarchique

Méthodes d'agrégation et résultats:

Exemple: Utilisez la méthode du lien simple et celle du lien complet pour regrouper les données décrites par la matrice de distance suivante. Puis tracer les dendrogrammes correspondants.

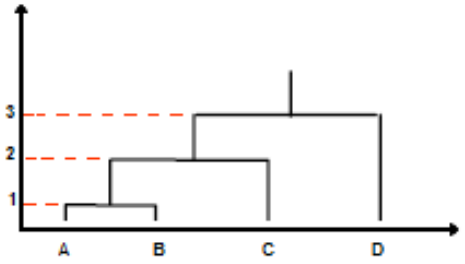
	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Correction:

- **Méthode du lien simple:**

	C	D	{A, B}
C	0	3	2
D		0	5
{A, B}			0

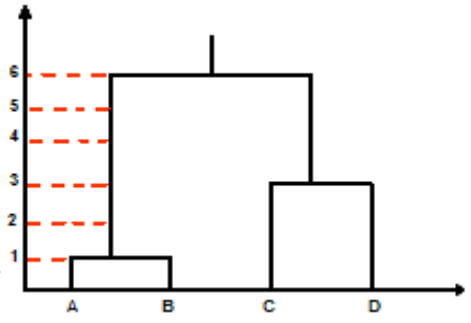
	D	{{A, B}, C}
D	0	3
{{A, B}, C}		0



- **Méthode du lien complet:**

	C	D	{A, B}
C	0	3	4
D		0	6
{A, B}			0

	{A, B}	{C, D}
{A, B}	0	6
{C, D}		0



Classification Ascendante Hiérarchique

- **Exercice 1 :**

	a	b	c	d	e	f	g
a	0	1	3	6	7	11	16
b		0	2	5	6	10	15
c			0	3	4	8	13
d				0	1	5	10
e					0	4	9
f						0	5
g							0

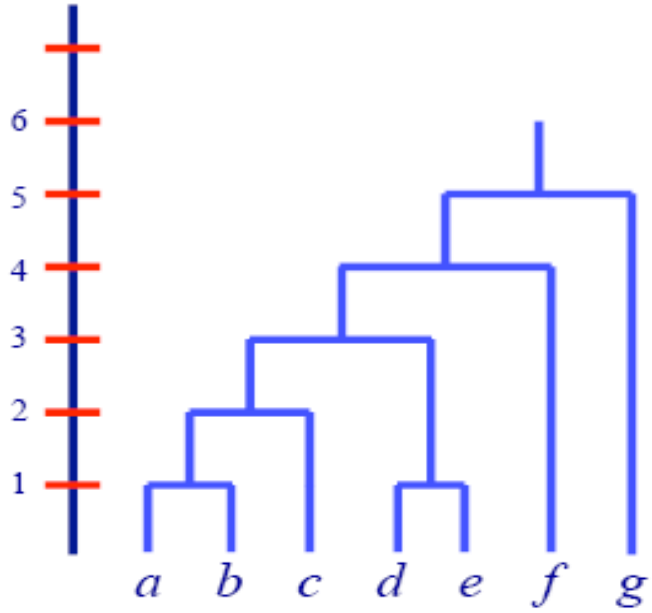
Lien Simple = Saut minimal

	ab	c	de	f	g
ab	0	2	5	10	15
c		0	3	8	13
de			0	4	9
f				0	5
g					0

	abc	de	f	g
abc	0	3	8	13
de		0	4	9
f			0	5
g				0

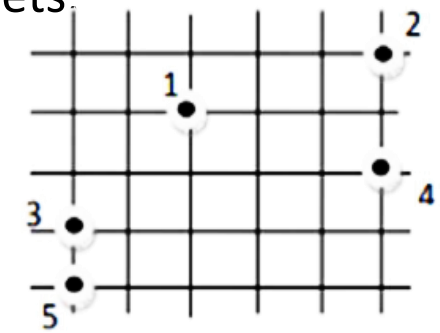
	abcde	f	g
abcde	0	4	9
f		0	5
g			0

	abcdef	g
abcdef	0	5
g		0



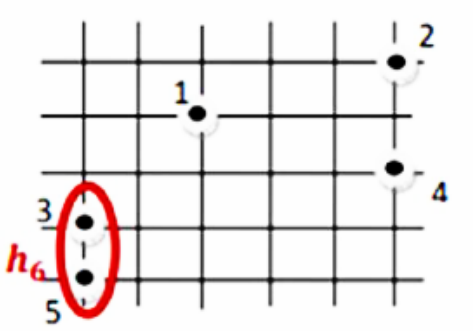
Classification Ascendante Hiérarchique

- **Exercice 2:** Soit un ensemble d'objets représentés par des points numérotés de 1 à 5, dans un repère euclidien. Notons la distance euclidienne mesurée entre les objets



d	1	2	3	4	5
1	0	$\sqrt{10}$	$\sqrt{8}$	$\sqrt{10}$	$\sqrt{13}$
2		0	$\sqrt{34}$	2	$\sqrt{41}$
3			0	$\sqrt{26}$	1
4				0	$\sqrt{29}$
5					0

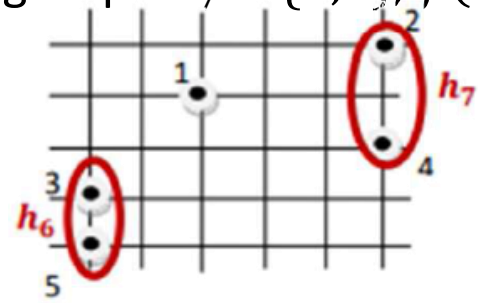
  - Fusionner les groupes 3 et 5, qui sont les groupes les plus proches (distance minimale), $d = 1$, et former un groupe $h_6 = \{3,5\}$. A ce groupe est associé son niveau (indice d'agrégation), qui est la distance entre ses deux sous-groupes 3 et 5 , $f(h_6) = 1$. Soit la **méthode du lien simple** pour la comparaison des groupes



d	1	2	4	h_6
1	0	$\sqrt{10}$	$\sqrt{10}$	$\sqrt{8}$
2		0	2	$\sqrt{34}$
4			0	$\sqrt{26}$
h_6				0

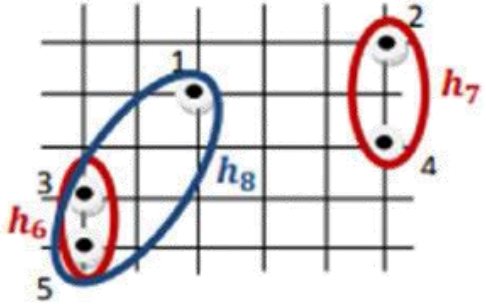
Classification Ascendante Hiérarchique

- **Exercice 2 …**
  - Fusionner les groupes 2 et 4, qui sont les groupes les plus proches, $d = 2$, et former un groupe $h_7 = \{2,4\}$, $f(h_7) = 2$.



d	1	h_6	h_7
1	0	$\sqrt{8}$	$\sqrt{10}$
h_6		0	$\sqrt{26}$
h_7			0

  - Fusionner les groupes 1 et h_6, qui sont les groupes les plus proches, $d = \sqrt{8}$, et former un groupe $h_8 = \{1, h_6\}$, $f(h_8) = \sqrt{8}$.

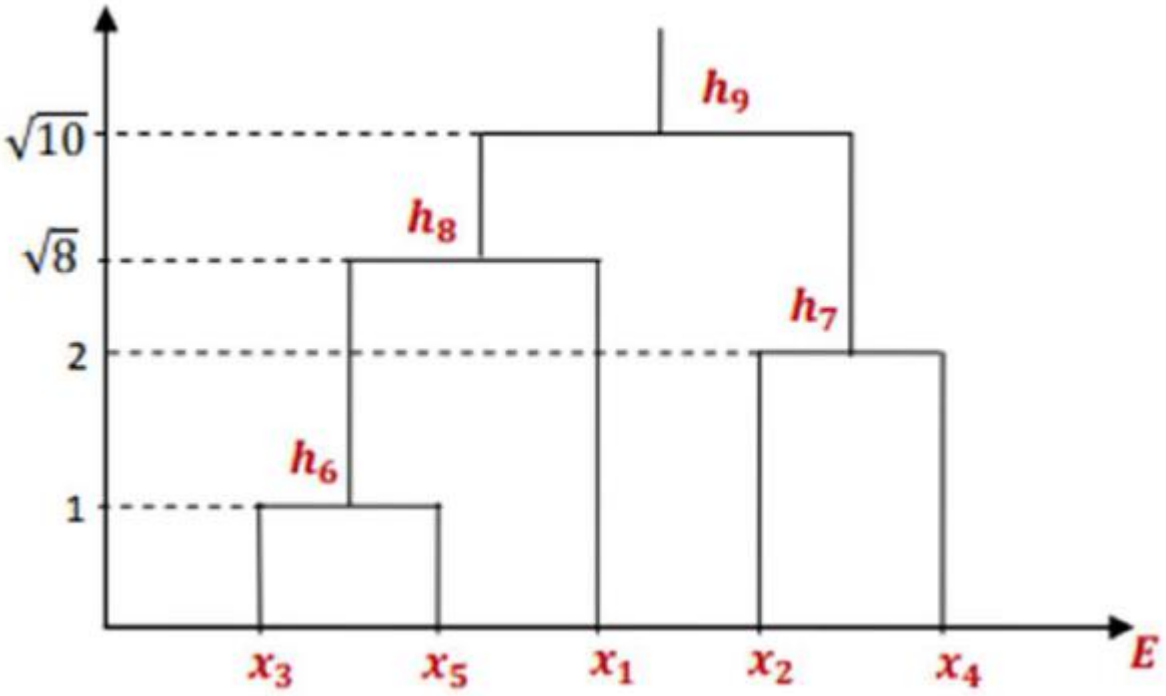


d	h_7	h_8
h_7	0	$\sqrt{10}$
h_8		0

  - Si on continue à la dernière étape de regroupement, tous les objets sont regroupés : $h_9 = \{h_7, h_8\} = \{1,2,3,4,5\}$, $f(h_9) = \sqrt{10}$

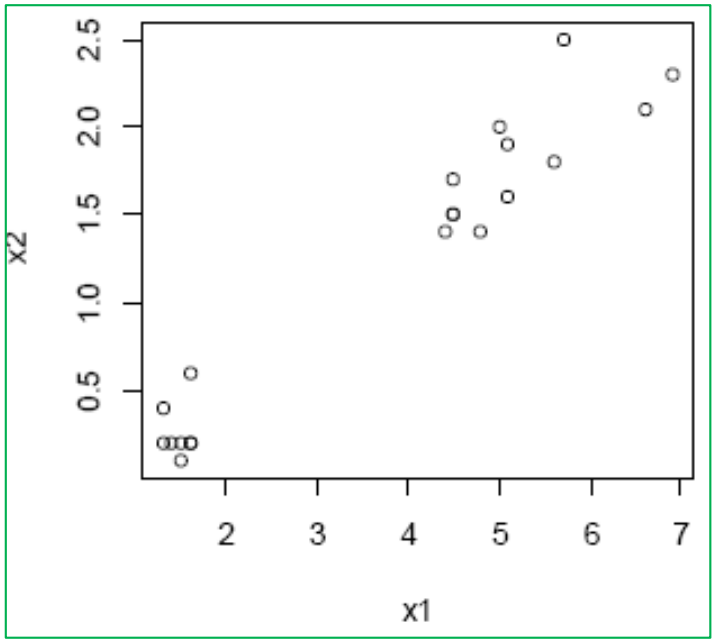
Classification Ascendante Hiérarchique

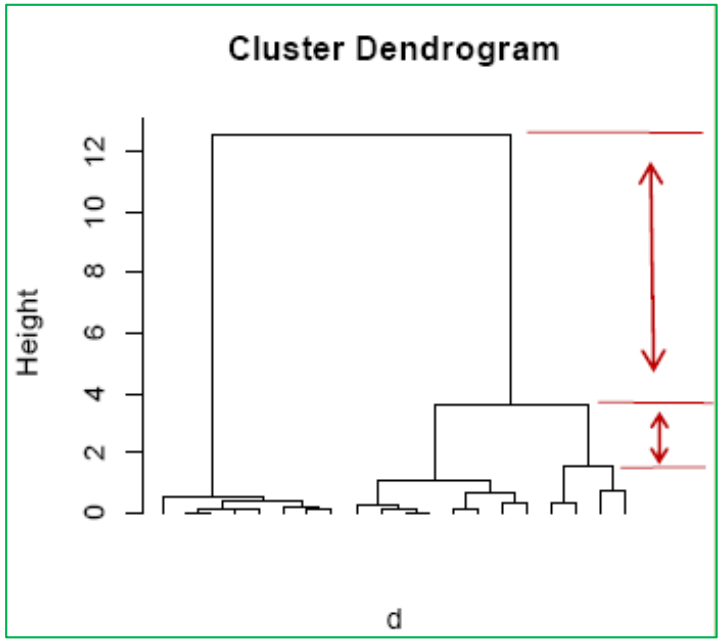
- **Exercice 2 …**
  - Cette hiérarchie de regroupement des objets peut être représentée par un dendrogramme, une représentation arborescente d'une hiérarchie.

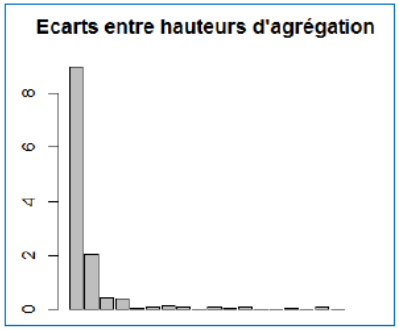


Classification hiérarchique

- **Qualité des groupes obtenus: (identification du « bon » nombre de clusters)**
  - Détectée par l'**écart entre paliers d'agrégations**
    - Des fortes différences entre deux niveaux d'agrégation successifs indique une modification «significative» de la structure des données lorsqu'on a procédé au regroupement.





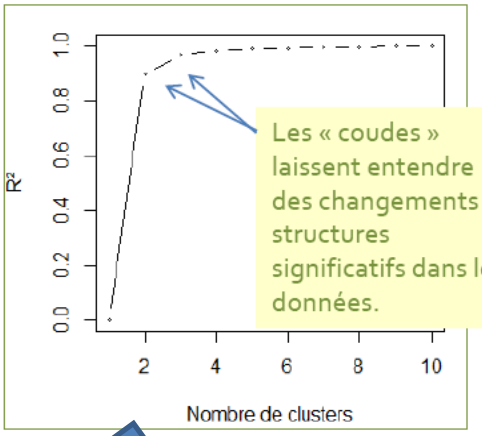


Classification hiérarchique

- **Qualité des groupes obtenus …**
  - Basée sur le calcul des inerties de la population: I_{intra} , I_{inter} , I_{total}
  - Elle peut être mesurée par:
    - La part d'inertie exprimé par:

$$R^2 = \frac{I_{inter}}{I_{total}}$$

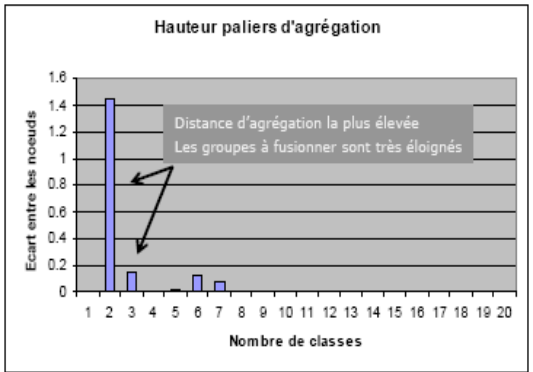
R² = 0, il y a un seul groupe.
R² = 1,Partition triviale: 1 individu = 1 groupe.

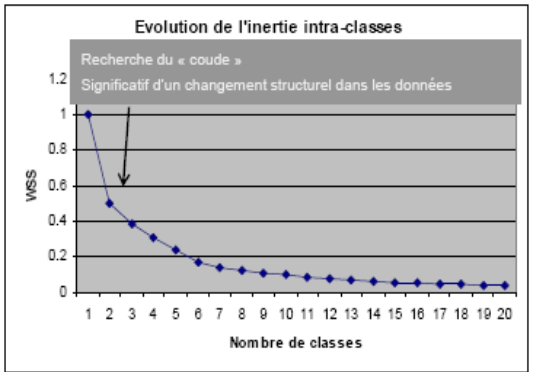


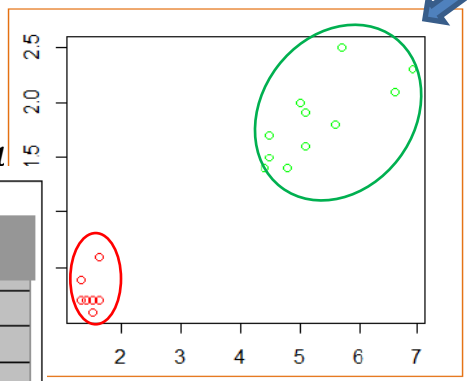
    - Le rapport de l'inertie inter par l'inertie intra:

$$Q = \frac{I_{inter}}{I_{intra}}$$

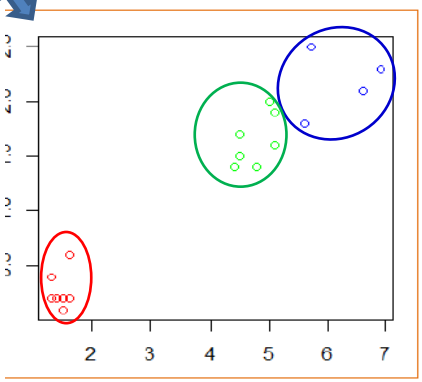
    - Ou parfois par l'inertie intra : I_{intra}







Partition en deux groupes.



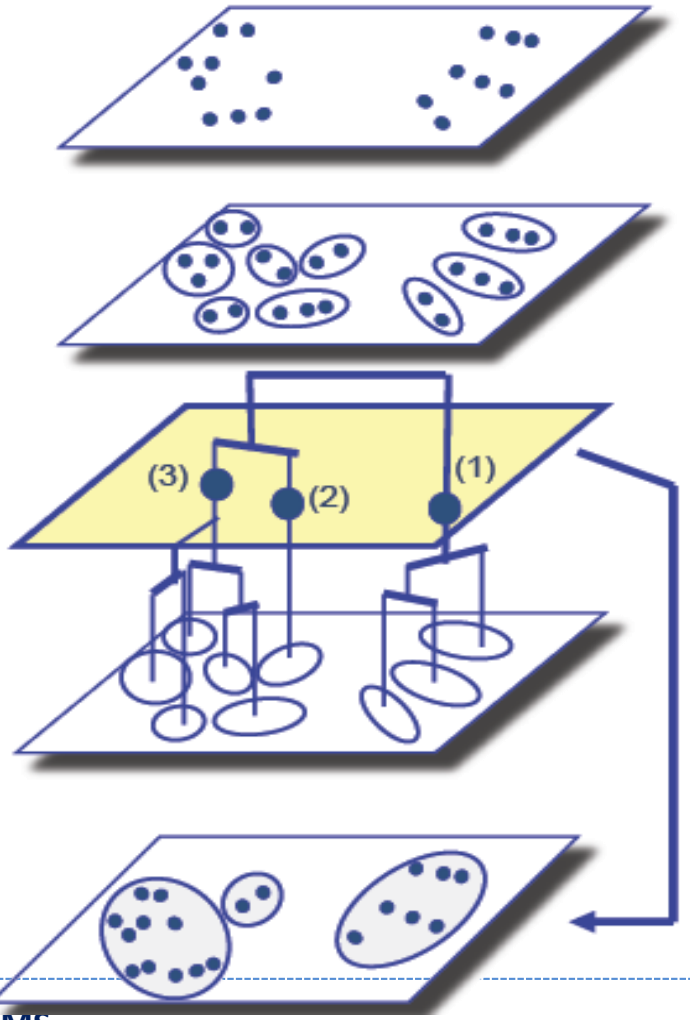
Partition en trois groupes.

Classification hiérarchique

- **Avantages**
  - Conceptuellement simple
  - Le groupement ne dépend pas d'une initialisation (comme pour les k-means)
  - Très utile quand on ne connaît pas *à priori* le nombre de classes
- **Inconvénients**
  - Les résultats dépendent de la méthode d'agrégation
  - Non adapté aux grandes volumes de données
  - Groupement obtenu est final (non modifiable par la suite) même obtenu avec des décisions erronées
- **On peut l'utiliser conjointement à une k-means:**
  - Classification hiérarchique puis k-means → pour permettre la réallocation des individus frontière
  - K-means puis Classification hiérarchique → si **N** est élevé avec un **k** de k-means grand

Classification hiérarchique

- **Clustering mixte pour gros volumes:**



1. Données avant classification
2. Partition préliminaire :
   - K-means avec K (grande) classes
3. Classification ascendante hiérarchique
   -Coupure du dendrogramme à K* classes
   Avec K*<<K
4. Partition finale (consolidation)
   - K-means avec K* (petit) classes

Conclusion

- Les méthodes de regroupement sont entièrement automatiques et se basent sur le principe d'apprentissage non supervisé
- Très utilisées en datamining:
  - Découpage du marché en sous-ensembles dont les éléments réagissent de façon similaires aux variations des variables d'action du marché
- Faciles à implémenter et généralement disponibles dans les logiciels de datamining
- S'appliquent sur tout type de données (même textuelles)
- **Inconvénients:**
  - Choix des bons paramètres:
    - Par exemple: le choix du k et des centres initiaux pour l'algorithme de k-means
  - Les performances dépendent du choix de la mesure de similarité (distance) utilisée
  - L'interprétation des résultats

Exercice: CAH & K-means

- Nous disposons de la base de données suivante :

	A1	A2	A3	A4
D1	C1	R	A	P
D2	C2	V	B	G
D3	C3	V	B	G
D4	C1	J	B	P
D5	C2	R	A	P
D6	C3	J	A	G

1) Effectuez l'étape de codage et de normalisation nécessaire sur ces données en vue de leur appliquer une modélisation descriptive.
2) Quelle distance peut-on utiliser avec ces données? Expliquez.
3) Tout d'abord, nous allons appliquer la méthode de Classification Ascendante Hiérarchique comme méthode de regroupement de ces données. Construisez la hiérarchie (dendrogramme) ascendante correspondante à ces données en détaillant toutes les étapes de calcul. On utilise la méthode du lien simple pour le calcul de la distance entre les groupes.
4) Nous allons appliquer, par la suite, la méthode des K-Means pour trouver un regroupement par ressemblance de ces données. Classez donc ces dernières en **trois** groupes tout en utilisant les centres initiaux suivants : **D1, D3** et **D4**.

Exercice: CAH & K-means

1) Nous remarquons que les attributs sont nominaux; ce qui nécessite une étape de codage et de normalisation avant l'application d'une modélisation descriptive. Cette étape consiste à une représentation horizontale ou éclatée des données.

	A1	A2	A3	A4
D1	C1	R	A	P
D2	C2	V	B	G
D3	C3	V	B	G
D4	C1	J	B	P
D5	C2	R	A	P
D6	C3	J	A	G

	C1	C2	C3	R	V	J	A	B	P	G
D1	1	0	0	1	0	0	1	0	1	0
D2	0	1	0	0	1	0	0	1	0	1
D3	0	0	1	0	1	0	0	1	0	1
D4	1	0	0	0	0	1	0	1	1	0
D5	0	1	0	1	0	0	1	0	1	0
D6	0	0	1	0	0	1	1	0	0	1

2) Les données sont binaires asymétriques, donc la distance qu'on peut utiliser est le coefficient de Jaccard

$$d_{jc}(x_i, x_j) = \frac{b + c}{a + b + c}$$

	1	0
1	a	b
0	c	d

Exercice: CAH & K-means

3) Classification Ascendante Hiérarchique

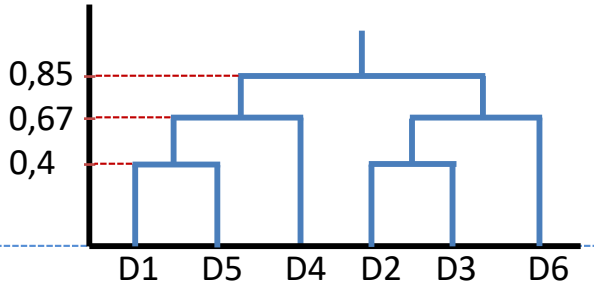
$d(D1,D2)=\frac{8}{8}=1$	$d(D2,D3)=\frac{2}{5}=0,4$	$d(D3,D5)=\frac{8}{8}=1$
$d(D1,D3)=\frac{8}{8}=1$	$d(D2,D4)=\frac{6}{7}=0,86$	$d(D3,D6)=\frac{4}{6}=0,67$
$d(D1,D4)=\frac{4}{6}=0,67$	$d(D2,D5)=\frac{6}{7}=0,86$	$d(D4,D5)=\frac{6}{7}=0,86$
$d(D1,D5)=\frac{2}{5}=0,4$	$d(D2,D6)=\frac{6}{7}=0,86$	$d(D4,D6)=\frac{6}{7}=0,86$
$d(D1,D6)=\frac{6}{7}=0,86$	$d(D3,D4)=\frac{6}{7}=0,86$	$d(D5,D6)=\frac{6}{7}=0,86$

	C1	C2	C3	R	V	J	A	B	P	G
D1	1	0	0	1	0	0	1	0	1	0
D2	0	1	0	0	1	0	0	1	0	1
D3	0	0	1	0	1	0	0	1	0	1
D4	1	0	0	0	0	1	0	1	1	0
D5	0	1	0	1	0	0	1	0	1	0
D6	0	0	1	0	0	1	1	0	0	1

	D1	D2	D3	D4	D5	D6
D1	0	1	1	0,67	**0,4**	0,86
D2		0	**0,4**	0,86	0,86	0,86
D3			0	0,86	1	0,67
D4				0	0,86	0,86
D5					0	0,86
D6						0

	{D1,D5}	{D2,D3}	D4	D6
{D1,D5}	0	0,86	**0,67**	0,86
{D2,D3}		0	0,86	**0,67**
D4			0	0,86
D6				0

	{{D1,D5},D4}	{{D2,D3}, D6}
{{D1,D5},D4}	0	**0,86**
{{D2,D3}, D6}		0



Exercice: CAH & K-means

4) K-Means avec les centres initiaux : **D1, D3** et **D4**

d(D1,D2)= $\frac{8}{8}$=1	d(D2,D3)= $\frac{2}{5}$=0,4	d(D3,D5)= $\frac{8}{8}$=1
d(D1,D3)= $\frac{8}{8}$=1	d(D2,D4)= $\frac{6}{7}$=0,86	d(D3,D6)= $\frac{4}{6}$=0,67
d(D1,D4)= $\frac{4}{6}$=0,67	d(D2,D5)= $\frac{6}{7}$=0,86	d(D4,D5)= $\frac{6}{7}$=0,86
d(D1,D5)= $\frac{2}{5}$=0,4	d(D2,D6)= $\frac{6}{7}$=0,86	d(D4,D6)= $\frac{6}{7}$=0,86
d(D1,D6)= $\frac{6}{7}$=0,86	d(D3,D4)= $\frac{6}{7}$=0,86	d(D5,D6)= $\frac{6}{7}$=0,86

	C1	C2	C3	R	V	J	A	B	P	G
D1	1	0	0	1	0	0	1	0	1	0
D5	0	1	0	1	0	0	1	0	1	0
G1	0	0	0	1	0	0	1	0	1	0
D2	0	1	0	0	1	0	0	1	0	1
D3	0	0	1	0	1	0	0	1	0	1
D6	0	0	1	0	0	1	1	0	0	1
G2	0	0	1	0	1	0	0	1	0	1
D4= G3	1	0	0	0	0	1	0	1	1	0

	D1	D2	D3	D4	D5	D6
D1=G1	**0**	1	1	0,67	**0,4**	0,86
D3=G2	1	**0,4**	**0**	0,86	1	**0,67**
D4=G3	0,67	0,86	0,86	**0**	0,86	0,86

C1={D1, D5}, G1=(0, 0, 0, 1, 0, 0, 1, 0, 1, 0)
C2={D2, D3, D6} , G2=(0, 0, 1, 0, 1, 0, 0, 1, 0, 1)=D3
C3={D4}, G3= D4

	D1	D2	D3	D4	D5	D6
G1	**0,25**	1	1	0,83	**0,25**	0,83
G2	1	**0,4**	**0**	0,86	1	**0,67**
G3	0,67	0,86	0,86	**0**	0,86	0,86

Les centres ne changent pas →Convergence