ENICARTHAGE
المدرسة الوطنية للمهندسين بقرطاج
Ecole Nationale d'Ingénieurs de Carthage

Année Universitaire 2020-2021

Analyse de données

Analyse de données

Les méthodes de classification non supervisées (K_Means, CAH, classification mixte)

Introduction

La classification a pour but de regrouper des individus en classes homogènes en fonction de l'étude de certaines caractéristiques des individus. Par classes homogènes, on entend regrouper les individus qui se ressemblent et séparer ceux qui sont éloignés.

Il y a alors deux approches distinctes :

– La classification automatique, qui fonctionne selon des algorithmes formalisés ;

– La classification subjective. Celle-ci est effectuée par les praticiens, en fonction de leurs études qualitatives et de leurs intuitions.

Comme souvent dans l'analyse de données, les meilleures solutions se trouveront dans une combinaison des deux approches. Dans ce cours, nous aborderons uniquement la classification automatique.

Introduction

La classification automatique se divise en deux catégories :

- La classification automatique hiérarchique : il s'agit d'effectuer une partition de classes de plus en plus vaste (classification hiérarchique ascendante) ou de moins en moins vaste (classification automatique descendante). Nous développerons l'algorithme de la classification hiérarchique ascendante.

- La classification automatique non-hiérarchique. Dans ce cas, le nombre de classes de la partition est fixé en avance. La méthode des centres mobiles illustrera cette approche. La généralisation de cette méthode, que nous n'introduirons pas ici, conduit à la méthode des nuées dynamiques.

I- Les Notions

◆ Indice de dissimilarité

Soit E l'ensemble des n objets à classer. Une dissimilarité **d** est une application de E×E dans R+ telle que :

1. d (i, i) = 0 $\quad \forall i \in E$

2. d (i, i') = d(i', i) $\quad \forall i, i' \in E \times E$

Une distance satisfait les propriétés d'un indice de dissimilarité.

◆ Matrice des distances

Pour un nuage d'individus, on peut résumer l'ensemble des distances entre individus au sein d'une matrice des distances que l'on note D. Chaque coefficient d_{ij} représente la distance entre l'individu M_i et l'individu M_j . Par exemple, si l'on choisit comme critère de ressemblance la distance euclidienne, on a $d_{ij} = d(M_i,M_j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
Avec deux points (M_1,M_2) qui ont 2 variables uniquement : (x_1, y_1) et (x_2, y_2).

Une matrice de distances est une matrice carré, symétrique ($d_{ij} = d_{ji}$), de coefficients positifs ($d_{ij} \geq 0$) et de coefficients nuls sur la diagonale ($d_{ii} = d(M_i,M_i) = 0$).

I- Les Notions

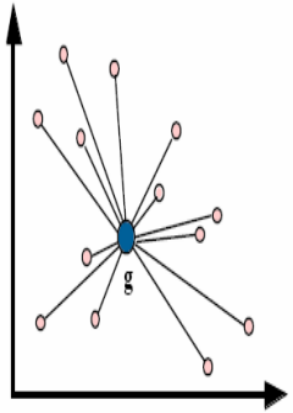
- ***La notion d'Inertie***

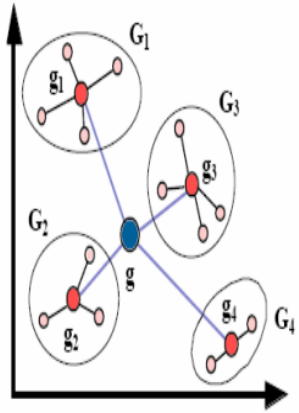
Soit une classification en k groupes d'effectifs n_1, ... ,n_k, les individus étant des points d'un espace euclidien. Notons les groupes G_1, ... ,G_k, et g_1, ... ,g_k leurs centres de gravité (gest le centre de gravité du nuage).

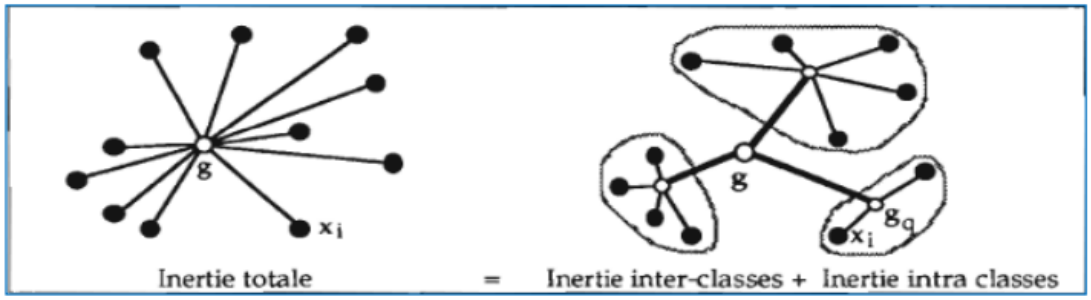
Inertie totale : $I_{tot} = \frac{1}{n} \sum_{i=1}^{n} d^2(e_i, g)$

Inertie interclasse : $I_{inter} = \frac{1}{n} \sum_{i=1}^{k} n_i \cdot d^2(g_i, g)$

Inertie intraclasse : $I_{intra} = \frac{1}{n} \sum_{i=1}^{k} \sum_{e \in G_i} d^2(e, g_i)$







I- Les Notions

Une partition pour être bonne doit satisfaire les deux critères suivants :

– Les individus proches doivent être regroupés : chaque classe doit être le plus homogène possible.

– Les individus éloignés doivent être séparé : les classes de la partition doivent être éloignées les unes des autres.

◆ L'inertie est une mesure de l'homogénéité d'un ensemble de points (nuage ou classe). Une classe (ou un nuage) sera d'autant plus homogène que son inertie totale sera faible.

◆ L'inertie intraclasse mesure l'homogénéité de l'ensemble des classes. Plus l'inertie intraclasse est faible, plus la partition est composée de classes homogènes.

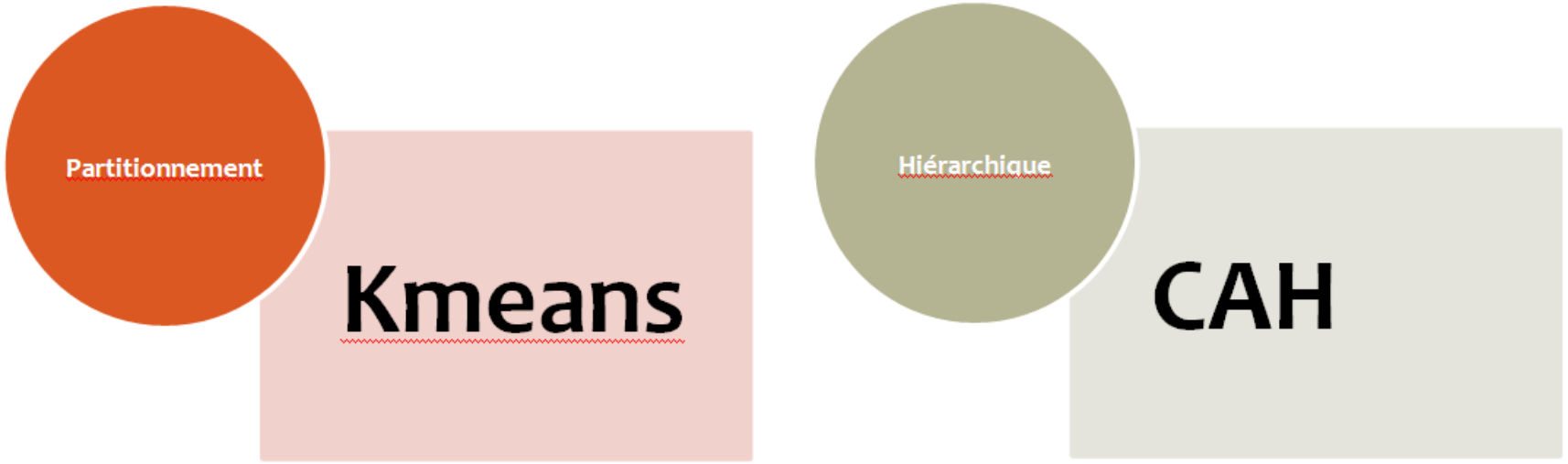
◆ L'inertie interclasse mesure la séparation entre les classes d'une partition. Plus l'inertie interclasse est grande plus les classes sont distinctement séparées.

Théorème de Huygens :

Inertie totale = Inertie inter-classe + Inertie intra-classe

$$I_{tot} = I_{inter} + I_{intra}$$

Choix de la méthode de classification



II- La classification K-means (Agrégation autour des centres mobiles)

- ✓ L'algorithme des K-moyennes permet de trouver des classes dans des données.
- ✓ les classes qu'il construit n'entretiennent jamais de relations hiérarchiques: une classe n'est jamais incluse dans une autre classe
- ✓ L'algorithme fonctionne en précisant le nombre de classes attendues.
- ✓ L'algorithme calcule les distances **Intra-Classe** et **Inter-Classe**.
- ✓ Il travaille sur des **variables continues**.

Principe Algorithmique

Algorithme K-Means

Entrée : *k le nombre de groupes cherchés*

DEBUT

Choisir aléatoirement les centres des groupes

REPETER

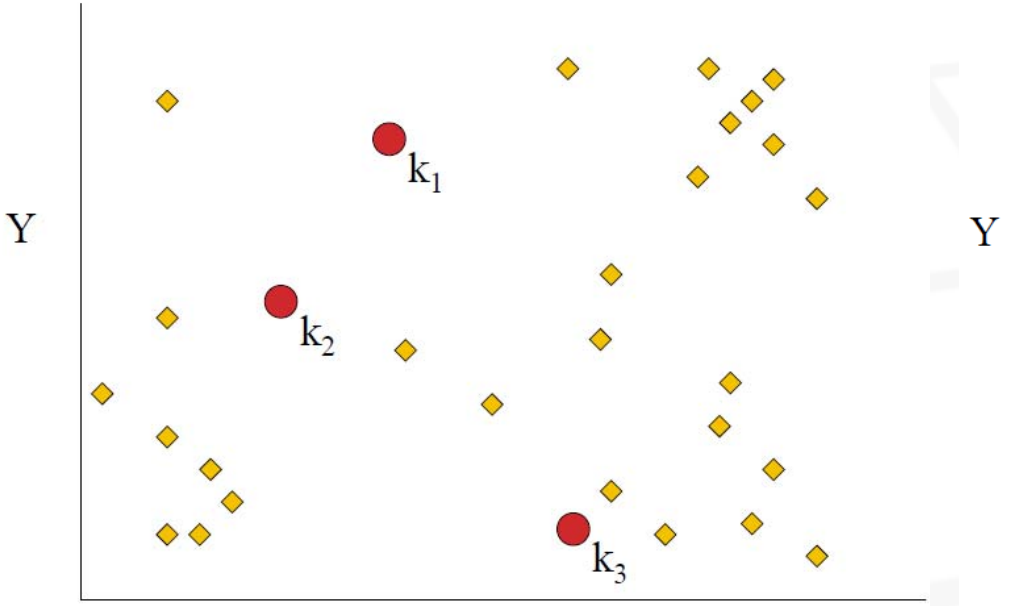
i. ***Affecter chaque cas au groupe dont il est le plus proche au son centre***
ii. ***Recalculer le centre de chaque groupe***

JUSQU'A *(stabilisation des **centres**)*

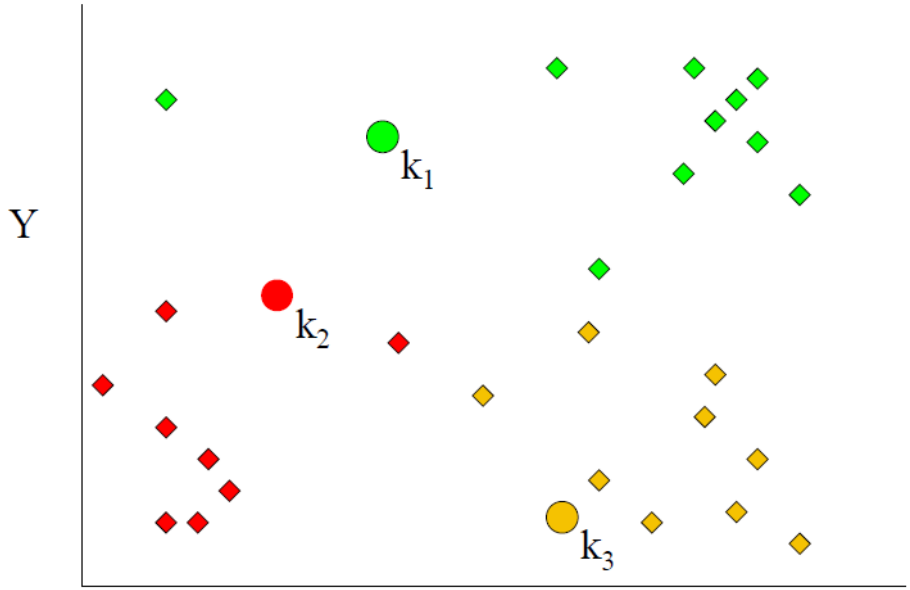
OU *(nombre d'itérations =**t**)*

FIN

II- La classification K-means (Agrégation autour des centres mobiles)

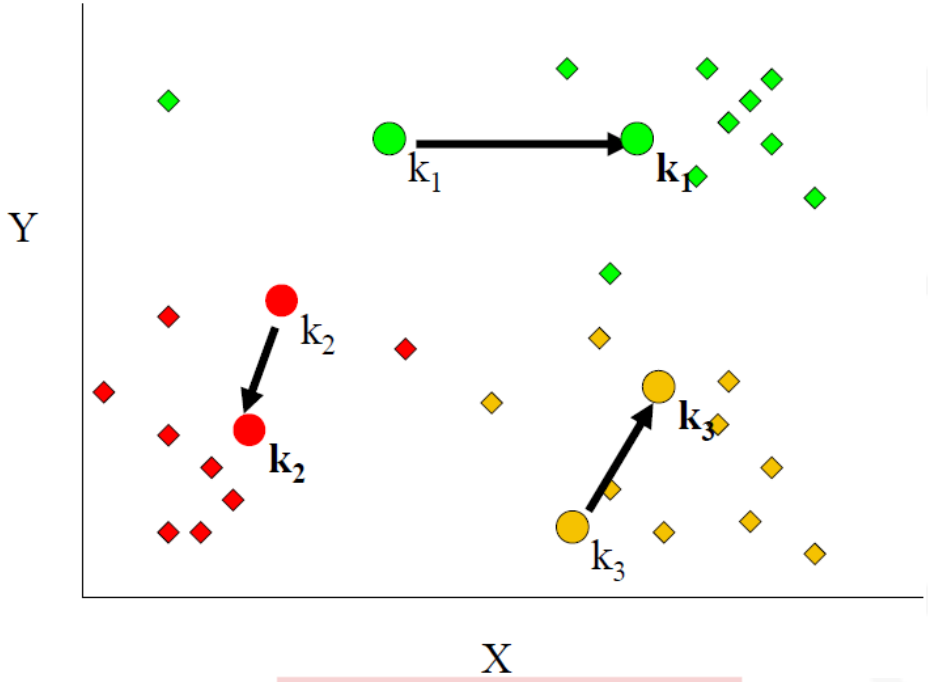


Choisir **3** Centres de classes (au hasard)

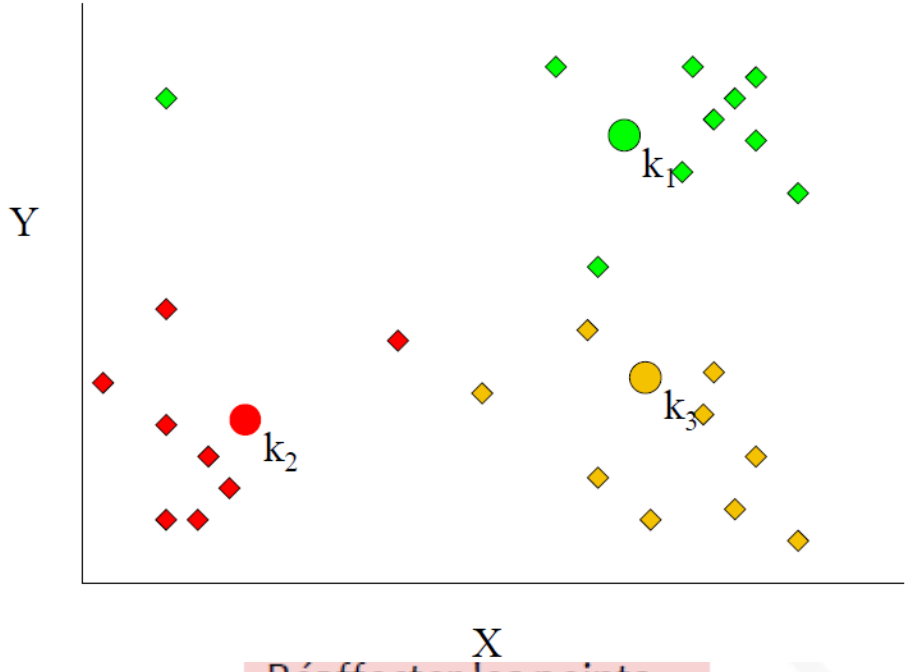


Affecter chaque point à la classe dont le centre est le plus proche

II- La classification K-means (Agrégation autour des centres mobiles)

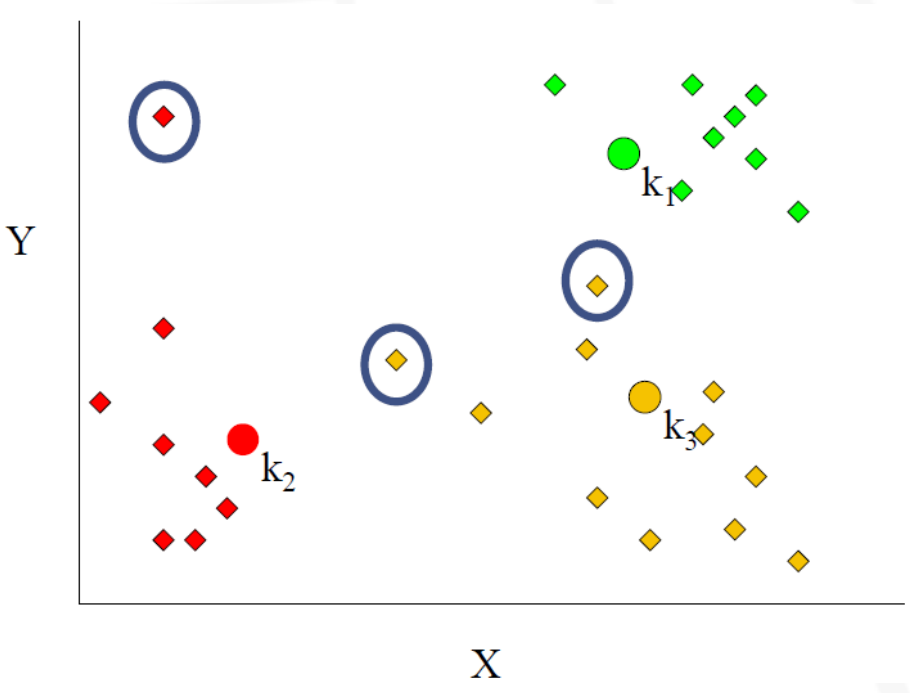


Déplacer chaque centre de classe vers la moyenne de chaque classe

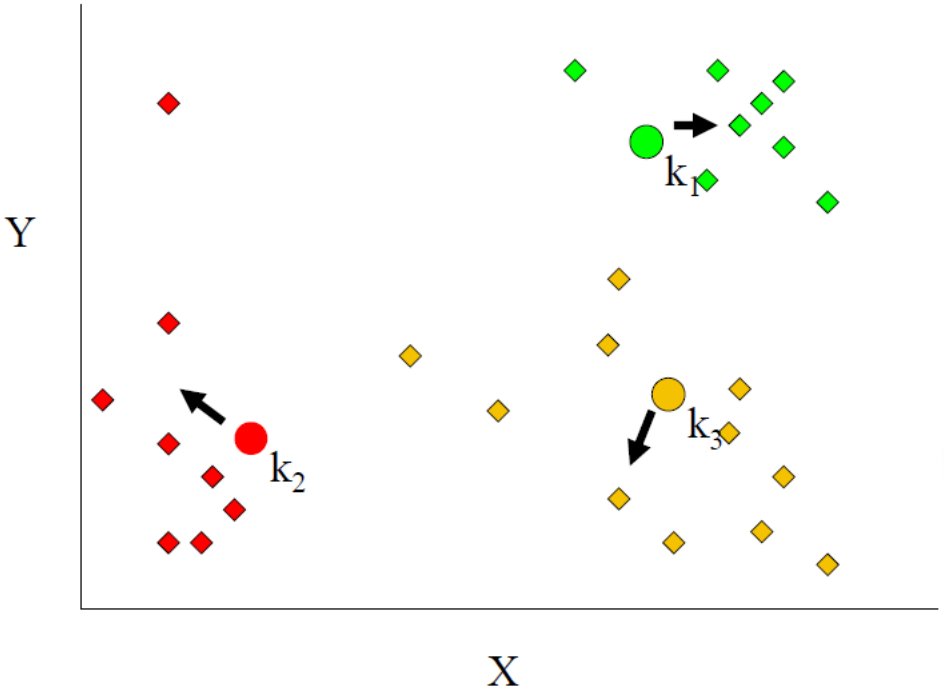


Réaffecter les points qui sont plus proches du centre d'une autre classe

II- La classification K-means (Agrégation autour des centres mobiles)

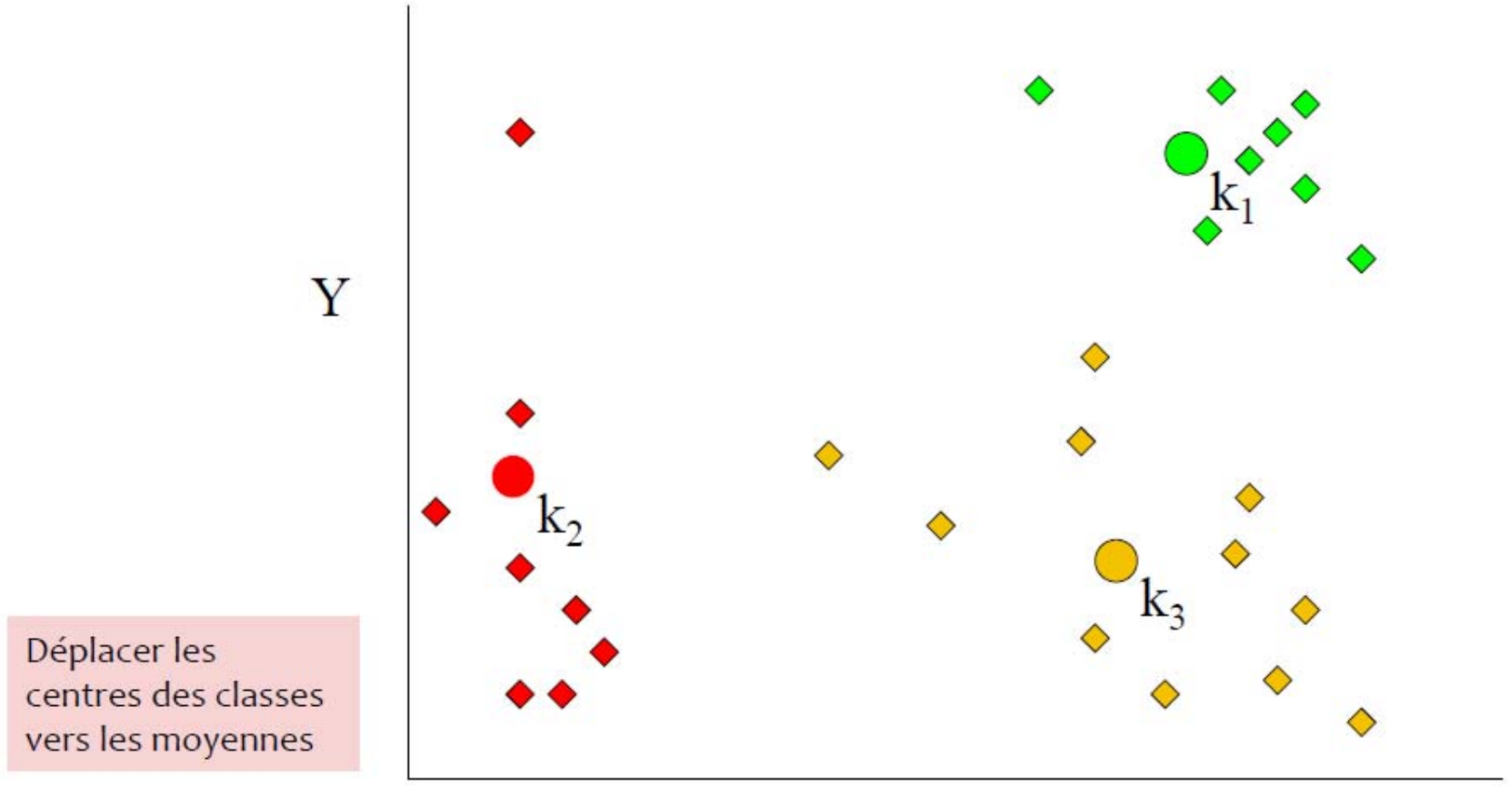


les trois points qui changent de classe



Re-calculer les moyennes des classes

II- La classification K-means (Agrégation autour des centres mobiles)



II- La classification K-means (Agrégation autour des centres mobiles)

Le processus se stabilise nécessairement et l'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition, soit lorsqu'un critère convenablement choisi (par exemple, la mesure de la variance intra-classes) cesse de décroître de façon sensible, soit encore parce qu'un nombre maximal d'itérations a été fixé à priori.
Généralement, la partition obtenue finalement dépend du choix initial des centres.

Justification élémentaire de l'algorithme

La variance intra-classes ne peut que décroître (Ou rester stationnaire) entre l'étape m et l'étape m+1. Des règles d'affectation permettent de faire en sorte que cette décroissance soit stricte et donc de conclure à la convergence de l'algorithme puisque l'ensemble de départ I est fini.

II- La classification K-means (Agrégation autour des centres mobiles)

On souhaite effectuer une partition en deux classes.

1. Initialisation : Les centres initiaux sont : $C_1(1) = M_1$ et $C_2(1) = M_2$.
2. Itération 1 : Matrice des distances entre les centres et les points :

$$\begin{pmatrix} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \\ C_1 & \mathbf{0} & 2 & 4 & 5.65 & 4.47 & 4.24 \\ C_2 & 2 & \mathbf{0} & \mathbf{2} & \mathbf{4.47} & \mathbf{4} & \mathbf{3.16} \end{pmatrix}$$

On repère la distance minimale entre chaque points et chaque centre. On place les points dans la classe correspondante. On obtient ainsi les deux classes suivantes : $\gamma_1(1) = M_1$ et $\gamma_2(1) = M_2; M_3; M_4; M_5; M_6$.
Les centres de gravité de ces deux nouvelles classes sont : $C_1(2) = (-2; 3)$ et $C_2(2) = (0.2; 0)$

II- La classification K-means (Agrégation autour des centres mobiles)

3. Itération 2 : Matrice des distances entre les centres et les points :

$$\begin{pmatrix} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \\ C_1 & \mathbf{0} & \mathbf{2} & 4 & 5.65 & 4.47 & 4.24 \\ C_2 & 3.7 & 2.4 & \mathbf{2.4} & \mathbf{2.06} & \mathbf{2.06} & \mathbf{0.8} \end{pmatrix}$$

On obtient les deux classes suivantes : $\gamma_1(2) = (M_1; M_2)$ et $\gamma_2(2) = (M_3; M_4; M_5; M_6)$.
Les centres de gravité de ces deux nouvelles classes sont : $C_1(3) = (-2; 2)$ et $C_2(3) = (0.75; -0.25)$

4. Itération 3 :
Matrice des distances entre les centres et les points :

$$\begin{pmatrix} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \\ C_1 & \mathbf{1} & \mathbf{1} & 3 & 5 & 4.12 & 3.60 \\ C_2 & 4.25 & 3.02 & \mathbf{2.85} & \mathbf{1.46} & \mathbf{1.77} & \mathbf{0.35} \end{pmatrix}$$

On obtient les deux classes suivantes : $\gamma_1(3) = (M_1; M_2)$ et $\gamma_2(3) = (M_3; M_4; M_5; M_6)$.
Comme la partition est la même lors de deux itérations successives, on arrête l'algorithme.

La partition obtenue est $\Gamma = \gamma_1; \gamma_2$ avec $\gamma_1 = (M_1; M_2)$ et $\gamma_2 = (M_3; M_4; M_5; M_6)$.

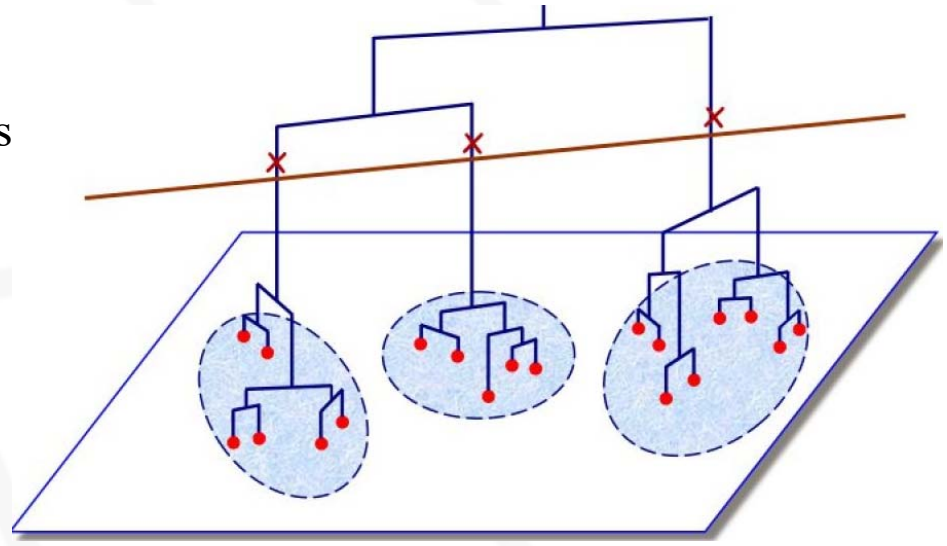
Faire le même raisonnement avec deux autres centres initiaux. Par exemple, si l'on prend $C_1(1) = M1$ et $C_2(1) = M6$ on obtiendra dans ce cas une partition composée des classes : $\gamma_1 = (M_1; M_2; M_3)$ et $\gamma_2 = (M_4; M_5; M_6)$.

III- La classification hiérarchique (classification hiérarchique ascendante)

Le principe de l'algorithme consiste à créer, à chaque étape, une partition obtenue en agrégeant deux a deux les éléments les plus proches. On désignera alors par éléments à la fois les individus et les regroupements d'individus générés par l'algorithme. Il y a différentes manières de considérer le nouveau couple d'éléments agrégés, d'ou un nombre important de variante de cette technique.

L'algorithme ne fournit pas une partition en q classes d'un ensemble de n objets mais une hiérarchie de partition, se présentant sous la forme d'arbres appelés également dendrogrammes et contenant n-1 partitions.

L'intérêt de ces arbres est qu'ils peuvent donner une idée du nombre de classes existant effectivement dans la population.



III- La classification hiérarchique (classification hiérarchique ascendante)

Distance entre éléments et entre groupes

On suppose au départ que l'ensemble des individus à classer est muni d'une distance. On construit alors une première matrice de distances entre tous les individus.

Une fois constitué un groupe d'individus, il convient de se demander ensuite sur quelle base on peut calculer une distance entre un individu et un groupe et par la suite une distance entre deux groupes. Ceci revient à définir une stratégie de regroupements des éléments, c'est-à-dire se fixer des règles de calcul des distances entre groupements disjoints d'individus, appelées critères d'agrégation. Cette distance entre groupements pourra en général se calculer directement à partir des distances des différents éléments impliqués dans le regroupement.

III- La classification hiérarchique (classification hiérarchique ascendante)

Par exemple si x, y, z sont trois objets, et si les objets x et y sont regroupés en un seul élément noté h, on peut définir la distance de ce groupement à z par la plus petite distance des divers éléments de h à z :

$$\mathbf{d(h,z) = Min \{d(x,z), d(y,z) \}}$$

Cette distance s'appelle le saut minimal (single linkage) (Sneath,1957 Johnson,1967) et constitue un critère d'agrégation.

On peut également définir la distance du saut maximal (ou : Diamètre) en prenant la plus grande distance des divers éléments de h à z :

$$\mathbf{d(h,z) = Max \{d(x,z), d(y,z)\}}$$

Une autre règle simple et fréquemment employée est celle de la distance moyenne; pour deux objets x et y regroupés en h :

$$\mathbf{d(h, z) = \{d(x, z) + d(y, z)\} / 2}$$

III- La classification hiérarchique (classification hiérarchique ascendante)

Algorithme de classification

L'algorithme fondamental de classification ascendante hiérarchique se déroule de la façon suivante :

Étape 1 : il y a n éléments à classer (qui sont les n individus);

Étape 2 : on construit la matrice de distances entre les n éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à n-1 classes;

Étape 3 : on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement n-1 éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec n-2 classes et qui englobe la première.

Étape m : on calcule les nouvelles distances jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

III- La classification hiérarchique (classification hiérarchique ascendante)

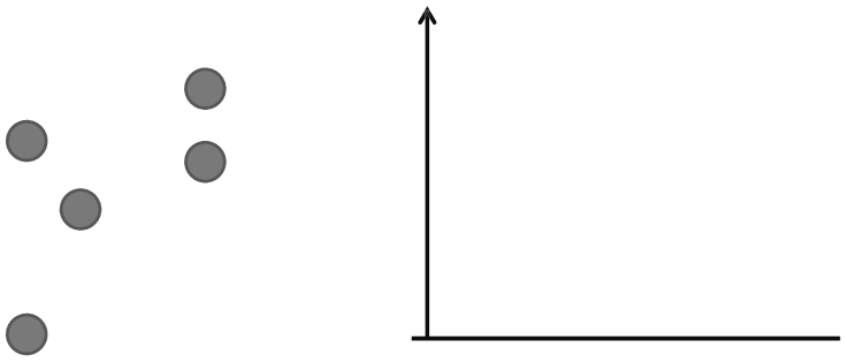
Le Dendogramme

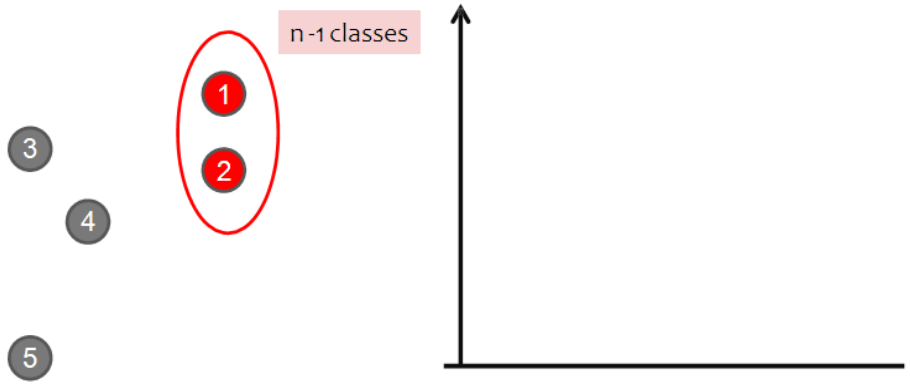
- ✓ Durant les étapes d'un algorithme de classification hiérarchique, on est en train de construire un dendrogramme.
- ✓ Le dendrogramme indique les objets et classes qui ont été fusionnées à chaque itération.
- ✓ Le dendrogramme indique aussi la valeur du critère choisi pour chaque partition rencontrée
- ✓ Il donne un résumé de la classification hiérarchique
- ✓ Chaque palier correspond à une fusion de classes
- ✓ Le niveau d'un palier donne une indication sur la qualité de la fusion correspondante
- ✓ Toute coupure horizontale correspond à une partition

III- La classification hiérarchique (classification hiérarchique ascendante)

Simulation du CAH

n individus / n classes

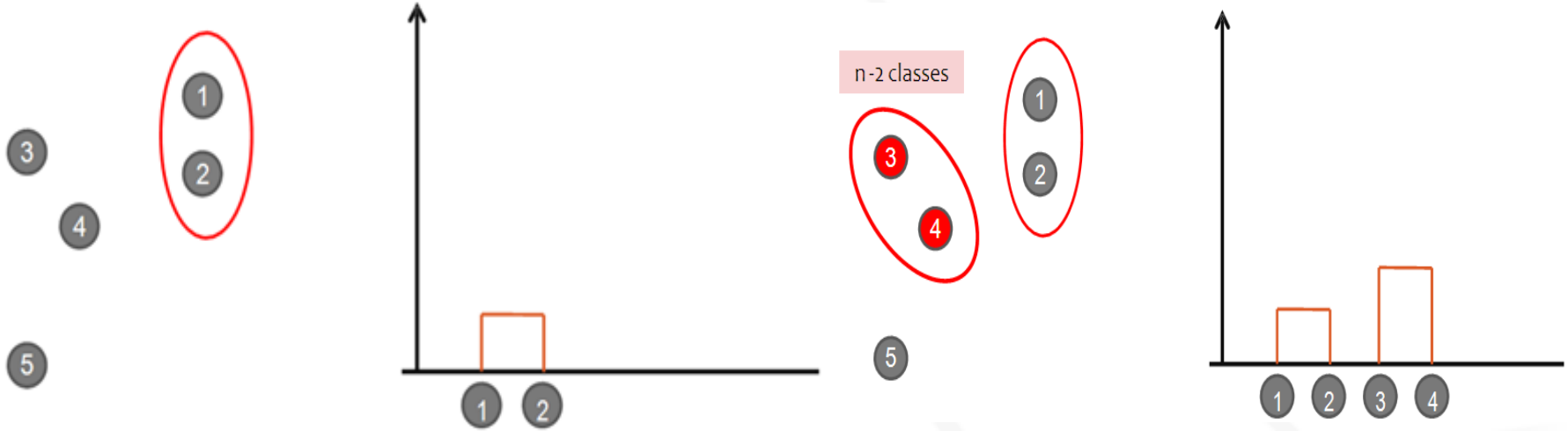




On construit la matrice de distance entre les n éléments
et on regroupe les 2 éléments les plus proches

III- La classification hiérarchique (classification hiérarchique ascendante)

Simulation du CAH

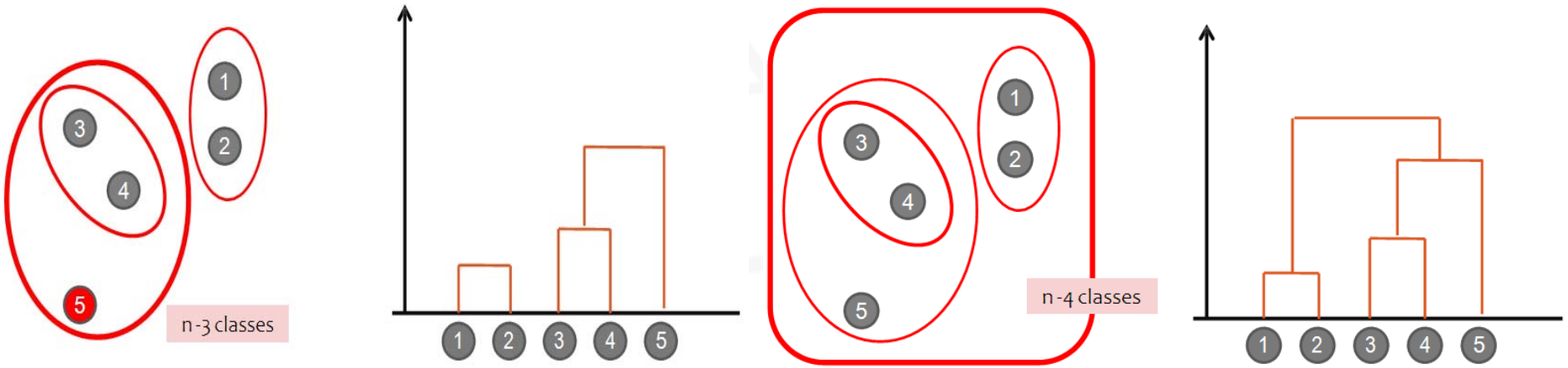


Comment mesurer la distance entre une classe et un élément individuel ?

Critères des centres de gravité, de la distance minimale, maximale, critère de Ward...

III- La classification hiérarchique (classification hiérarchique ascendante)

Simulation du CAH

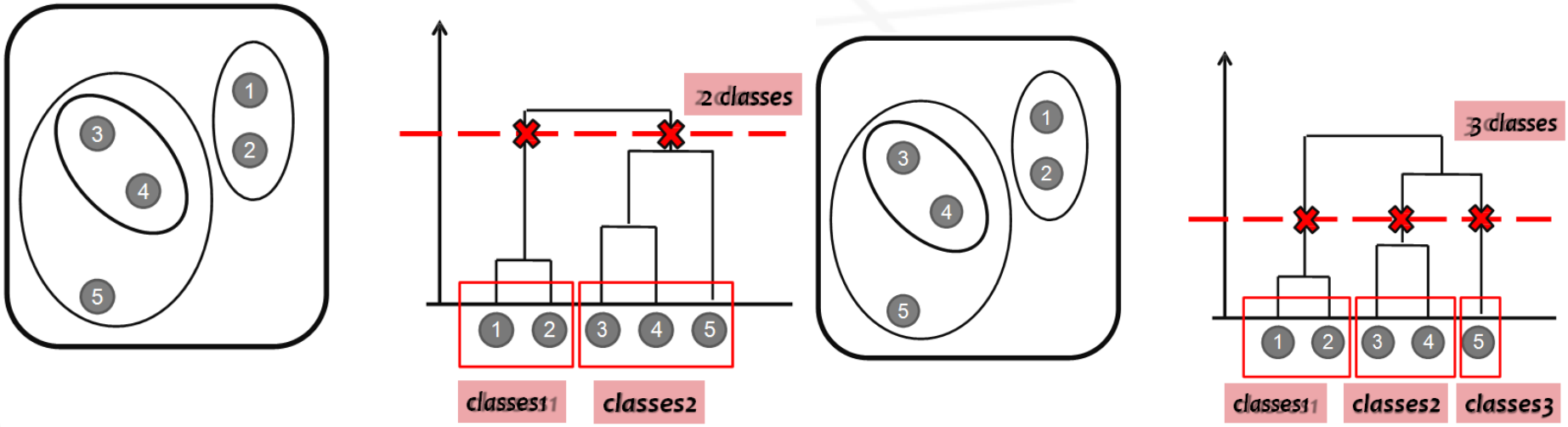


Comment mesurer la distance entre une classe et un élément individuel ?

Critères des centres de gravité, de la distance minimale, maximale, critère de Ward...

III- La classification hiérarchique (classification hiérarchique ascendante)

Coupure du Dendrogramme



Comment mesurer la distance entre une classe et un élément individuel ?

Critères des centres de gravité, de la distance minimale, maximale, critère de Ward...

III- La classification hiérarchique (classification hiérarchique ascendante)

Critère d'agrégation selon la variance

A l'étape initiale, l'inertie Intra-classes est nulle et l'inertie inter-classes est égale à l'inertie totale du nuage puisque chaque élément terminal constitue à ce niveau une classe. A l'étape finale, c'est l'inertie inter-classes qui est nulle et l'inertie intra-classes est équivalente à l'inertie totale puisque l'on dispose à ce niveau d'une partition en une seule classe. Par conséquent, au fur et à mesure que l'on effectue des regroupements, l'inertie intra-classes augmente et l'inertie inter-classes diminue.

Le principe de l'algorithme d'agrégation selon la variance consiste à rechercher à chaque étape une partition telle que la variance interne de chaque classe soit minimale et par conséquent la variance entre les classes soit maximale.

III- La classification hiérarchique (classification hiérarchique ascendante)

Exercice Application

Agrégation selon le lien minimum

	a	*b*	*c*	*d*	*e*
a	0	23	35	43	50
b	23	0	21	32	45
c	35	21	0	11	25
d	43	32	11	0	17
e	50	45	25	17	0

Tableau des dissimilarités

$G_1 = \{c, d\} \Rightarrow$

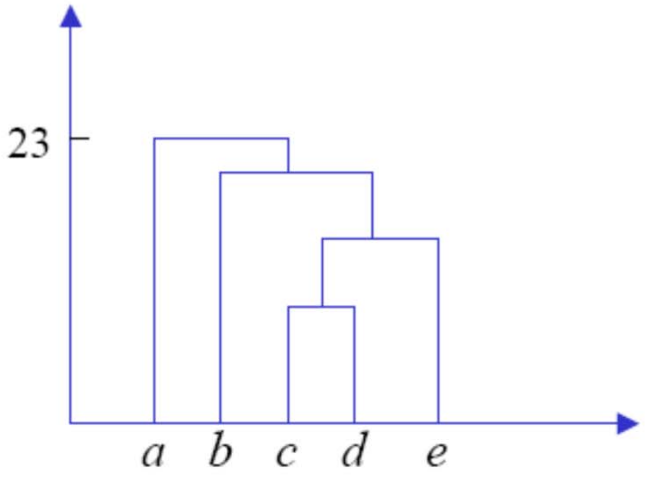
	a	*b*	*e*	G_1
a	0	23	50	35
b	23	0	45	21
e	50	45	0	17
G_1	35	21	17	0

$G_2 = \{e, G_1\} \Rightarrow$

	a	*b*	G_2
a	0	23	35
b	23	0	21
G_2	35	21	0

$G_3 = \{b, G_2\} \Rightarrow$

	a	G_3
a	0	23
G_3	23	0



III- La classification hiérarchique (classification hiérarchique ascendante)

Exercice Application

Agrégation selon le lien maximum

	a	*b*	*c*	*d*	*e*
a	0	23	35	43	50
b	23	0	21	32	45
c	35	21	0	11	25
d	43	32	11	0	17
e	50	45	25	17	0

Tableau des dissimilarités

$G_1 = \{c, d\} \Rightarrow$

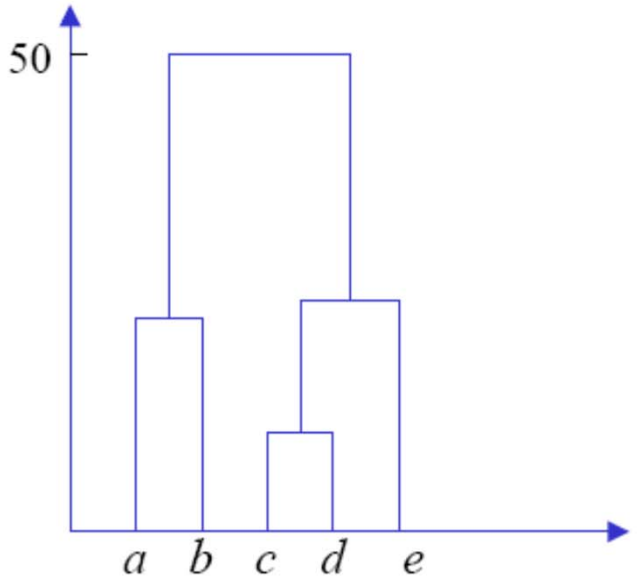
	a	*b*	*e*	G_1
a	0	23	50	43
b	23	0	45	32
e	50	45	0	25
G_1	43	32	25	0

$G_2 = \{a, b\} \Rightarrow$

	e	G_1	G_2
e	0	25	50
G_1	25	0	43
G_2	50	43	0

$G_3 = \{e, G_1\} \Rightarrow$

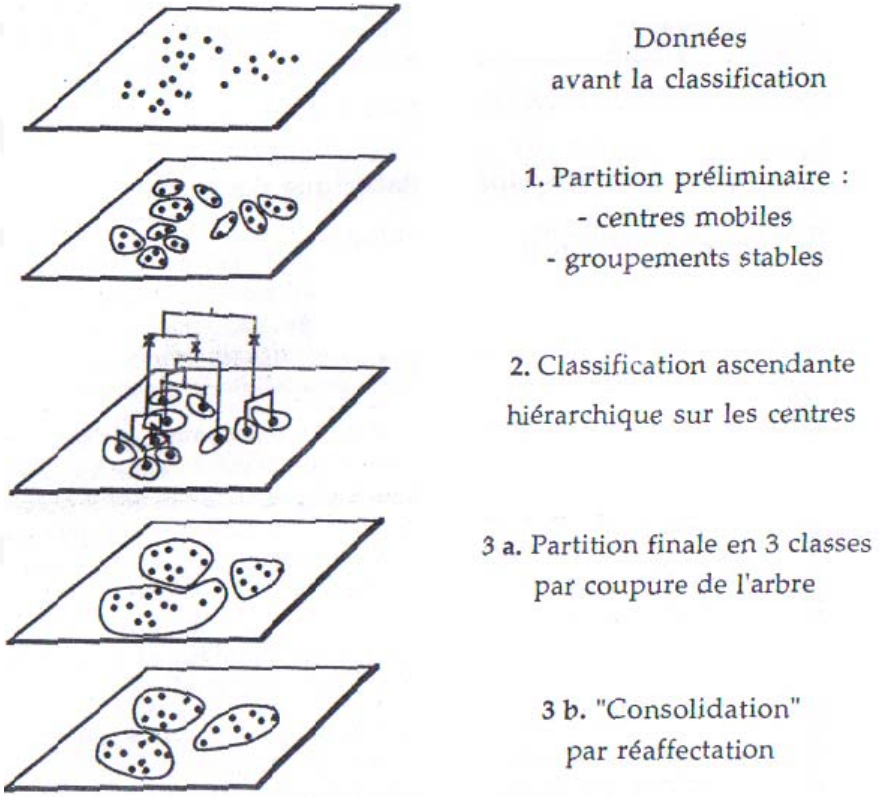
	G_2	G_3
G_2	0	50
G_3	50	0



IV- La classification Mixte

Principe de la classification mixte

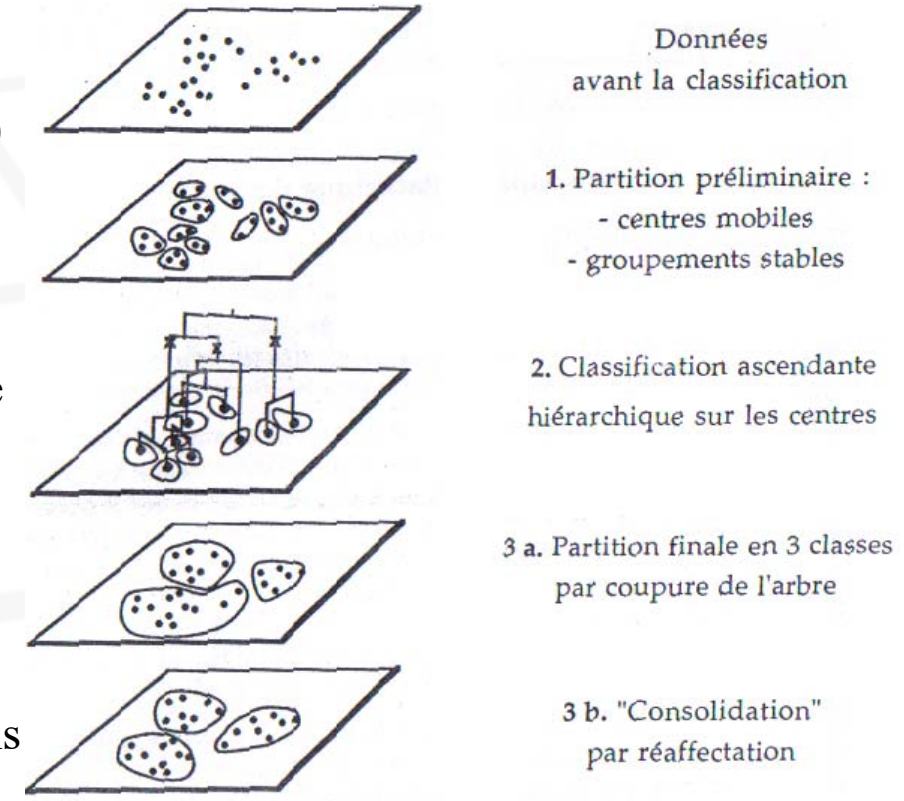
L'algorithme de classification mixte procède en trois phases: l'ensemble des éléments à classer subit un partitionnement initial (centres mobiles) de façon à obtenir quelques dizaines, voire quelques centaines de groupes homogènes; on procède ensuite à une agrégation hiérarchique de ces groupes, dont le dendrogramme suggérera éventuellement le nombre de classes finales à retenir; et enfin, on optimise (encore par la technique des centres mobiles) la ou les partitions correspondant aux coupures choisies de l'arbre.



IV- La classification Mixte

Principe de la classification mixte

L'algorithme de classification mixte procède en trois phases: l'ensemble des éléments à classer subit un partitionnement initial (centres mobiles) de façon à obtenir quelques dizaines, voire quelques centaines de groupes homogènes; on procède ensuite à une agrégation hiérarchique de ces groupes, dont le dendrogramme suggérera éventuellement le nombre de classes finales à retenir; et enfin, on optimise (encore par la technique des centres mobiles) la ou les partitions correspondant aux coupures choisies de l'arbre.



IV- La classification Mixte

Les étapes de la classification mixte

Cette première étape vise à obtenir, rapidement et à un faible coût, une partition des n objets en k classes homogènes, où k est largement plus élevé que le nombre de classes désiré dans la population, et largement plus petit que n. Nous utilisons, pour ce partitionnement initial en quelques dizaines de classes, l'algorithme d'agrégation autour de centres mobiles. Cette procédure augmente l'inertie entre les classes à chaque itération et produit une partition en un nombré fixé au préalable de classes mais qui dépend du choix initial des centres. Ces groupes d'individus ou d'éléments qui apparaissent toujours dans les mêmes classes seront les éléments de base de l'étape suivante.

La seconde étape consiste à effectuer une classification ascendante hiérarchique où les éléments terminaux de l'arbre sont les k classes de la partition initiale. Le but de l'étape d'agrégation hiérarchique est de reconstituer les classes qui ont été fragmentées et d'agréger des éléments apparemment dispersés autour de leurs centres d'origine.

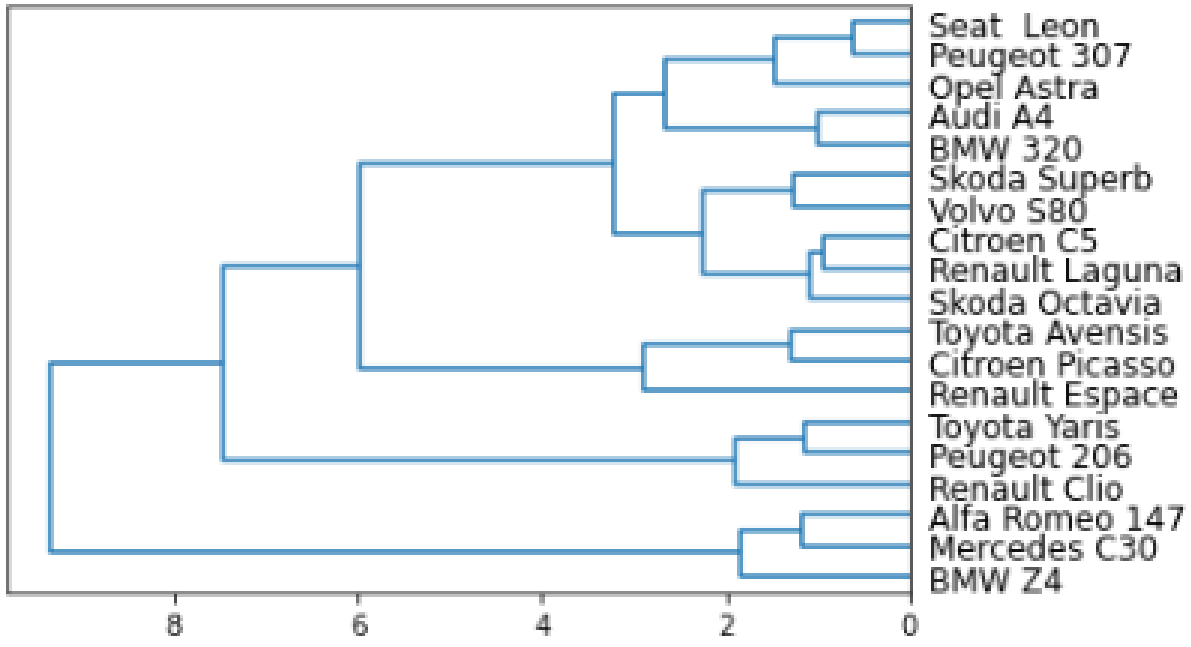
La partition finale de la population est définie par coupure de l'arbre de la classification ascendante hiérarchique. L'homogénéité des classes obtenues peut être optimisée par réaffectations.

CAH avec Python – exemple voiture (suite)

```
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage

#générer la matrice des liens
Z = linkage(data_cr,method='ward',metric='euclidean')

#affichage du dendrogramme plt.title("CAH")
dendrogram(Z,labels=data.index,orientation='left',color_threshold=0)
plt.show()
```



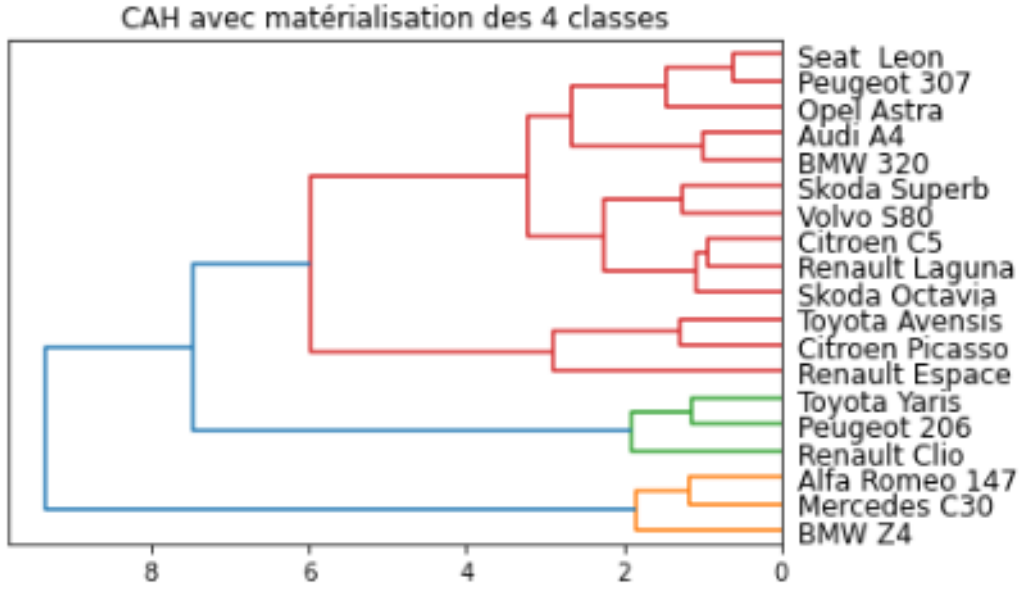
CAH avec Python – exemple voiture (suite)

```
#matérialisation des 4 classes (hauteur t = 7)
plt.title('CAH avec matérialisation des 4 classes')
dendrogram(Z,labels=data.index,orientation='left',color_threshold=7)
plt.show()

#découpage à la hauteur t = 7 ==> identifiants de 4 groupes obtenus
groupes_cah = fcluster(Z,t=7,criterion='distance')
print(groupes_cah)

#index triés des groupes import numpy as np
idg = np.argsort(groupes_cah)

#affichage des observations et leurs groupes
print(pandas.DataFrame(data.index[idg],groupes_cah[idg]))
```



K_Means avec Python – exemple voiture (suite)

```
#k-means sur les données centrées et réduites
from sklearn import cluster
import numpy as np
kmeans = cluster.KMeans(n_clusters=4)
kmeans.fit(data_cr)

#index triés des groupes
idk = np.argsort(kmeans.labels_)

#affichage des observations et leurs groupes
print(pandas.DataFrame(data.index[idk],kmeans.labels_[idk]))

#distances aux centres de classes des observations
print(kmeans.transform(data_cr))

#correspondance avec les groupes de la CAH
pandas.crosstab(groupes_cah,kmeans.labels_)
```

```
                Voiture
0     Skoda Octavia
0      Skoda Superb
0           Audi A4
0          Volvo S80
0         Citroen C5
0            BMW 320
0         Seat  Leon
0    Renault Laguna
0       Peugeot 307
0         Opel Astra
1    Alfa Romeo 147
1      Mercedes C30
1             BMW Z4
2    Toyota Avensis
2  Citroen Picasso
2    Renault Espace
3      Toyota Yaris
3       Peugeot 206
3     Renault Clio
```