Université de Lille

Méthodes de classification

M. Fourcot
marie.fourcot@univ-lille.fr

2021-2022

Plan

1 Introduction

2 Classification ascendante hiérarchique

3 Méthodes de partitionnement

4 Interprétation des clusters

Données

Id	Prénom	Age	Enfants	Pointure	Taille	Malade
1	Michel	50	3	43	181	non
2	Ultrogoth	28	1	42	175	non
3	Jacqueline	72	2	37	160	non
4	Georges	20	0	46	190	oui

Nous disposons de données pour n indidividus, avec p variables.
Nous voudrions constituer k classes.

Remarque : on pourrait aussi parler d'apprentissage

Deux types de classification

Id	Prénom	Age	Enfants	Pointure	Taille	Malade
1	Michel	50	3	43	181	non
2	Ultrogoth	28	1	42	175	non
3	Jacqueline	72	2	37	160	non
4	Georges	20	0	46	190	oui

Classification supervisée

- l'étiquette y_i est connue pour les n individus de la table
- k classes sont construites à partir des p variables explicatives, afin de regrouper les données ayant la même étiquette
- un nouvel individu recevra l'étiquette de la classe dans laquelle il tombe (prédicteur)

Deux types de classification

Id	Prénom	Age	Enfants	Pointure	Taille
1	Michel	50	3	43	181
2	Ultrogoth	28	1	42	175
3	Jacqueline	72	2	37	160
4	Georges	20	0	46	190

Classification non-supervisée

- pas d'étiquette
- k classes sont construites à partir des p variables descriptives, afin de représenter les données
- profilage des classes

Objectif : Partitionner un ensemble de n individus décrits par p variables en k sous groupes (clusters) qui sont les plus homogènes possibles

Inertie

Définition : inertie

On appelle inertie du nuage de points par rapport à l'origine G

$$I_G = \sum_{i=1}^{n} p_i d_M^2(\mathrm{i}, \mathrm{G})$$

Théorème de Huygens $I_T = I_B + I_W$

Minimisation de l'inertie intra-classe $\Rightarrow$ rassembler les individus "proches" et dissocier les individus éloignés.

Inertie

Exemple : De qui Michel est-il le plus proche ?

Id	Prénom	Age	Enfants	Pointure	Taille
1	Michel	50	3	43	181
2	Ultrogoth	28	1	42	175
3	Jacqueline	72	2	37	160
4	Georges	20	0	46	190

Distance euclidienne entre Michel et Ultrogoth :

$$d_{12} = \sqrt{(50-28)^2 + (3-1)^2 + (43-42)^2 + (181-175)^2}$$

Les deux variables "Pointure" et "Enfants" n'ont que peu de poids dans le calcul de la distance d'où la nécessité de centrer réduire.

Différentes méthodes

Il existe plusieurs méthodes de classification non supervisée. Le choix d'une méthode par rapport à une autre dépend de plusieurs facteurs :

- taille de l'échantillon
- nature des variables

Étapes

- construire la partition $C_1, \ldots, C_k$
- vérifier la séparabilité des clusters
- interpréter les classes obtenues

Approches classiques

- Méthodes hiérarchiques
- Méthodes de partionnement
- Méthodes mixtes (hiérarchique + partitionnement)

Plan

1 Introduction

2 Classification ascendante hiérarchique

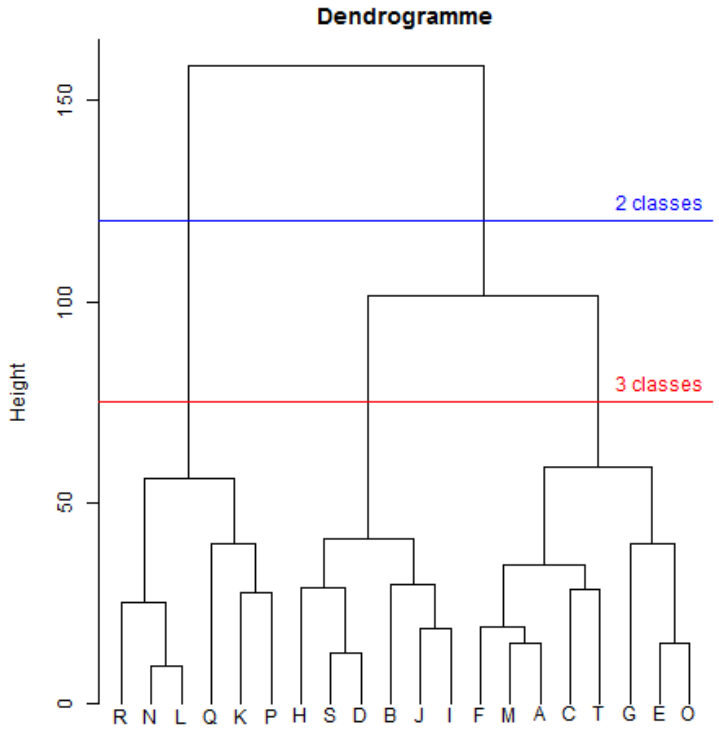
3 Méthodes de partitionnement

4 Interprétation des clusters

Principe

Étapes :

- Construction d'une hiérarchie de partitions par agrégation successive d'individus les plus "proches"
- Choix du nombre de classes
- Obtention de la classification par découpage de l'arbre



Algorithme

Étape 1 :

- Calcul des distances entre chaque élément de l'échantillon des données

$$\begin{bmatrix} 0 & d(i=1,j=2) & \dots & \dots & d(i=1,j=n) \\ d(i=2,j=1) & 0 & \dots & \dots & \dots \\ \dots & \dots & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ d(i=n,j=1) & \dots & \dots & \dots & 0 \end{bmatrix}$$

- Regroupement des deux individus les plus proches
  $\rightarrow$ obtention de $n-1$ classes

Algorithme

Étape j ($1 \leqslant j \leqslant n-1$) :

- Calcul des dissemblances entre chaque partie obtenue à l'étape $j-1$
- Regroupement des deux parties les plus proches
  $\rightarrow$ obtention de $n-j$ classes

On itère n fois l'algorithme jusqu'à agrégation complète.
On représente les agrégations progressives sous forme d'arbre (dendrogramme).

Questions

- Choix d'une distance entre les individus
- Choix d'un critère d'agrégation entre les classes
- Choix du nombre de classes

Distance entre individus

Calcul de la matrice D en fonction du type de variables

- Les variables $X_1, \ldots, X_p$ sont numériques. Dans ce cas, on utilise la distance euclidienne.
- Les variables $X_1, \ldots, X_p$ sont binaires. Soit on utilise un indice de similarité (simple comptage - ne tient pas compte de la structure des variables et des liens), soit on utilise une méthode factorielle et on travaille sur les axes issus de la transformation.
- Tableau de contingence : on utilise la distance du χ^2 (considérée comme euclidienne pour la méthode de Ward) pour calculer la distance entre les lignes ou colonnes du tableau.

Indices de similarité

	présent	absent
présent	a	b
absent	c	d

- Jaccard

$$\frac{a}{a+b+c}$$

- Dice Sorensen

$$\frac{2a}{2a+b+c}$$

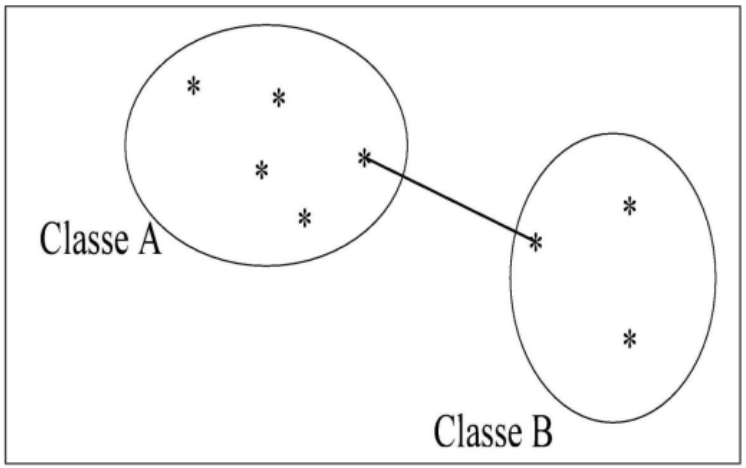
- Russel et Rao

$$\frac{a}{a+b+c+d}$$

Distance d'agrégation

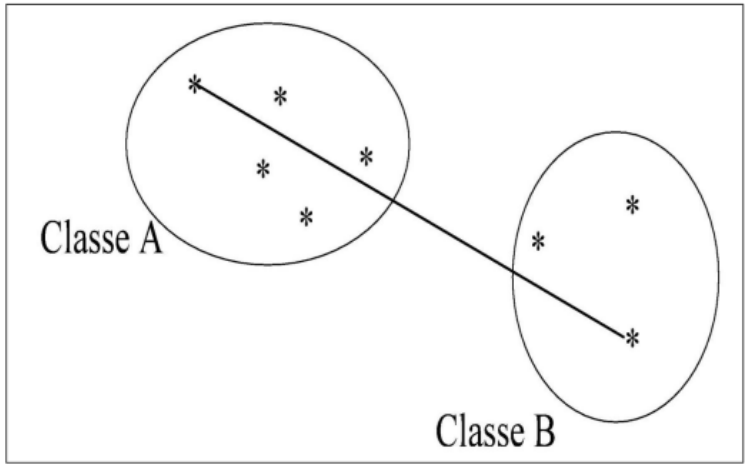
Choix d'une distance permettant de quantifier la dissemblance entre les parties.
Le choix de l'un ou l'autre des critères d'agrégation peut donner des dendrogrammes différents.

- saut minimum (single linkage) : plus petite distance
  $d(A, B) = \min_{a \in A, b \in B} d(a, b)$



Distance d'agrégation

- lien maximal (complete linkage) $d(A, B) = \max_{a \in A, b \in B} d(a, b)$



- distance moyenne (average linkage)

$$d(A, B) = \frac{1}{n_A n_B} \sum_{a \in A, b \in B} d(a, b)$$

- méthode de Ward

Distance d'agrégation

Principe de la méthode de Ward :

On définit :

- L'inertie intra-classe est la moyenne des carrés des distances des points de la classe au centre de gravité
- L'inertie inter-classe est la moyenne des carrés des distances des différents centres de gravité

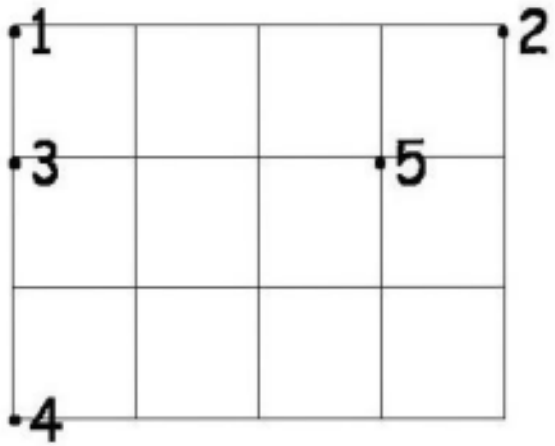
On cherche à faire varier au minimum l'inertie intra-classe.
À chaque itération, on regroupe les classes pour lesquelles est minimale la perte :

$$d(A,B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$$

avec p_A, p_B les poids des classes A et B et g_A, g_B les centres de gravité des classes A et B.

Exemple

- Cinq points dans un plan :

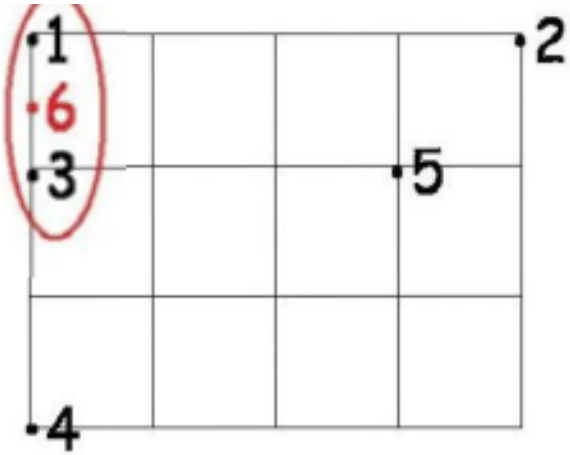


- Carrés des distances euclidiennes

	1	2	3	4	5
1	0	16	1	9	10
2		0	17	25	2
3			0	4	9
4				0	13
5					0

Exemple

- Regroupement de 1 et 3 $\rightarrow$ nouvel individu 6

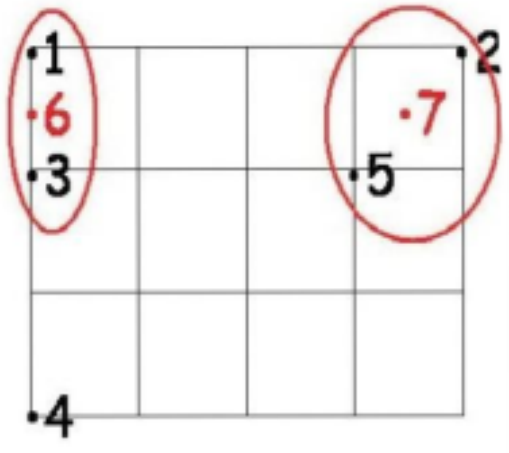


- Carrés des distances euclidiennes (arrondis)

	2	4	5	6
2	0	25	2	16
4		0	13	6
5			0	7
6				0

Exemple

- Regroupement de 2 et 5 $\rightarrow$ nouvel individu 7

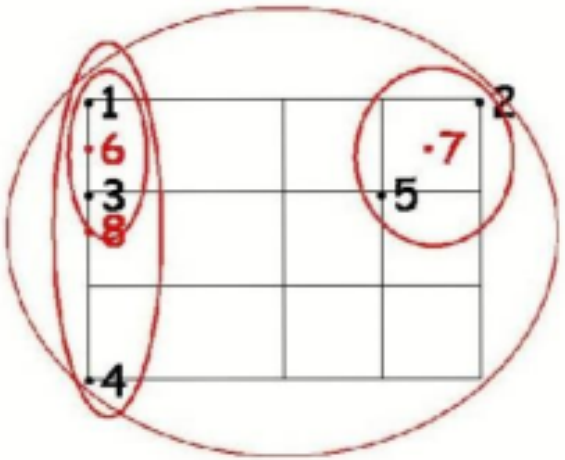


- Carrés des distances euclidiennes (arrondis)

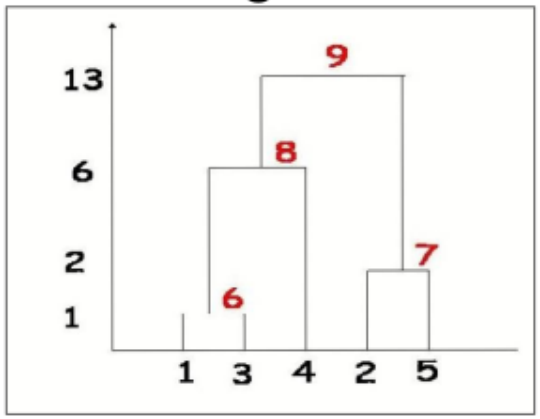
	4	6	7
4	0	6	19
5		0	9
7			0

Exemple

- Regroupement de 4 et 6 $\rightarrow$ nouvel individu 8



Dendrogramme :



Avantages et inconvénients

Avantages :

- Fournit à la fois les classes et leur nombre

Inconvénients :

- Souvent malaisé de choisir la coupure significative sur le dendrogramme
- Partition non-optimale en raison de sa structure hiérarchique
- Fort coût algorithmique lorsque le nombre d'individus devient grand

Plan

1 Introduction

2 Classification ascendante hiérarchique

3 Méthodes de partitionnement

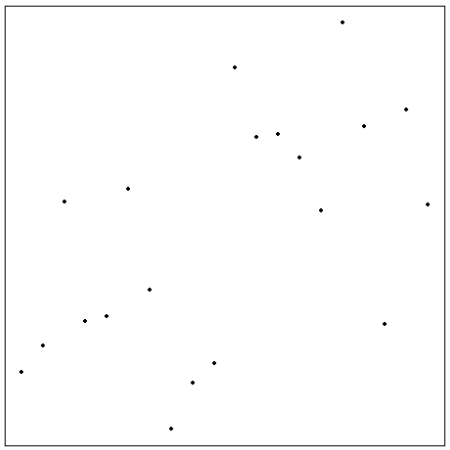
4 Interprétation des clusters

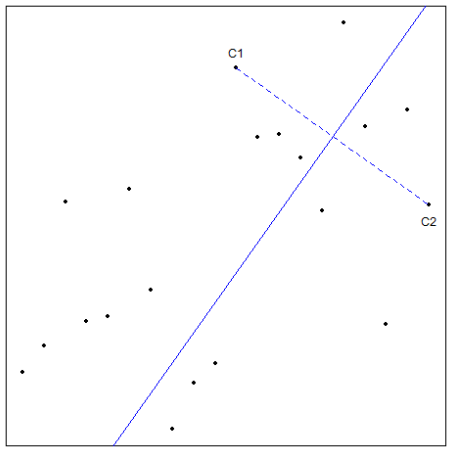
Méthodes de partitionnement

Principe des centres mobiles (Forgy 1965) :

Partitionner un ensemble de n individus décrits par p variables **numériques** en k sous groupes (clusters) qui sont les plus homogènes possibles

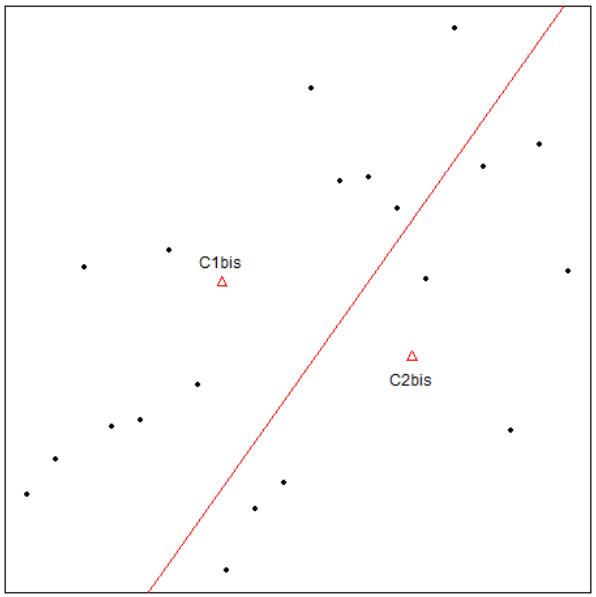
- fixer k, le nombre de clusters / classes souhaité
- choisir des centres initiaux (au hasard ou non) et affecter chaque individu au centre le plus proche

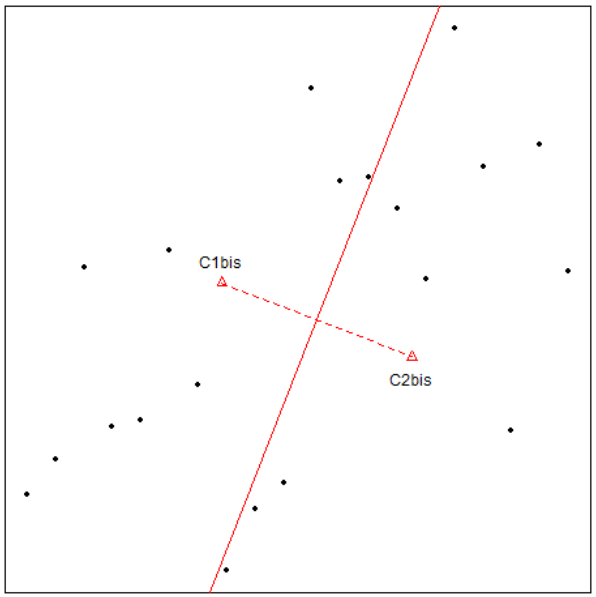




Méthodes de partitionnement

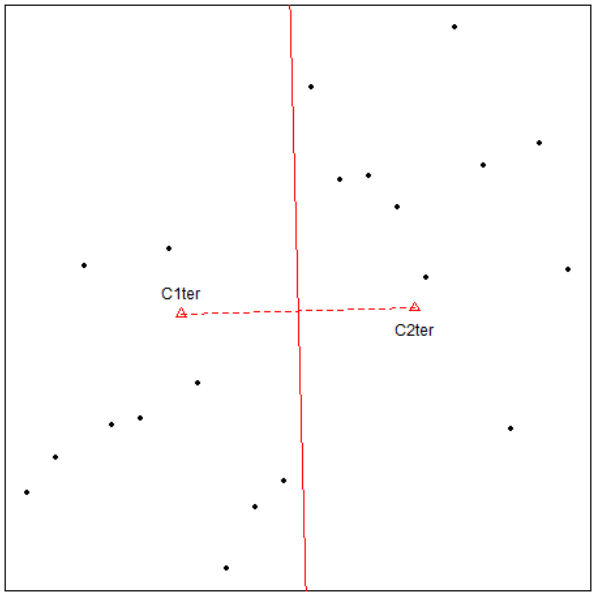
- calculer les nouveaux centres, centres de gravité de chaque classe
- affecter chaque individu au centre le plus proche

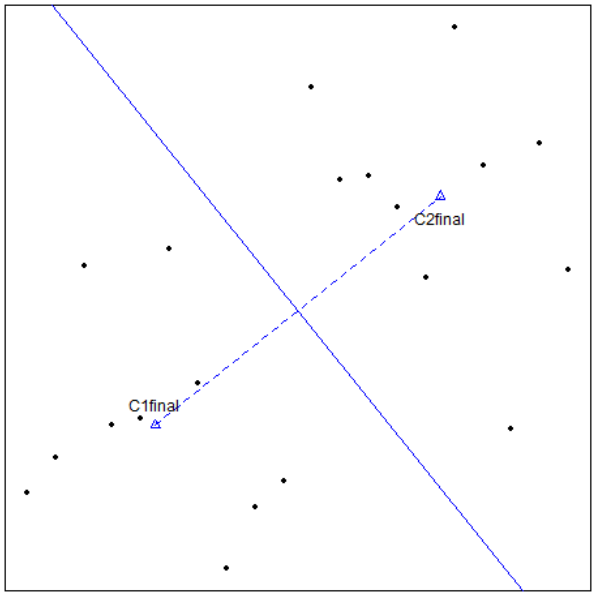




Méthodes de partitionnement

- réitérer l'opération jusqu'à obtenir des classes stables ou un nombre donné d'itérations





Variantes

- k-means
  Remarque : de manière générale k-means = centres mobiles
  mais parfois k-means = k-means incrémental
- nuées dynamiques
- k-medoids

Variantes

k-means incrémental (Mc Queen 1967) :
les observations sont ajoutées les unes après les autres et le logiciel recalcule les barycentres après chaque ajout
$\Rightarrow$ l'algorithme est plus rapide.

nuées dynamiques (Diday 1971) :
chaque classe n'est plus représentée par son barycentre (éventuellement extérieur à la population) mais par un sous-ensemble de la classe, appelé noyau, qui sera plus représentatif de la classe que son barycentre s'il est bien composé (des individus les plus centraux, par exemple).

Variantes

k-medoids PAM : partitioning around medoids
(Kaufman et Roussew 1990)
Un médoide est le représentant d'une classe, choisi comme son objet le plus central, ce dont on s'assure en permutant systématiquement un représentant et un autre objet de la population choisi au hasard. On regarde si la qualité de la classification croît, c'est à dire si la somme des distances de tous les objets à leurs représentants décroit. L'algorithme s'arrête lorsque plus aucune permutation n'améliore la qualité.

Avantage : plus robuste aux données aberrantes.
Inconvénient : temps de calcul plus long

Avantages et inconvénients

Avantages :

- méthodes rapides (sauf PAM) qui convergent tout le temps.
- fonctionnent avec de gros volumes de données.

Inconvénients :

- il faut fixer le nombre de classes au départ
- les résultats sont dépendants des centres initiaux
- méthodes limitées aux variables numériques

Avantages et inconvénients

Solutions :

- choix du nombre de classes : visualiser la structure de la population par une ACP (variables quantitatives) ou une ACM (variables qualitatives) pour faire apparaître un certain nombre de nuages plus ou moins formés.
- problème du choix des centres initiaux $\Rightarrow$ 2 solutions :
  1. effectuer plusieurs tirages aléatoires des centres initiaux puis comparer les classifications obtenues et croiser les classes pour constituer des formes fortes.
  2. faire deux classifications : la première fournit des centres finaux qui seront utilisés comme centre initiaux dans la deuxième classification.
- variables qualitatives : faire une AFC ou ACM puis utiliser les coordonnées des observations (souvent individus) sur les axes factoriels comme variables quantitatives.

Distance/Qualité

Critère de qualité :

$$R^2 = \frac{\text{inertie inter-classe}}{\text{inertie totale}}$$

Ce critère augmente avec le nombre de classes et la classification perd en robustesse. On définit donc le Pseudo F qui doit être le plus élevé possible :

$$\text{Pseudo F} = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}}$$

Plan

1 Introduction

2 Classification ascendante hiérarchique

3 Méthodes de partitionnement

4 Interprétation des clusters

Formes fortes

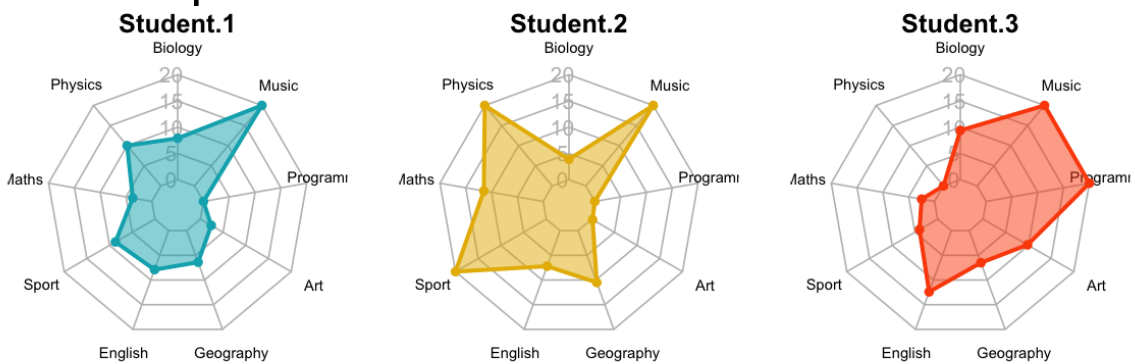
Si on exécute plusieurs fois un algorithme de classification sur les mêmes données, on peut aboutir à des résultats différents. On appelle formes fortes les regroupements stables.

Les formes fortes permettent de valider les calculs des clusters, on vérifie aussi qu'il n'y a pas de clusters qui correspondent à des individus aberrants.

Visualisation

Pour donner une interprétation clinique, on croise la variable "cluster" avec les variables de départ pour voir les caractéristiques de chaque cluster.

- si les variables sont numériques, on représente chaque cluster par son centre de gravité et on utilise une représentation en "radar" pour visualiser les clusters.



- si les variables sont binaires, on représente les pourcentages.

On peut également s'intéresser à la dispersion au sein des clusters pour voir si les clusters sont homogènes ou non.

Interprétation des clusters

Méthodes de classification et méthodes factorielles

Ces approches sont complémentaires : la classification permet de segmenter la population en groupes homogènes dont les méthodes factorielles fournissent l'interprétation.

Résumé

But classification : faire en sorte que la différenciation entre les groupes soit maximale, former des groupes dans lesquels les éléments se ressemblent le plus.

- Méthodes de partitionnement : partir d'un nombre de groupes fixé a priori.
- Classification hiérarchique : rassembler des éléments en une suite de partitions emboîtées en procédant pas à pas.