Classification

Introduction

- Définitions :
  - Classification : action de constituer ou construire des classes
  - Classe : ensemble d'individus (ou d'objets) possédant des traits de caractères communs (groupe, catégorie)

- Deux types de classification :
  - hiérarchique : arbre, CAH
  - méthode de partitionnement : partition

Données

- Les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés de poids;
  - Les observations de p variables quantitatives sur ces n individus ;
  - Un tableau de distances (ou dissimilarités, ou mesures de dissemblance), n × n, entre les individus pris deux à deux ;

Objectifs

- L'objectif d'une méthode de classification est exploratoire.
- C'est la recherche de partitions, ou répartition des individus en classes, ou catégories.
- On vise donc à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible.
- Cet objectif est à distinguer des procédures de discrimination, ou encore de classement (en anglais classification) pour lesquelles une typologie est a priori connue.
- Nous sommes dans une situation d'apprentissage non-supervisé, ou en anglais de clustering.

Démarche classique

Former des groupes homogènes à l'intérieur d'une population

- Etant donné un ensemble de points, chacun ayant un ensemble d'attributs, et une mesure de similarité définie sur eux,
- trouver des groupes tels que :
  - Les points à l'intérieur d'un même groupe sont très similaires entre eux.
  - Les points appartenant à des groupes différents sont très dissimilaires.
- Le choix de la mesure de similarité est important.

Problématiques

- Nature des observations : données binaires, textuelles, numériques, ... ?
- Notion de similarité (ou de dissimilarité entre observations)
- Définition d'une classe
- Evaluation de la validité d'une classe.
- Nombre de classes pouvant être identifiées dans les données
- Quels algorithmes ?
- Comparaison de différents résultats de classification.

Dissimilarité et Similarité

L'homogénéité d'un groupe d'observations, est lié au degré de la ressemblance entre deux observations.

- Dissimilarité
  - Une fonction de dissimilarité est une fonction d qui à tout couple (x_1, x_2) associe une valeur dans R telle que :
  - $d(x_1, x_2) = d(x_2, x_1) \geq 0$, et $d(x_1, x_2) = 0 \Rightarrow x_1 = x_2$
  - Plus la mesure est faible, plus les points sont similaires.
- Similarité
  - Une fonction de similarité est une fonction s qui à tout couple (x1, x2) associe une valeur dans R telle que :
  - $s(x1, x2) = s(x2, x1) \geq 0$, et $s(x1, x1) \geq s(x1, x2)$
  - Plus la mesure est grande, plus les points sont similaires

Représentation de données

Une fois ces préliminaires accomplis, nous nous retrouvons donc avec :

- soit un tableau de mesures quantitatives n × p, associé à une matrice de produit scalaire p × p définissant une métrique euclidienne,
- soit directement un tableau n × n de dissemblances ou de distances entre individus.

N.B. si n est grand, la deuxième solution peut se heurter rapidement à des problèmes de stockage en mémoire pour l'exécution des algorithmes.

Qualité d'une classification

Quand une partition est-elle bonne?

- Si les individus d'une même classe sont proches
- Si les individus de 2 classes différentes sont éloignés

En d'autres termes :

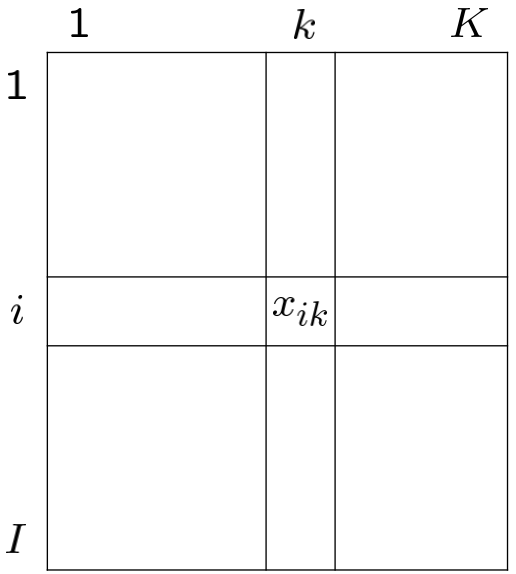
- Variabilité intra-classe petite
- Variabilité inter-classes grande

Classification Ascendante Hiérarchique

»

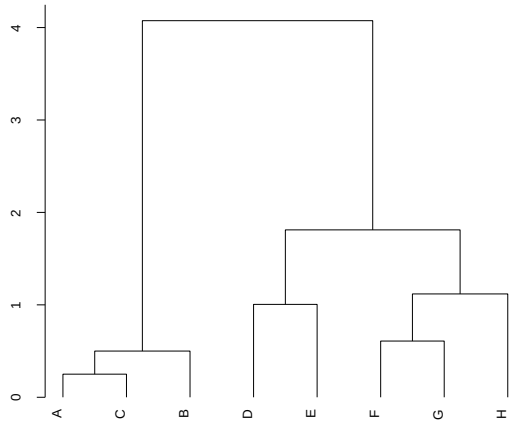
Quelles données pour quels objectifs?

Tableaux de données individus × variables quantitatives

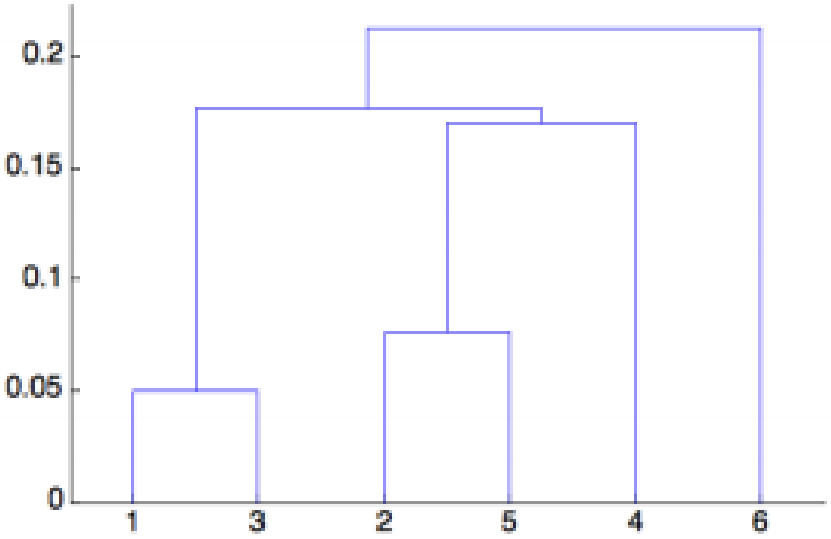


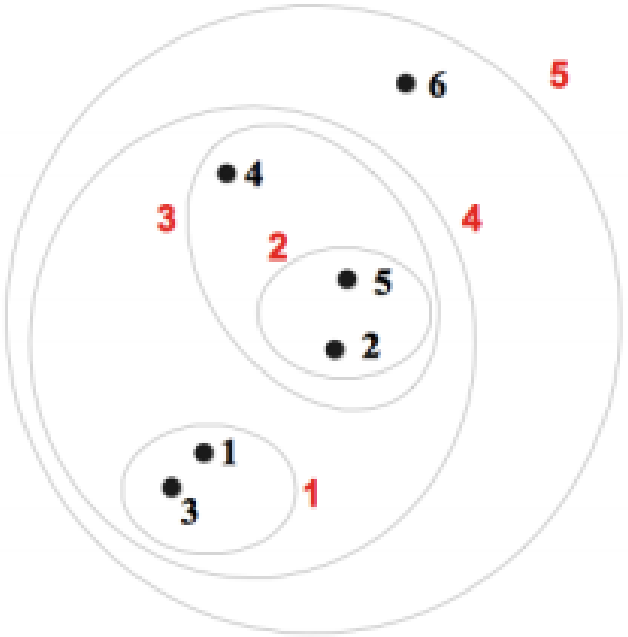
Objectifs : Structure arborescence permettant :

- la mise en évidence de liens hiérarchiques entre individus ou groupes d'individus
- la détection de classes « naturelle » au sein de la population



Classification hiérarchique





Principe

- Chaque individu représente un groupe.
- Trouver les deux groupes les plus proches.
- Grouper les deux groupes en un nouveau groupe.
- Itérer jusqu'à N groupes.

Classification Ascendante Hiérarchique

Principe : Chaque point ou classe est progressivement absorbé par la classe la plus proche.

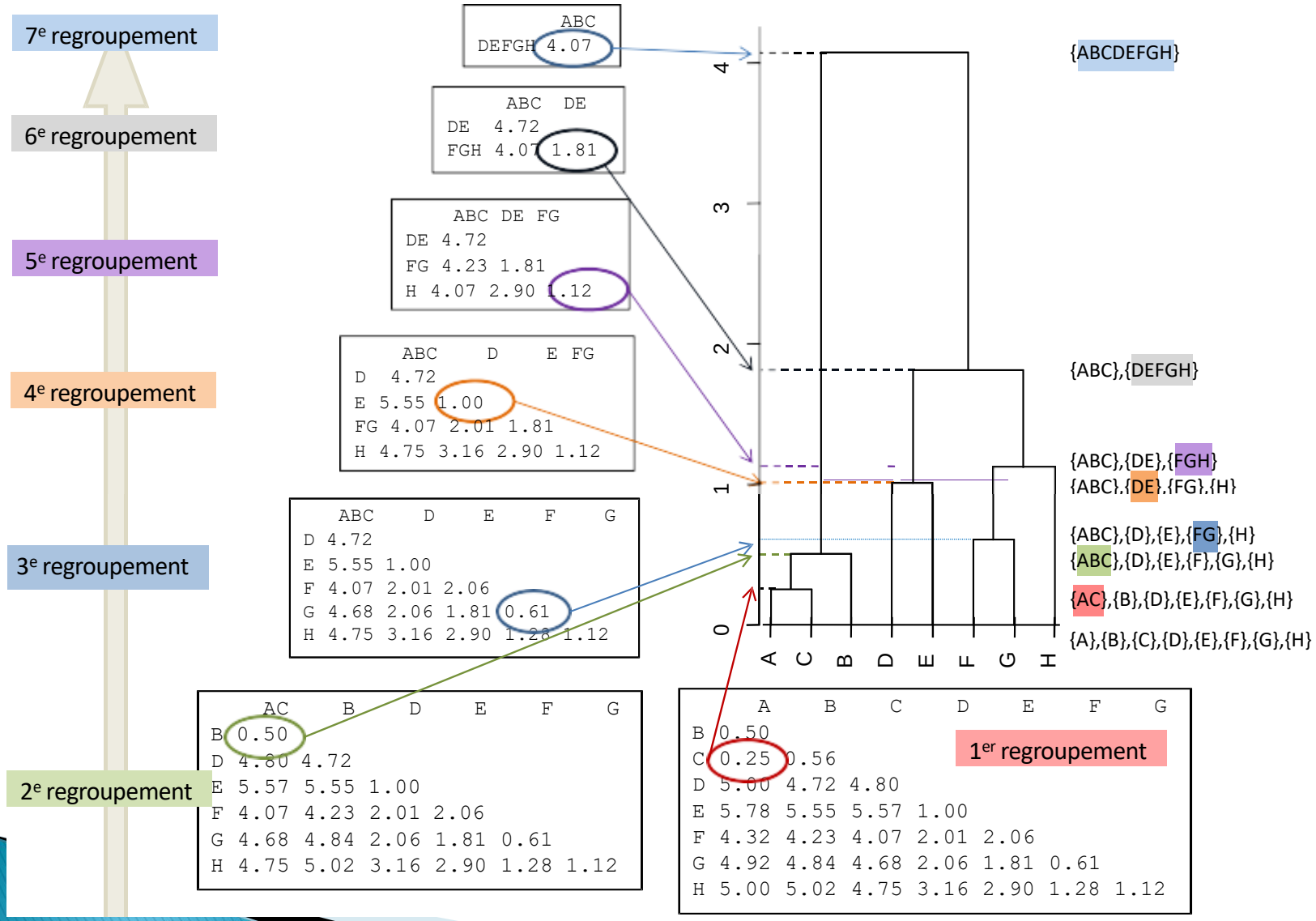
Les quatre étapes de la méthode :

- Choix des variables représentant les individus
- Choix d'un indice de dissimilarité
- Choix d'un indice d'agrégation
- Algorithme de classification et résultat produit

Classification ascendant hiérarchique (CAH) : Algorithme

- Initialisation
  - Chaque individu est placé dans sa propre classe.
  - Calcul de la matrice de ressemblance M entre chaque couple de classe
- Répéter
  - Sélection dans M des deux classes les plus proches C_i et C_j .
  - Fusion de C_i et C_j pour former une classe C_g.
  - Mise à jour de M en calculant la ressemblance entre C_g et les clusters existants.
- Jusqu'à fusion des 2 dernières classes.

Algorithme


7e regroupement
6e regroupement
5e regroupement
4e regroupement
3e regroupement
2e regroupement
ABC
DEFGH 4.07
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
1er regroupement
4
3
2
1
0
A C B D E F G H
{ABCDEFGH}
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

Résultat représenté du dendrogramme

Une hiérarchie de classes telles que :

- toute classe est non vide
- tout individu appartient à une ou plusieurs classes
- deux classes distinctes sont disjointes, sinon elles vérifient une relation d'inclusion
- toute classe est la réunion des classes qu'elle inclus.

Quelques distances ou indices de dissimilarité

- Distance Euclidienne : $d(I_i,I_j)=\sqrt{\sum_k (x_{ik}-x_{jk})^2}$

- Distance Euclidienne au carré : $d(I_i,I_j)=\sum_k (x_{ik}-x_{jk})^2$

- Distance du City-block (Manhattan) : $d(I_i,I_j)=\sum_k \left|x_{ik}-x_{jk}\right|$

- Distance de Tchebychev : $d(I_i,I_j)=Max\ \left|x_{ik}-x_{jk}\right|$

- Distance à la puissance : $d(I_i,I_j)=\left(\sum_k \left|x_{ik}-x_{jk}\right|^p\right)^{1/r}$

Choix de l'indice de dissimilarité entre les individus

- Ce choix est lié aux données étudiées et aux objectifs.
  - Distance Euclidienne : le plus utilisé. Il s'agit d'une distance géométrique dans un espace multidimensionnel.
  - Distance Euclidienne au carré : Permet de "sur-pondérer" les objets atypiques (éloignés).
  - Distance du City-block (Manhattan) : cette distance est simplement la somme des différences entre les dimension.

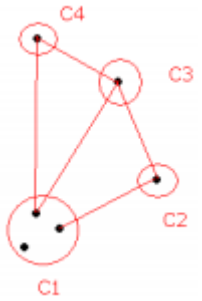
Distance : données qualitatives à valeurs discrètes

- Mesure de la distance $d(x_1, x_2)$ entre deux points x_1 et x_2.
  - Similarité entre individus : Codage disjonctif complet permettant de se ramener à un tableau de variables binaires.
  - Similarité entre variables : tableau de contingence

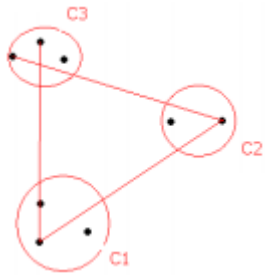
Classification ascendant hiérarchique : métrique

Problème : Trouver la métrique entre les classes la plus proche de celle utilisée entre les individus : min, max, moyenne, ...

- Saut minimal :
  - Tendance à produire des classes générales
  - Sensibilité aux outliers et individus bruités



- Saut maximal :
  - Tendance à produire des classes spécifiques (des classes très proches).
  - Sensibilité aussi aux individus bruités.



Classification ascendant hiérarchique : métrique

- Saut moyen : se base sur la distance $d_{moy}(C1, C2)$
  - Tendance à produire des classes de variance proche
- Barycentre : se base sur la distance $d_{cg}(C1, C2)$
  - Bonne résistance au bruit

Distance Euclidienne au carré et méthode de Ward

✓ Inertie totale = Inertie « intra » + Inertie « inter »

$$I = \sum_{j=1}^{g} \sum_{i=1}^{n_j} G_j M_{ij}^2 + \sum_{j=1}^{g} n_j G G_j^2$$

✓ A chaque étape, on réunit les deux classes de façon à augmenter le moins possible l'inertie « intra »

$$\text{Inertie totale} = \sum_{\text{les classes}} \text{Inertie dans} + \text{Inertie des points moyens pondérés par les effectifs des classes}$$

Saut minimal

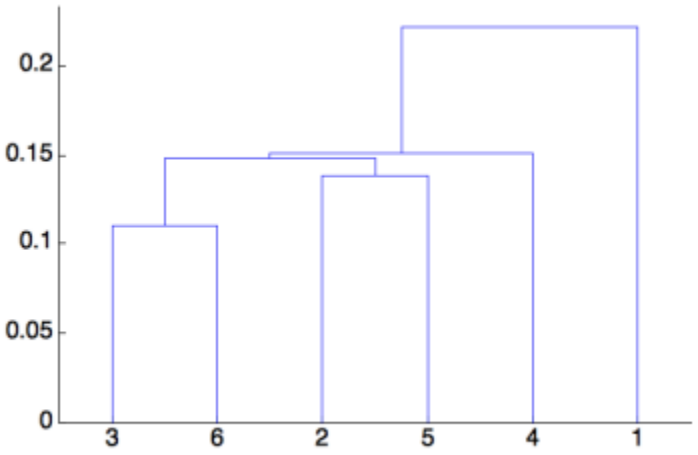
la similarité est déterminée par un lien dans le graphe de proximité

d(3, 6) = 0.11

d({3, 6}, {2, 5}) = min(d(3, 2), d(6, 2), d(3, 5), d(6, 5)) = min(0.15, 0.25, 0.28, 0.39) = 0.15

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 8.4. Euclidean distance matrix for 6 points.

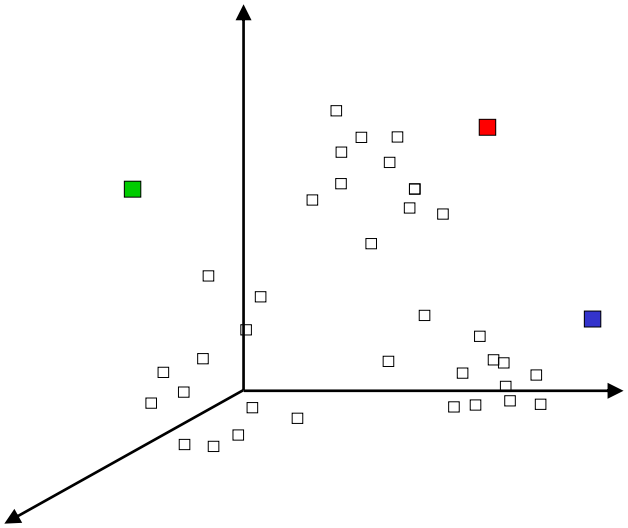


Classifcation par K-means

Classification par K-means : Algorithme

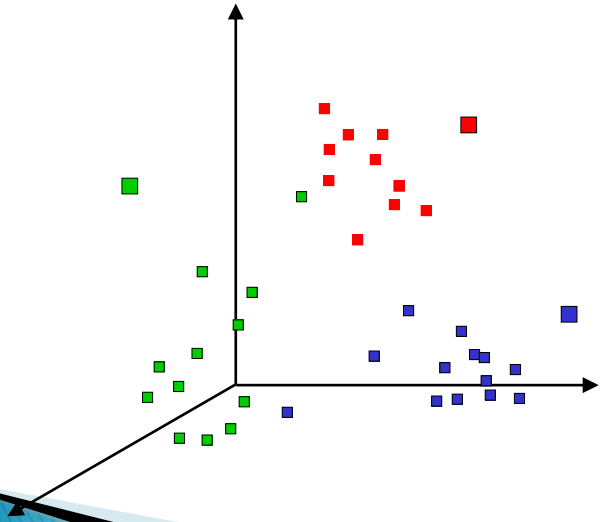
- Initialisation
  - Tirer au hasard, ou sélectionner k points dans l'espace des individus comme centres initiaux de classes
- Itérer les deux étapes suivantes, jusqu'à ce qu'il y ai convergence, c'est-à-dire jusqu'à ce qu'il n'y ai plus aucun individu à changer de classe (stabilisation des classes).
  - Attribuer chaque individu à la classe la plus proche au sens de la métrique choisie; on obtient ainsi, à chaque étape, une classification en k classes.
  - Calculer le centre de gravite de chaque classe : il devient le nouveau noyau ; si une classe s'est vidée, on peut éventuellement retirer aléatoirement un noyau complémentaire.

K-means : illustration de l'algorithme



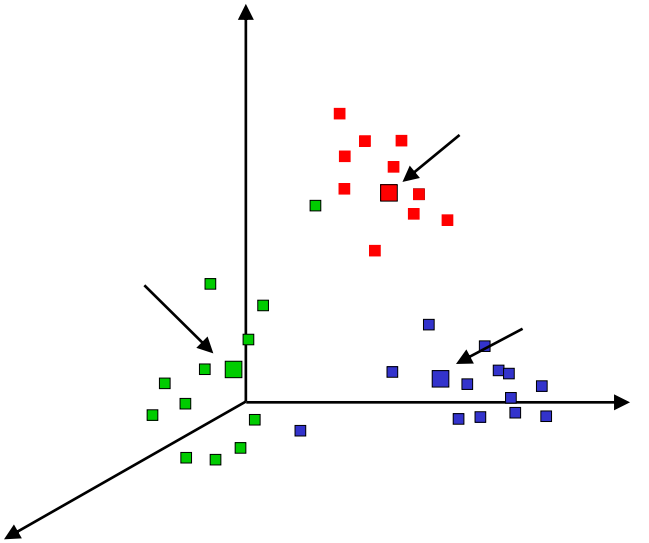
Au départ

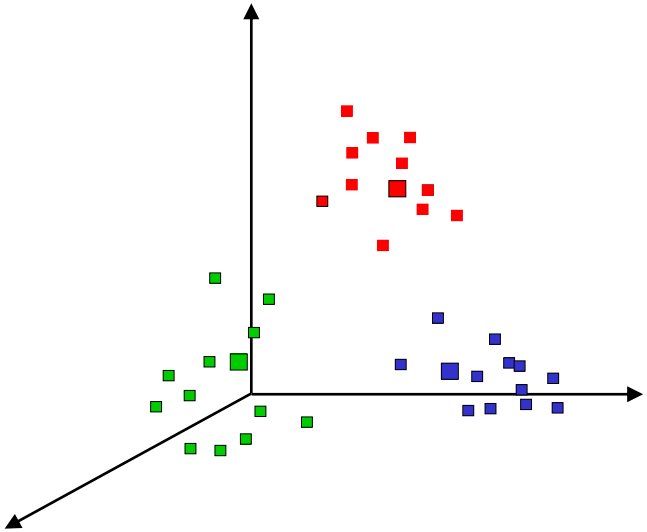
Création aléatoire de centres de gravité.



Etape 1

Chaque observation est classée en fonction de sa proximité aux centres de gravités.



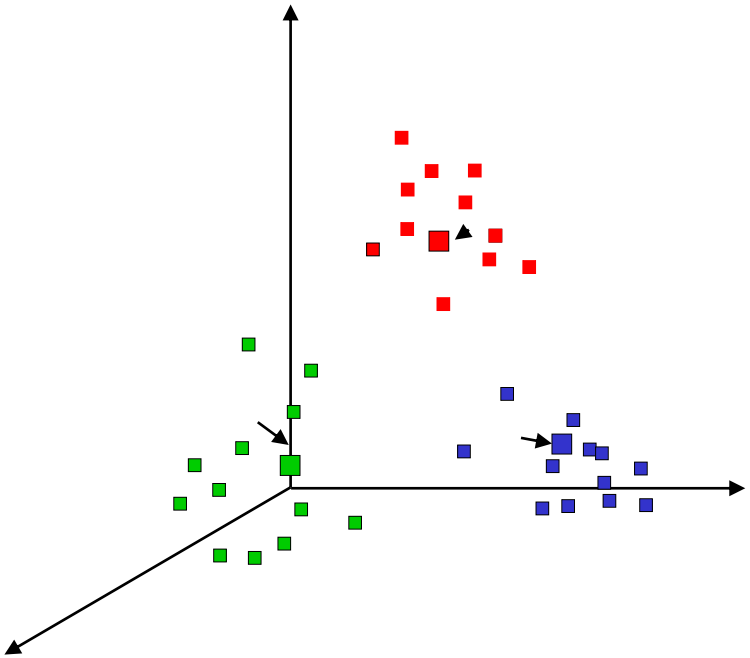


Etape 2

Chaque centre de gravité est déplacé de manière à être au centre du groupe correspondant

Etape 12

On répète l'étape 1 avec les nouveaux centres de gravité.

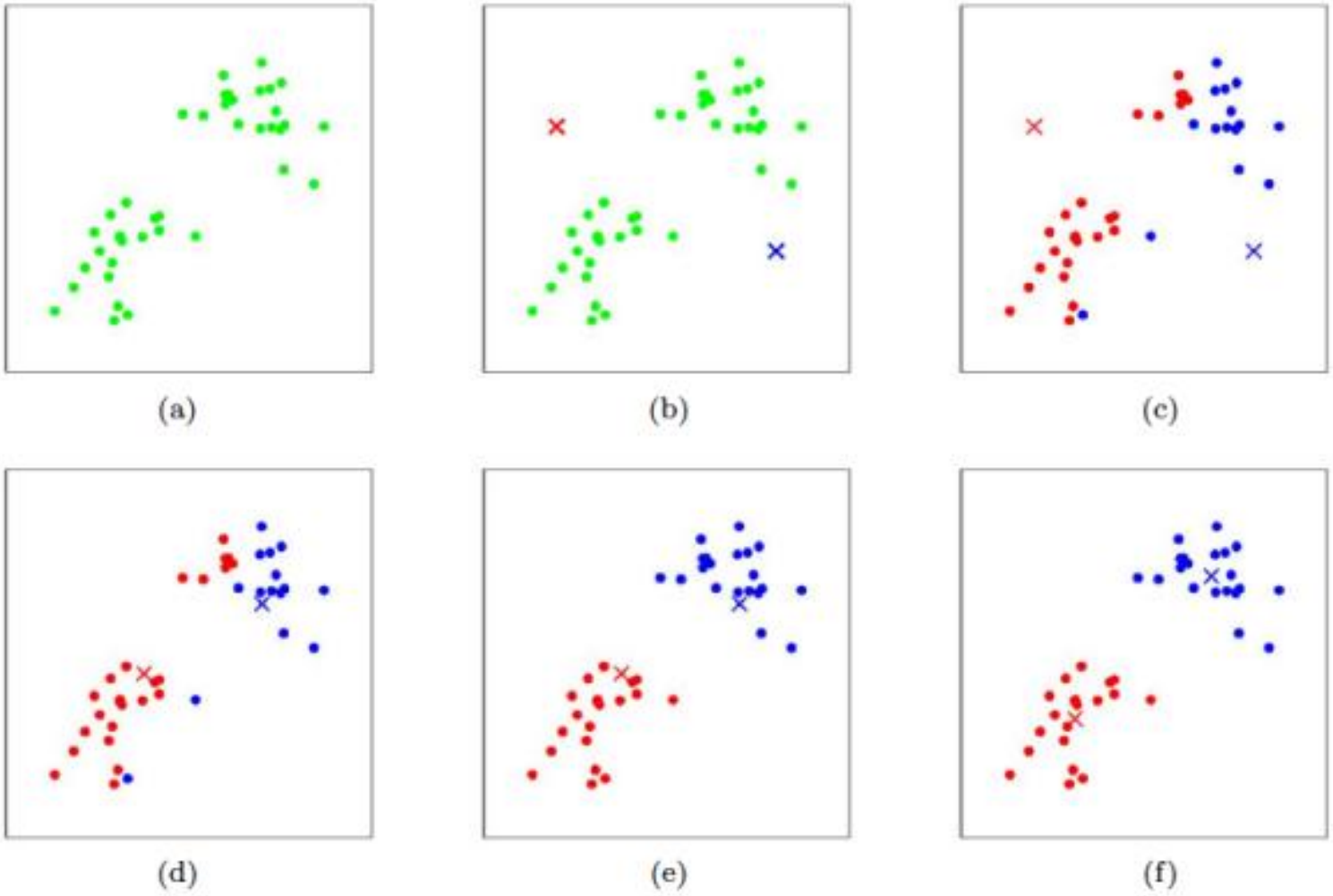


Etape 22

De nouveau, chaque centre de gravité est recalculé.

On continue jusqu'à ce que les centres de gravité ne bougent plus.

Illustration d'un clustring en 2 classes

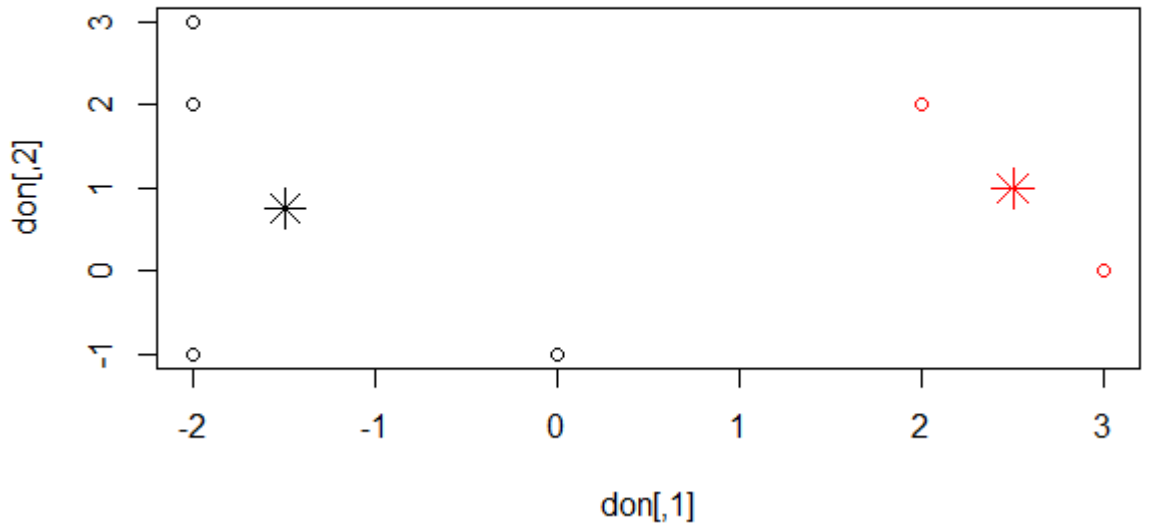


K-means : Détails

- Les centres initiaux sont choisis aléatoirement :
  - Dans l'intervalle de définition des x_i
  - Dans l'ensemble x_i, les clusters résultant peuvent donc varier.
- Le centre est la moyenne des points dans un groupe.
- La proximité est calculée à l'aide d'une distance euclidienne, cosinus, corrélation, ..
- L'algorithme converge souvent en quelques itérations.
- Complexité : O(NKld) avec N nombre de d'individus, d nombre d'attributs (variables), K nombre de clusters et l le nombre d'itérations.

Programme en R

- x = c(-2,-2,0,2,-2,3)
- y = c(2, -1,-1,2,3,0)
- don = matrix(data=c(x,y), nr=6, nc=2)
- ctre = c(-1,2,-1,3)
- ctre1 = matrix(data=ctre, nr=2, nc=2)
- cl1 = kmeans(don,ctre1,algorithm="Lloyd")
- plot(don, col = cl1$cluster)
- points(cl1$centers, col = 1 : 2, pch = 8, cex=2)



Programme R

- Classification hiérarchique des données ”iris” avec le saut moyen :
- data(iris)
- don=iris[,1 :4]
- # # Classification par la C.A.H. #
- hc<-hclust(dist(don), ”ave”)
- plot(hc)

Conclusion

- La classification s'applique à des tableaux individus × variables quantitatives
  - ⇒ L'ACM transforme des variables qualitatives en variables quantitatives
- CAH donne un arbre hiérarchique ⇒ nombre de classes
- K-means consolide les classes