CLUSTERING

CLUSTERING

- Définition générale
  - Données initiales
    - Un ensemble d'objets ou individus à organiser
  - Objectif
    - Organiser l'ensemble des objets ou individus en un ensemble de clusters respectant les conditions suivantes :
      - La similarité entre deux objets ou individus appartenant à un même cluster doit être importante
      - La similarité entre deux objets ou individus appartenant à des clusters différents doit par contre être faible
  - Clustering = Classification non supervisée : pas de classes prédéfinies, à découvrir automatiquement

Applications

- 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'observation 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 :

Qualité d'un clustering

- Une bonne méthode de regroupement permet de garantir
  - Une grande similarité intra-groupe
  - Une faible similarité inter-groupe
- La qualité d'un regroupement dépend donc de la mesure de similarité utilisée par la méthode et de son implémentation
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "patterns" cachés.

Types de données

- Matrices de données

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Matrices de proximité ou individus x individus

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Mesurer la qualité d'un clustering

- Métrique pour la similarité: La similarité est exprimée par le biais d'une mesure de distance

- Les définitions de distance sont très différentes que les variables soient des intervalles (continues), catégories, booléennes ou ordinales

- En pratique, on utilise souvent une pondération des variables

Mesures de similarité

- Fonctions des types des variables descriptives
  - Variables continue sur un intervalle (ex: poids,taille)
  - Variables binaires
  - Variables nominales (ex: couleur)
  - Variables ordinales
- Problème posé
  - Comment prendre en compte des variables descriptives de différents types?

Variables numériques

- Standardiser les données (Normaliser)
  - Pour ne pas favoriser certaines variables par rapport à d'autres
  - Calculer l'écart absolu moyen:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

où

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$$

  - Calculer la mesure standardisée (*z-score*)

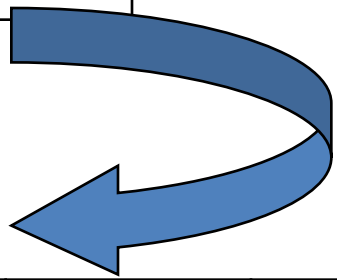
$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Exemple

	Age	Salaire
Personne1	50	11000
Personne2	70	11100
Personne3	60	11122
Personne4	60	11074

$$M_{Age} = 60 \qquad S_{Age} = 5$$

$$M_{salaire} = 11074 \qquad S_{salaire} = 148$$



	Age	Salaire
Personne1	-2	- 0,5
Personne2	2	0,175
Personne3	0	0,324
Personne4	0	2

Variables numériques

- Les distances expriment une similarité
- Comment calculer une distance?
  - Par l'utilisation d'une distance de Minkowski
- la *distance de Minkowski* :

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q)}$$

où $i = (x_{i1}, x_{i2}, ..., x_{ip})$ et $j = (x_{j1}, x_{j2}, ..., x_{jp})$ sont deux objets p-dimensionnels et q un entier positif

- Si $q = 1$, d est la distance de Manhattan

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

Similarité entre objets(I)

- *Si q = 2, d est la distance Euclidienne* :

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

  - Propriétés
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$

Exemple : distance de Manhattan

	Age	Salaire
Personne1	50	11000
Personne2	70	11100
Personne3	60	11122
Personne4	60	11074

d(p1,p2)=120

d(p1,p3)=132

Conclusion: p1 ressemble plus à p2 qu'à p3 ?

	Age	Salaire
Personne1	-2	-0,5
Personne2	2	0,175
Personne3	0	0,324
Personne4	0	2

d(p1,p2)=4,675

d(p1,p3)=2,324

Conclusion: p1 ressemble plus à p3 qu'à p2 !!

Variables binaires

- Différents types
  - Symétrique si les deux états possibles sont équivalents
    - Cas d'une variable indiquant le sexe d'un individu
  - Asymétrique si un état est plus important qu'un autre : Ex. Test HIV. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre. Généralement, on code par 1 la modalité la moins fréquente
    - 2 personnes ayant la valeur 1 pour le test sont *plus similaires* que 2 personnes ayant 0 pour le test

Variables binaires

- Une table de contingence pour données binaires

Objet *i* \ Objet *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 i a 1 et j a 1

- Exemple o_i=(1,1,0,1,0) et o_j=(1,0,0,0,1)

a=1, b=2, c=1, d=1

Mesures de distances

- Coefficient d'appariement (matching) simple (Similarité invariante si la variable binaire est symétriques):

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

Exemple o_i=(1,1,0,1,0) et o_j=(1,0,0,0,1)

$d(o_i, o_j)$=3/5

- Coefficient de Jaccard (Similarité non invariante si la variable binaire est asymétriques):

$$d(i,j) = \frac{b+c}{a+b+c}$$

$d(o_i, o_j)$=3/4

Variables binaires

- Exemple

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Ibra	M	Y	P	N	N	N	N

- Sexe est un attribut symétrique
- Les autres attributs sont asymétriques
- Y et P ≡ 1, N ≡ 0, la distance n'est mesurée que sur les asymétriques : calcul de la distance basée sur le Coefficient de Jaccard

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, Ibra) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Ibra, mary) = \frac{1+2}{1+1+2} = 0.75$$

Les plus similaires sont Jack et Mary⇒atteints du même mal

Variables nominales

- Généralisation des variables binaires avec plus de 2 états:
  - Une variable nominale peut prendre une valeur parmi « k » possibles avec k>2
    - Couleur : rouge, vert , bleu, …
- Comment calculer une distance?
  - Méthode 1: Matching simple
    - *Si deux individus ont les mêmes caractéristiques selon « m » variables nominales parmi « p », alors :*
    - *m=nombre de correspondances, p=nombre total de variables*

$$d(i,j) = \frac{p-m}{p}$$

- Méthode 2: utiliser un grand nombre de variables binaires
  - Créer une variable binaire pour chaque modalité (ex: variable rouge qui prend les valeurs vrai ou faux)

Variables ordinales

- Différence par rapport aux variables nominales
  - Un ordre peut être défini sur l'ensemble des valeurs possibles de la variable en question
    - Exemple : *Niveau ∈ { Insuffisant, Passable, Bien, Très Bien} avec Insuffisant < Passable < Bien < Très Bien*
- Comment calculer une distance?
  - Même méthode que celle utilisée pour les variables numériques
  - Remplacer chaque valeur « x_{ij} » par son rang « r_{ij} » dans l'ordre défini sur l'ensemble des valeurs
    - *Exemple : Insuffisant = 1, Passable = 2, Bien = 3, Très Bien = 4*
  - Remplacer toute variable « x_j » avec « k_j » modalités par une nouvelle variable « y_j »

$$y_{ij} = \frac{r_{ij} - 1}{kj - 1}$$

  - Utiliser une distance pour calculer la similarité

Mesures de similarité

- Prise en compte de variables de types différents
  - Dans une même mesure de similarité

$$d(i,j) = \frac{\sum_{k=1}^{p} \delta_k^{(i,j)} d_k^{(i,j)}}{\sum_{k=1}^{p} \delta_k^{(i,j)}}$$

  - Avec $\delta_k^{(i,j)}$ = 0 ou 1, $\delta_k^{(i,j)}$ = 0 dans les cas suivants
    - « x_{ik} » ou « x_{jk} » sont inconnus
    - « $x_{ik} = x_{jk} = 0$ » et la variable « k » est binaire asymétrique
  - Avec $d_k^{(i,j)}$ calculé en fonction du type de la variable « x_k »
    - Binaire ou nominale
      - $d_k^{(i,j)}$ = 0 ssi « $x_{ik} = x_{jk}$ »
    - Numérique ou ordinale
      - $d_k^{(i,j)} = | x_{ik} - x_{jk} | / (\max_h(x_{hk}) - \min_h(x_{hk}))$

Mesures de similarité : exemple

- soit les deux individus suivants :

  - Jean : age = 30 ans, poids = 80kg, sexe = masculin, actif = non, nationalité = française
  - Marie : age = 28 ans, poids = ?, sexe = féminin, actif = non, nationalisé = allemande

sachant que :

- le ? signifie que la donnée est inconnue,
- que les variables binaires 'sexe' et 'actif' sont considérées comme asymétriques (ce qui dépend de l'application considérée),
- que masculin est codé par la valeur 1 (et féminin par la valeur 0),
- que pour la variable actif, "oui" est codée par la valeur 1 et non par la valeur 0,
- que pour les variables 'age' et 'poids' les valeurs minimale et maximale sont respectivement 0 et 100

Mesures de similarité : exemple

Dans ce cadre, nous avons :

- d_age (Jean,Marie) = 30 - 28 / 100 - 0 = 0,02 et delta_age(Jean, Marie) = 1,
- d_poids(Jean,Marie) est indéterminée (on supposera nul), car le poids de Marie est inconnue, et donc delta_poids(Jean,Marie) = 0,
- d_sexe(Jean,Marie) = 1 car leur sexe sont différents et delta(Jean,Marie) = 1
(car le sexe de Jean est associée à la valeur 1),
- d_actif(Jean, Marie) = 0 (même valeur) et delta_actif(Jean,Marie) = 0 car à la fois pour Jean et Marie, la valeur associée à la variable actif est nulle (non est codé par 0).
- d_nationalité(Jean,Marie) = 1 (car leurs nationalités sont différentes) et delta_nationalité(Jean,Marie) = 1

En appliquant la formule, on obtient finalement :

- d(Jean,Marie) = (1*0,02 + 0*0 + 1*1 + 0*0 + 1*1) / (1 + 0 + 1 + 0 + 1) = 2,02 / 3, soit environ 0,66

Approches de clustering

- Méthodes de partitionnement
  - Construire plusieurs partitions puis les évaluer selon certains critères
- Méthodes hiérarchiques
  - Créer une décomposition hiérarchique des objets selon certains critères
  - Par division ou agglomération successive des données
- Méthodes basées sur la densité
  - Par recherche de sous-ensembles denses de données

Méthodes de partitionnement

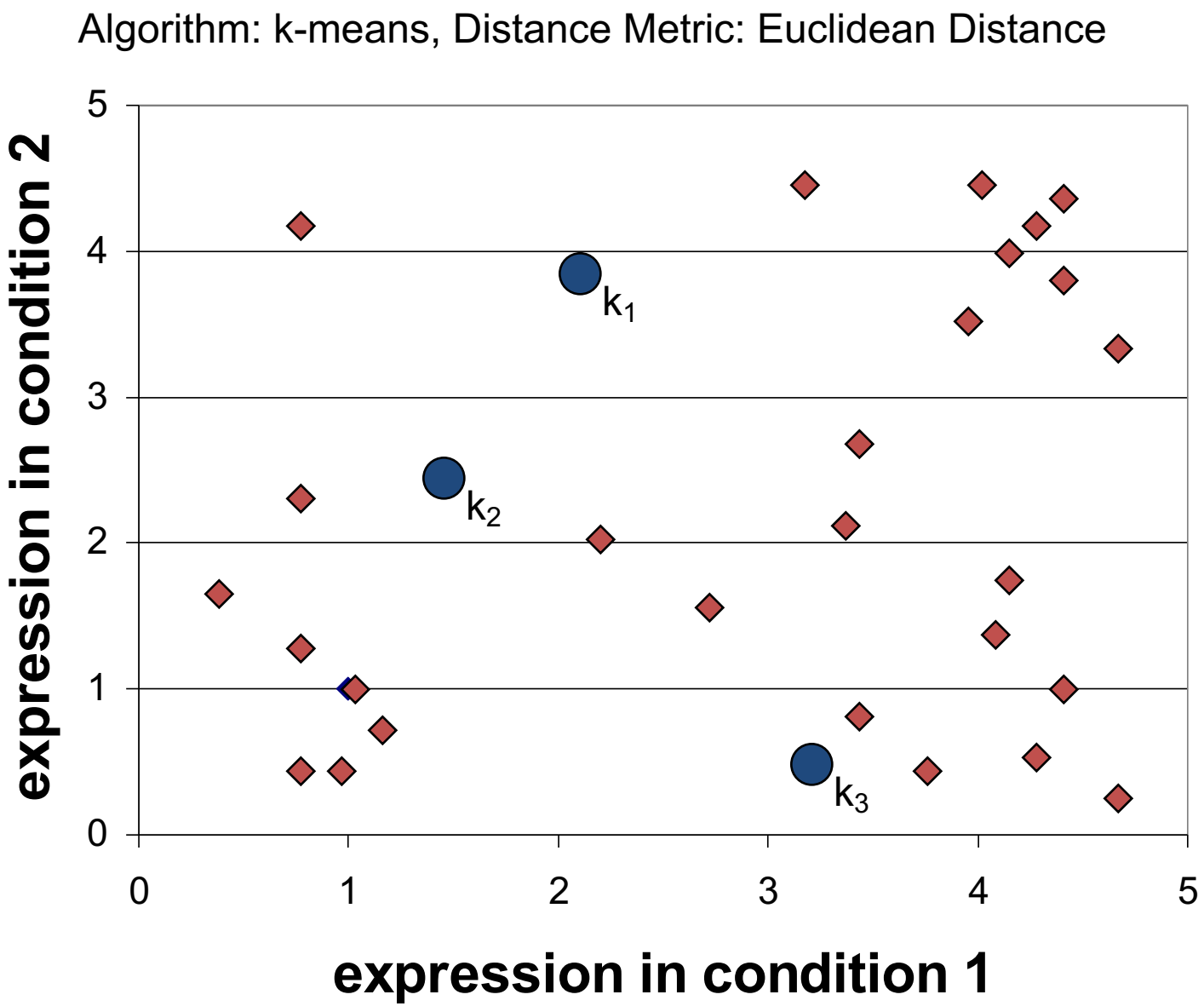
- Problème posé
  - Comment construire une partition en K classes d'un ensemble de N objets ou individus ?
    - Etant donné le nombre K de partitions à déterminer
    - Comment déterminer la partition dont la qualité est la plus grande?
      - Minimiser les distances intra-classes (à l'intérieur des différentes classes)
      - Maximiser les distances inter-classes (entre les différentes classes)
- Solutions proposées
  - **Méthode exacte**
    - Par évaluation de toutes les partitions possibles
  - **Méthodes approchées**
    - **K-means** (MacQueen'67) : classe représentée par son centre de gravité
    - **K-medoids** (Kaufman & al'87) : classe représentée par son élément le plus représentatif

Méthode des k-means

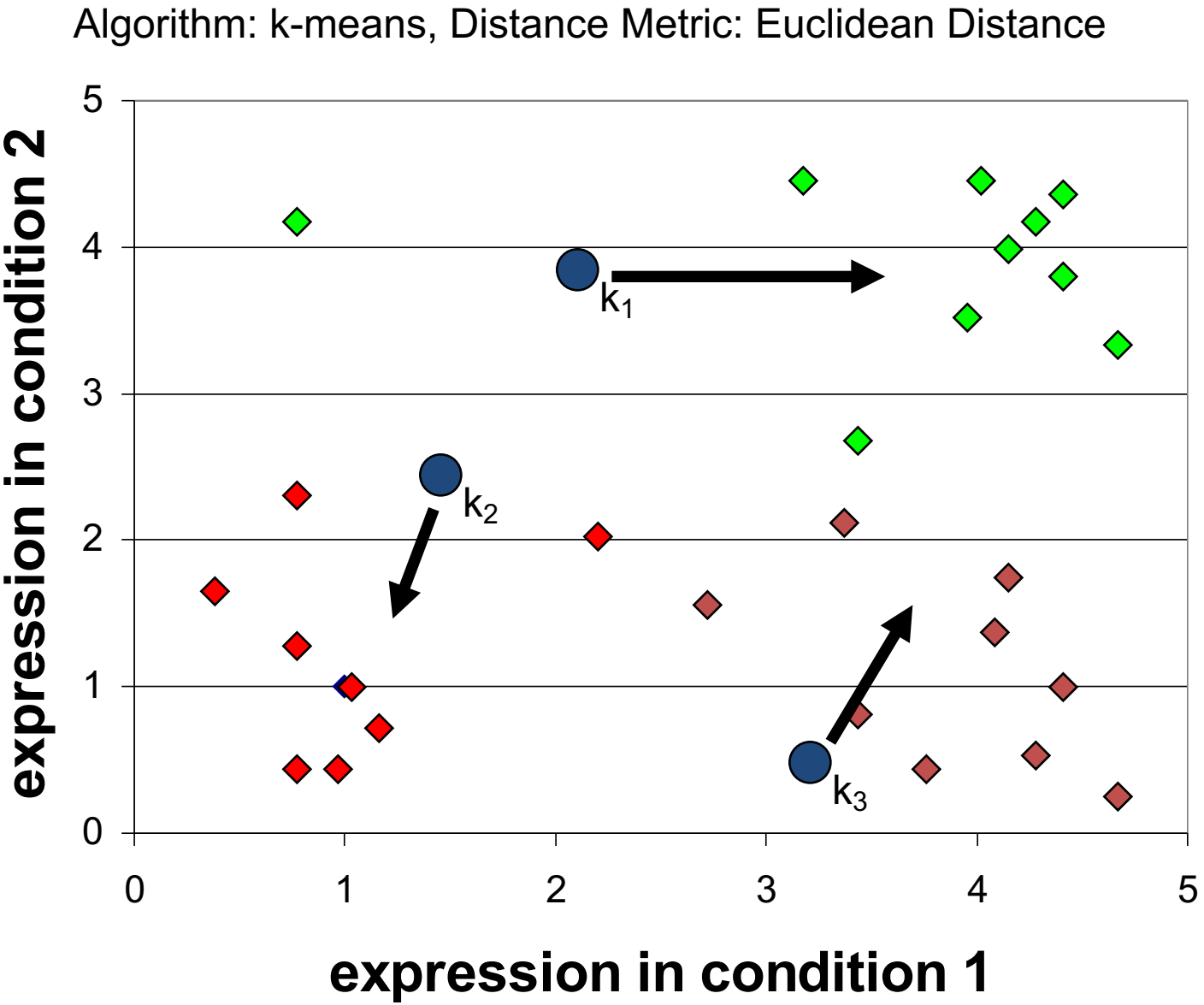
- Principe de base
  - Une classe est représentée par son centre de gravité
  - Un objet O_i appartient à la classe du centre de gravité le plus proche de O_i

1. **Découper les données** en K sous-ensembles P_1, P_2, …, P_K
2. **Calculer les centres de gravité** C_1, C_2, …, C_K de chaque sous-ensemble P_1, P_2, …, P_K
3. **Mettre à jour les sous-ensembles** P_1, P_2, …, P_K tel que pour tout objet O_i, $O_i \in P_j$ si C_j est le centre de gravité le plus proche de O_i
4. **Retourner à l'étape (2)** si un ou plusieurs sous-ensembles P_j ont été modifiés à l'étape (3)

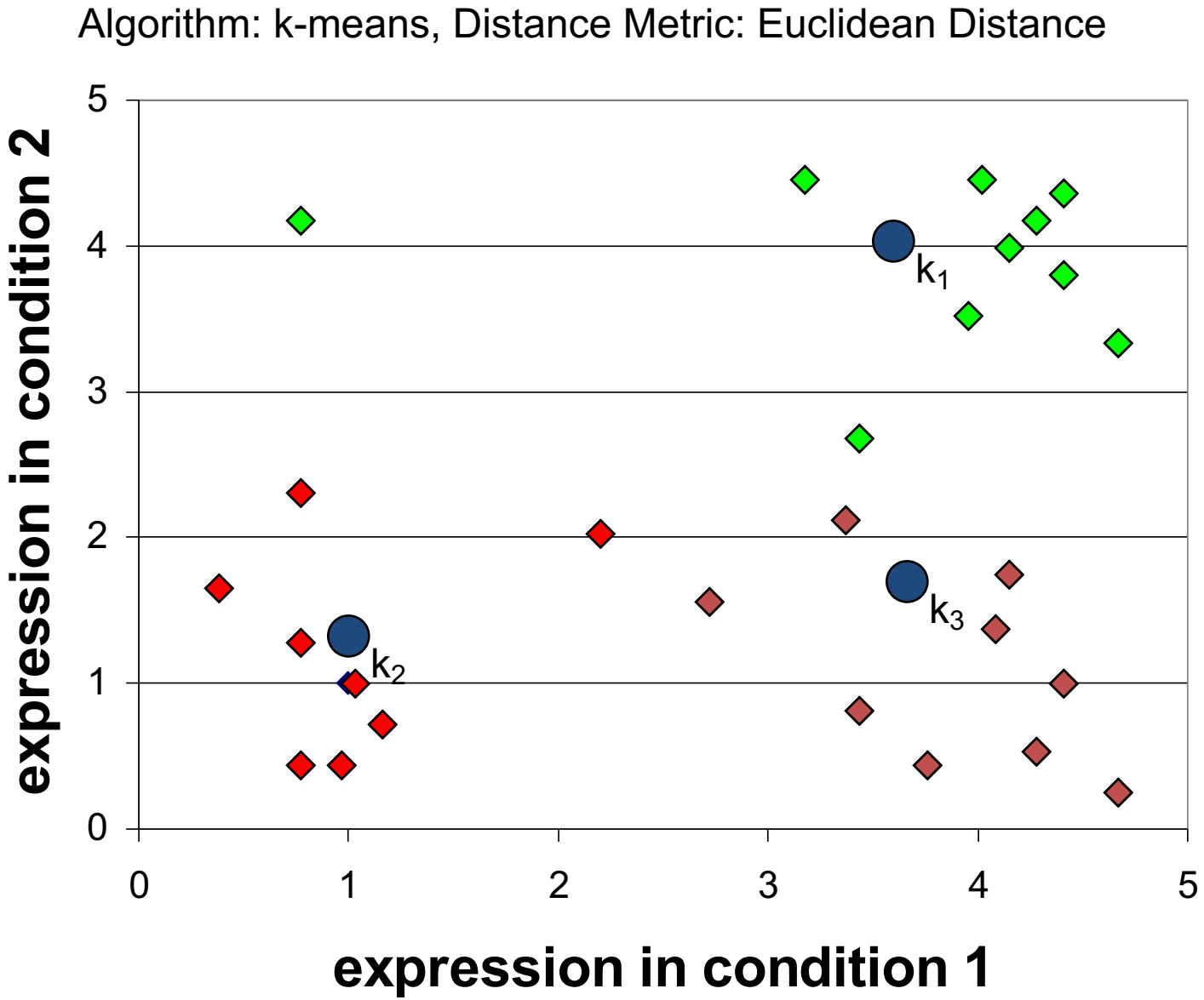
Méthode des k-means



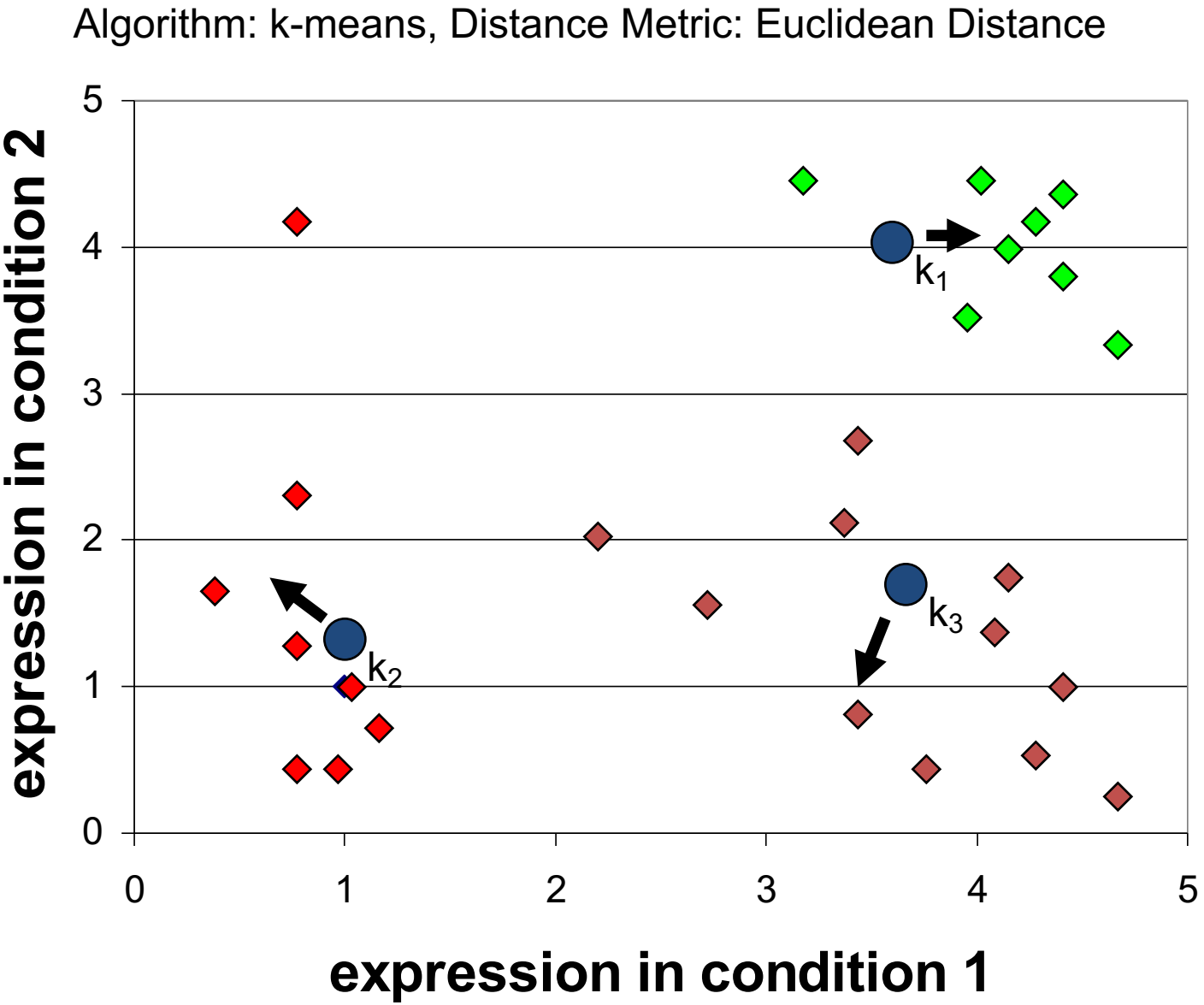
Méthode des k-means



Méthode des k-means

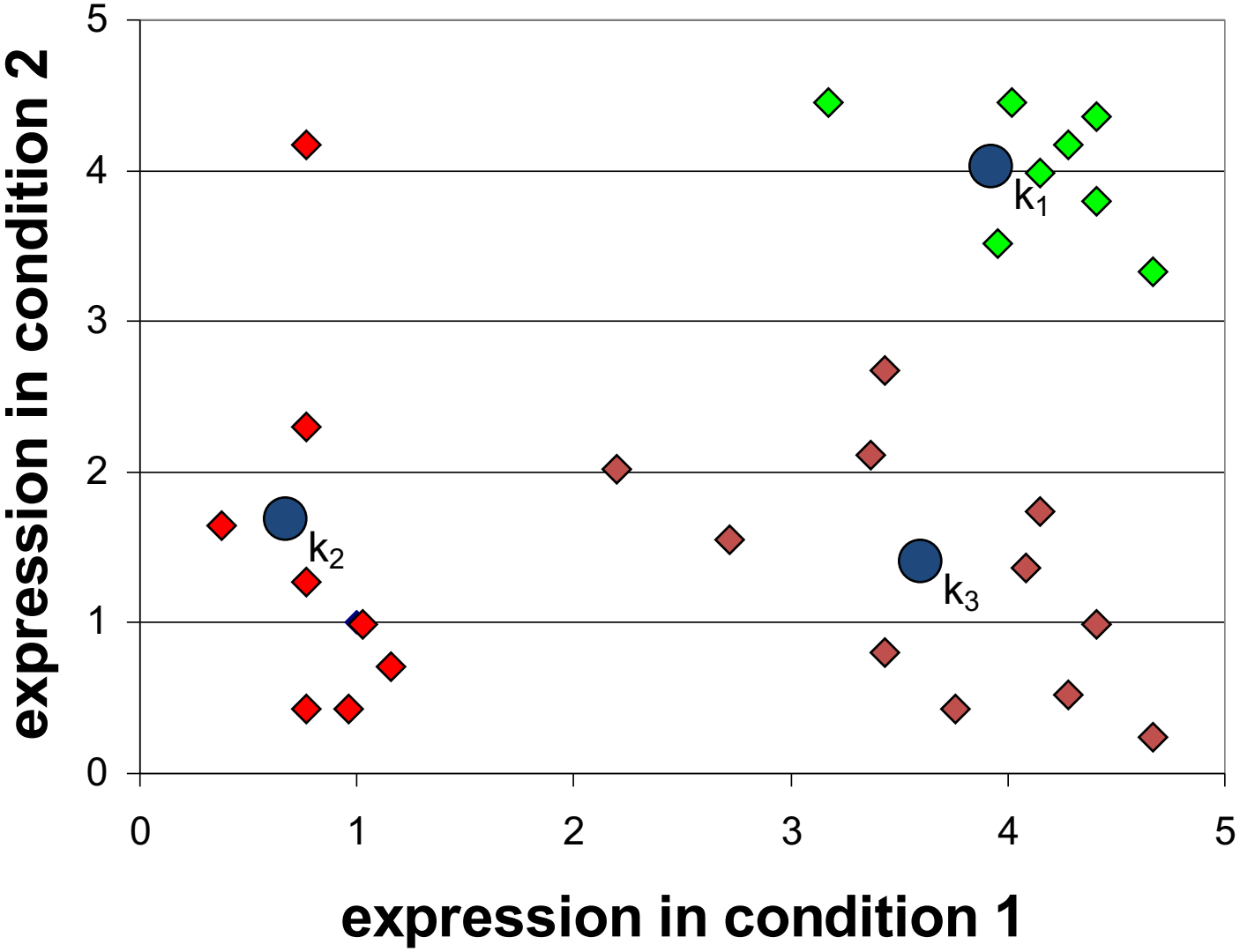


Méthode des k-means



Méthode des k-means

Algorithm: k-means, Distance Metric: Euclidean Distance



K-Means :Exemple

- 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. C_1={1}, M_1=1, C_2={2}, M_2=2, C_3={3} et M_3=3

- Chaque objet O est affecté au cluster au milieu duquel, O est le plus proche. 6 est affecté à C_3 car dist(M_3,6)<dist(M_2,6) et dist(M_3,6)<dist(M_1,6)

  On a C_1={1}, M_1=1,

  C_2={2}, M_2=2

  C_3={3, 6,7,8,13,15,17}, M3=69/7=9.86

K-Means :Exemple (suite)

- dist(3,M2)<dist(3,M3)➔3 passe dans C_2. Tous les autres objets ne bougent pas. C_1={1}, M_1=1, C_2={2,3}, M_2=2.5,C_3={6,7,8,13,15,17} et M_3= 66/6=11

- dist(6,M_2)<dist(6,M_3)➔6 passe dans C2. Tous les autres objets ne bougent pas. C_1={1}, M_1=1, C_2={2,3,6}, M_2=11/3=3.67, C3={7,8,13,15,17}, M_3= 12

- dist(2,M1)<dist(2,M2)➔2 passe en C1. dist(7,M2)<dist(7,M3)➔ 7 passe en C2. Les autres ne bougent pas. C1={1,2}, M1=1.5, C2={3,6,7}, M2=5.34, C3= {8,13,15,17}, M3=13.25

- dist(3,M1)<dist(3,M2)➔3 passe en 1. dist(8,M2)<dist(8,M3)➔8 passe en 2 C1={1,2,3}, M1=2, C2={6,7,8}, M2=7, C3={13,15,17}, M3=15

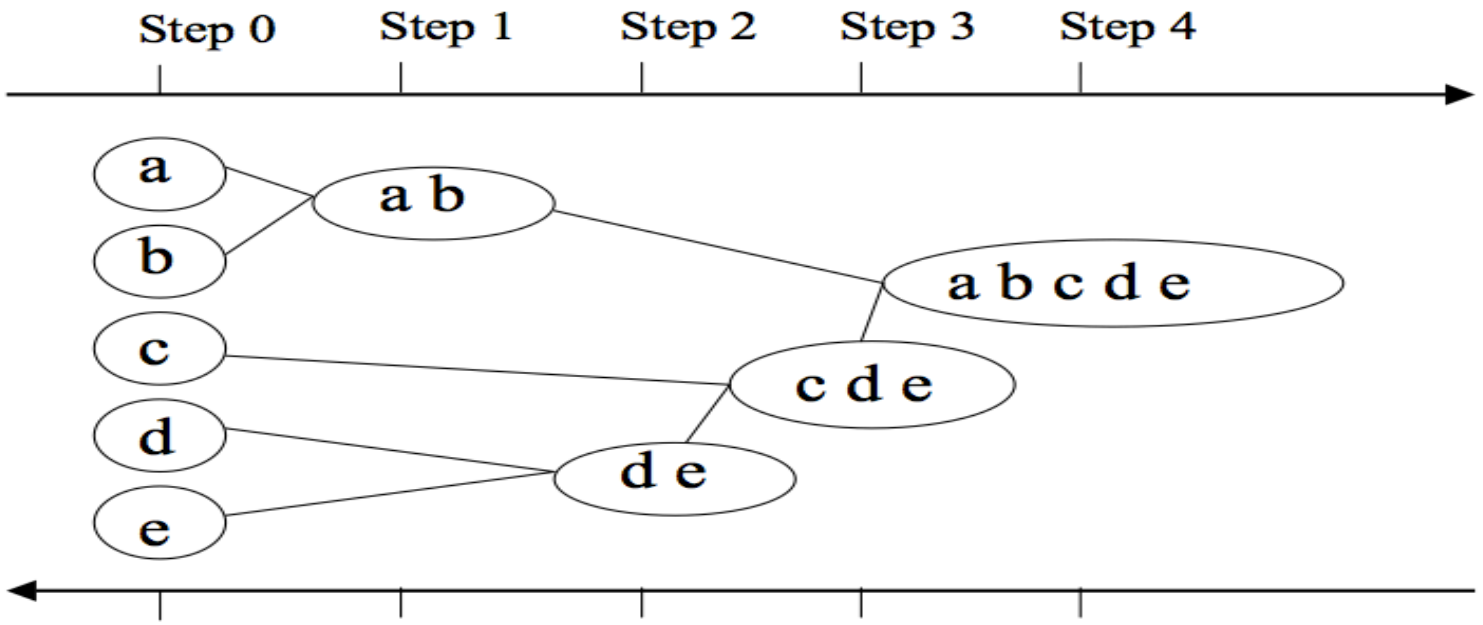
Plus rien ne bouge

K-means : enjeux

- Forces
  - *Relativement efficace*
  - Passage à l'échelle : *complexité linéaire en nombre d'objets*
- Faiblesses
  - N'est pas applicable en présence d'attributs qui ne sont pas du type intervalle (moyenne=?)
  - On doit spécifier k (nombre de clusters)
  - Les clusters sont construits par rapports à des objets inexistants (les milieux)

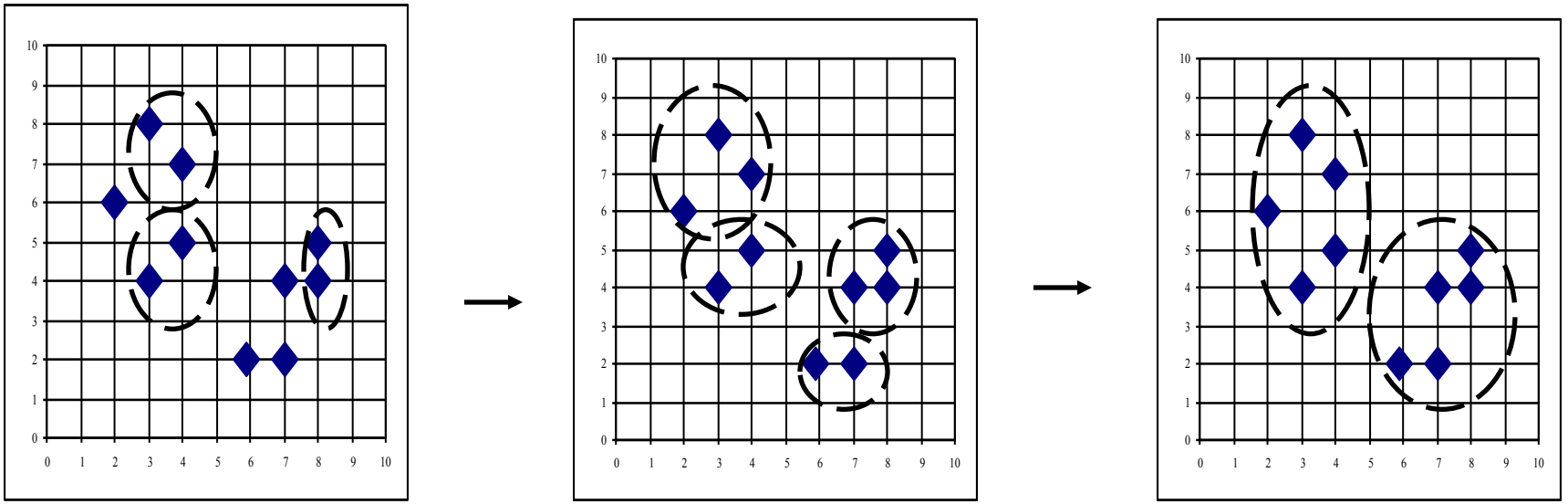
Méthodes hiérarchiques

1. On commence avec m clusters (cluster = 1 enregistrement)
2. Grouper les deux clusters les plus « proches ».
3. S'arrêter lorsque tous les enregistrements sont membres d'un seul groupe
4. Aller en 2.



AGNES (Agglomerative Nesting)

- Principe de base
  - Introduit par Kaufmann et Rousseeuw (1990)
  - Au départ : un objet = une classe
  - A chaque étape : regroupement des classes les plus proches
  - On peut se retrouver dans la situation où tous les nœuds sont dans le même groupe



Méthodes hiérarchiques

Exemple

Variables	Individus						
	A	B	C	D	E	F	G
V1	3	4	4	2	6	7	6
V2	2	5	7	7	6	7	4

Méthodes hiérarchiques
Matrice des distances Euclidiennes pour l'exemple

	A	B	C	D	E	F	G
A	.	3.162277660	5.099019514	5.099019514	5.000000000	6.403124237	3.605551275
B	3.162277660	.	2.000000000	2.828427125	2.236067977	3.605551275	2.236067977
C	5.099019514	2.000000000	.	2.000000000	2.236067977	3.000000000	3.605551275
D	5.099019514	2.828427125	2.000000000	.	4.123105626	5.000000000	5.000000000
E	5.000000000	2.236067977	2.236067977	4.123105626	.	1.414213562	2.000000000
F	6.403124237	3.605551275	3.000000000	5.000000000	1.414213562	.	3.162277660
G	3.605551275	2.236067977	3.605551275	5.000000000	2.000000000	3.162277660	.

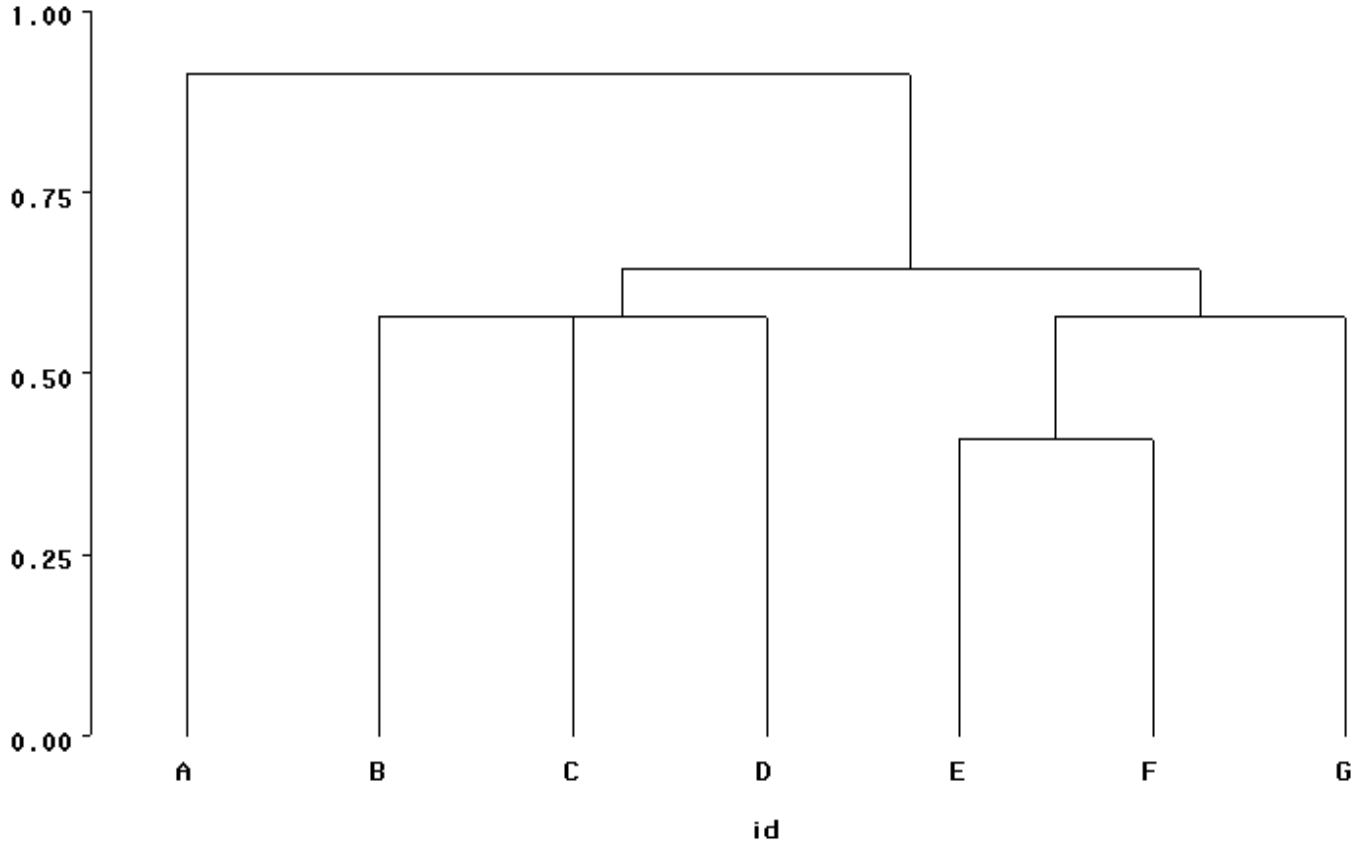
Méthodes hiérarchiques

Formation des groupes

- Solution initiale: chaque individu forme un groupe.
- Première étape: distance minimum= 1,414; les individus E et F sont regroupés.
- Deuxième étape: distance minimum entre les points de groupes différents= 2,0 entre les points B-C, C-D et E-G.
  - G est regroupé avec E et F.
  - B et C sont regroupés.
  - D est regroupé avec B et C.
- Nous avons maintenant 3 groupes: (A), (B C D), (E F G).
  - Distance minimum entre les points de groupes différents= 2,236 entre B-E et C-E.
  - Les groupes (B C D) et (E F G) sont regroupés.
- Nous avons maintenant 2 groupes: (A), (B C D E F G).
  - Distance minimum = 3,162 entre A-B.
  - Étape finale: tous les points sont regroupés en un seul groupe.

Méthodes hiérarchiques

Représentation graphique (dendogram)



Méthodes hiérarchiques : enjeux

- Points faibles des méthodes hiérarchiques
  - Passage à l'échelle difficile : *complexité quadratique*
  - Pas de remise en question possible des divisions ou agglomérations
    - Deux classes agglomérées à un niveau ne peuvent pas être séparées à un autre niveau
- Recherches en cours
  - Algorithme BIRCH (1996) : basée sur la représentation d'une classe par ses traits caractéristiques

Méthodes basées sur la densité

- Principe de base
  - Le regroupement d'objets au sein d'une même classe est basé sur un critère local évaluant la densité d'objets en un point donné de l'espace d'entrée
- Caractéristiques de ces méthodes
  - Découvre des classes de formes quelconques
  - Ne sont pas sensibles à la présence de bruit ou points isolés
  - Nécessite seulement un parcours des données
- Différentes propositions
  - **DBSCAN** : *introduite par Ester, et al. (KDD '96)*