Ministère de l'enseignement supérieur et de la recherche scientifique
Université de JIJEL
Faculté des Sciences Exactes et Informatique
Département d'Informatique

CAH

Classification Ascendante Hiérarchique

Définition

La **classification** a pour but de **regrouper** des **individus** en **classes homogènes** en fonction de l'étude de certaines **caractéristiques des individus.**

→ **Classes homogènes** : consiste à **regrouper les individus** qui **se ressemblent** et **séparer** ceux qui **sont éloignés**.

Définition

Principe

Les diverses techniques de classification visent **à répartir n individus**, caractérisés par p variables **X1, X2, …, Xp** en un certain nombre **m** de **sous-groupes** aussi **homogènes** que possible.

Techniques de classification

La classification non hiérarchique ou **partitionnement**:

la décomposition de l'ensemble de tous les individus en m ensembles disjoints ou classes d'équivalence **mais le nombre m de classes est fixé à l'avance**.

La classification non hiérarchique

Partitionner un ensemble d'observations (E) consiste à :

1-Regrouper les observations en *classes homogènes* (les sous-ensemble partagent des caractéristiques communes.

2-Regrouper les observations selon un critère:

-**Critère de similarité (la proximité)** pour inclure les observations qui se ressemblent dans une classe.

-**Critère de différenciation (la distance)** pour les exclure

La classification non hiérarchique

Notion de ***Partition***

Soit un ensemble E = { A.B.C.D}

Une partition de E est un ensemble de classe qui satisfait deux conditions:

Deux classes A et B : soient disjointes, soient confondues

→A et B sont disjointes, Si A∩B=Ø

→A et B sont confondues, si AUB= A=B

→L'Union de toutes les classe correspond à l'ensemble E, AUBUCUD=E

La classification hiérarchique

Elle consiste à fournir un ensemble de partitions de E en classes de moins en moins fines obtenues par regroupements successifs de parties

→Pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

Classification supervisée

Contexte : On considère une population divisée en q groupes d'individus différents.

Ces groupes sont distinguables suivant les valeurs de p caractères $X_1, \ldots, X_p$, sans que l'on ait connaissance des valeurs de $X_1, \ldots, X_p$ les caractérisant. On dispose de:

Classification supervisée

- n individus avec, pour chacun d'entre eux, les valeurs de X1, . . . , Xp et son groupe d'appartenance,
- un individu $\omega*$ de la population avec ses valeurs de X1, . . . , Xp, sans connaissance de son groupe d'appartenance.

- **Objectif :** Partant des données, l'objectif est de déterminer à quel groupe l'individu $\boldsymbol{\omega*}$ *a le* plus chance d'appartenir.

Classification non supervisée

- **Contexte** : On considère n individus extraits au hasard d'une population. Pour chacun d'entre eux, on dispose de p valeurs de p caractères X1, . . . , Xp.

- **Objectif** : Partant des données, l'objectif est de regrouper/classer les individus qui se ressemblent le plus/qui ont des caractéristiques semblables.

Classification non supervisée

- Ce regroupement peut avoir des buts divers :

  - tenter de séparer des individus appartenant à des sous-populations distinctes,

  - décrire les données en procédant à une réduction du nombre d'individus pour communiquer,

  - simplifier, exposer les résultats. . .

Méthodes

Pour atteindre l'objectif, plusieurs méthodes sont possibles. Parmi elles, il y a:

- l'algorithme de Classification Ascendante Hiérarchique (CAH),
- l'algorithme des centres mobiles,
- l'algorithme de Classification Descendante Hiérarchique (CDH),
- la méthode des nuées dynamiques (partitionnement autour d'un noyau),
- la méthode de classification floue,

Algorithme de Classification Ascendante Hiérarchique (CAH)

Objectif: obtenir une hiérarchie, c'est-à-dire une collection de groupes d'observations.

Étude de la ressemblance

1. **Ressemblance : On dira que des individus se ressemblent si les points associés sont** proches les uns des autres/si les distance qui les séparent sont petites. Ainsi, on souhaite rechercher dans N les zones denses pouvant correspondre à des groupes d'individus qu'il s'agira d'interpréter par la suite.

Étude de la ressemblance

2. Distances : On peut donc aborder le problème de la ressemblance entre individus par le biais de la notion de distance.

On appelle distance sur un ensemble *M toute application* $d : M^2 \rightarrow [0, \infty[$ *telle que :*

- **pour tout** $(x, y) \in M^2$**, on a** $d(x, y) = 0$ **si, et seulement si,** $x = y$**,**
- **pour tout** $(x, y) \in M^2$**, on a** $d(x, y) = d(y, x)$**,**
- **pour tout** $(x, y, z) \in M^3$**, on a** $d(x, y) \leq d(x, z) + d(z, y)$**.**

Étude de la ressemblance

3. Exemples de Distances : soient x = (x1, . . . , xm) ∈ R^m et y = (y1, . . . , ym) ∈ R^m.

*** distance euclidienne : On appelle distance euclidienne entre *x et y la distance :***

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}.$$

*** distance de Manhattan : On appelle distance de Manhattan entre *x et y la distance***

$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i$$

Étude de la ressemblance

*** distance de Canberra : On appelle distance de Canberra entre *x et y la distance :***

$$d(x,y) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}$$

*** distance maximum : On appelle distance maximum entre *x et y la distance :***

$$d(x,y) = \sup_{i \in \{1,\dots,m\}} |x_i - y_i|$$

Étude de la ressemblance

4. **Tableau des distances :** Soit d une distance. On appelle tableau des distances associées aux individus (ω1, . . . , ωn) le tableau :

$\mathbf{D} =$

	ω_1	ω_2	. . .	ω_{n-1}	ω_n
ω_1	0	$d_{1,2}$	. . .	$d_{1,n-1}$	$d_{1,n}$
ω_2	$d_{2,1}$	0	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
ω_{n-1}	$d_{n-1,1}$	. . .	. . .	0	$d_{n-1,n}$
ω_n	$d_{n,1}$	. . .	. . .	$d_{n,n-1}$	0

Étude de la ressemblance

5. Écarts

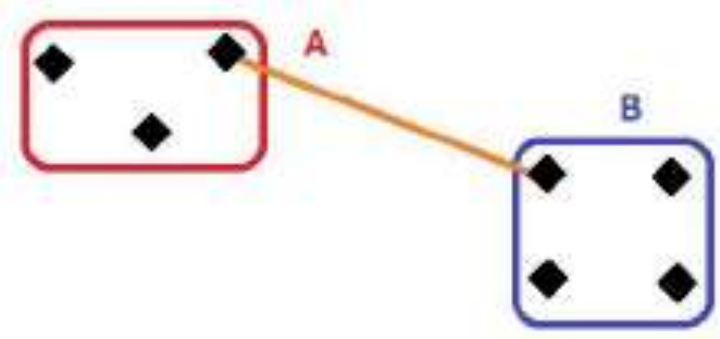
- Soit P(Γ) l'ensemble des parties de Γ. On appelle écart toute application $e : P(\Gamma)^2 \rightarrow [0, \infty[$ définie à partir d'une distance et évaluant la ressemblance entre deux groupes d'individus.

- **Règle centrale : Plus l'écart entre deux éléments est petit, plus ils se ressemblent.**

Étude de la ressemblance

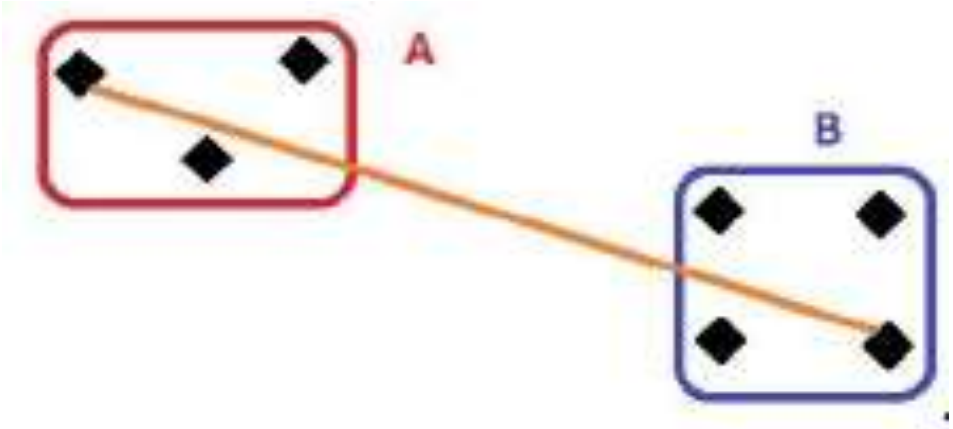
- **Écarts usuels :** Parmi les écarts usuels entre deux groupes A et B/méthodes usuelles mesurant la ressemblance entre deux groupes *A et B, il y a :*

  - Écart simple (single linkage)/Méthode du plus proche voisin : L'écart entre deux groupes A et B est caractérisé par la distance la plus faible entre un point de A et un point de B :



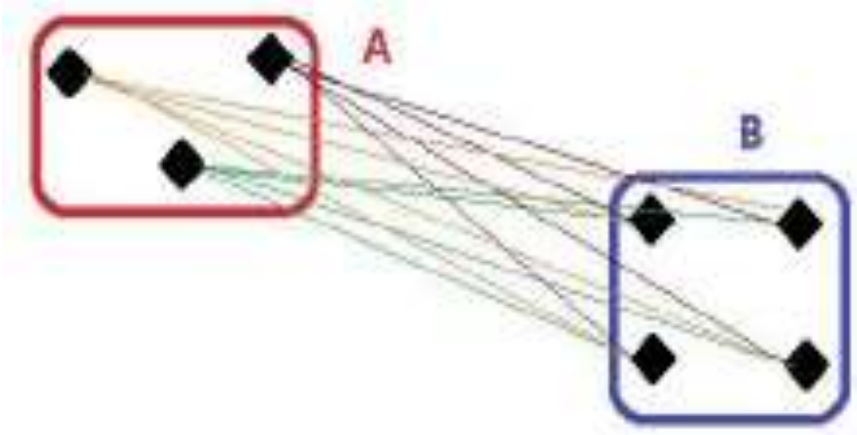
Étude de la ressemblance

- Écart complet (complete linkage)/Méthode du voisin le plus éloigné : L'écart entre deux groupes *A et B est caractérisé par la distance la plus forte entre un* point de *A et un point de B :*



Étude de la ressemblance

- Écart moyen (average linkage)/Méthode de la distance moyenne : L'écart entre deux groupes *A et B est caractérisé par la distance moyenne* entre les points de *A et B :*



Étude de la ressemblance

- Écart de Ward : Soit d la distance euclidienne. La méthode de Ward considère l'écart :

$$e(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B),$$

- où *gA est le centre de gravité de A, et gB celui de B.*
- Cette méthode prend en compte à la fois la dispersion à l'intérieur d'un groupe et la dispersion entre les groupes. Elle est utilisée par défaut dans la plupart des programmes informatiques. Elle fera l'objet d'un chapitre à venir.

Étude de la ressemblance

- Tableau des écarts : Soit e un écart défini par une des méthodes précédentes. On appelle tableau des écarts associé aux groupes d'individus (A1, . . . , An) le tableau :

$\mathbf{E} =$

	A_1	A_2	. . .	A_{n-1}	A_n
A_1	0	$e_{1,2}$	. . .	$e_{1,n-1}$	$e_{1,n}$
A_2	$e_{2,1}$	0	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
A_{n-1}	$e_{n-1,1}$	. . .	. . .	0	$e_{n-1,n}$
A_n	$e_{n,1}$	. . .	. . .	$e_{n,n-1}$	0

où,

pour tout $(u, v) \in \{1, .. , n\}2$ *avec* $u \neq v$, $e_{u,v} = e(Au, Av)$.

Algorithme de Classification Ascendante Hiérarchique (CAH)

- L'idée de l'algorithme de Classification Ascendante Hiérarchique (CAH) est de créer, à chaque étape, une partition de Γ = {*ω1, . . . , ωn} en regroupant les deux éléments les plus* proches. Le terme "élément" désigne aussi bien un individu qu'un groupe d'individus.

Algorithme de Classification Ascendante Hiérarchique (CAH)

- **Objectif :** On veut :

  - mettre en relief les liens hiérarchiques entre les individus ou groupe d'individus,

  - détecter les groupes d'individus qui se démarquent le plus.

Algorithme CAH: description

1. On choisit un écart. On construit le tableau des écarts pour la partition initiale des n individus de Γ : P0 = ({ω1}, . . . , {ωn}). Chaque individu constitue un élément.
2. On parcourt le tableau des écarts pour identifier le couple d'individus ayant l'écart le plus petit. Le regroupement de ces deux individus forme un groupe A. On a donc une partition de Γ de n − 1 éléments : A et les n − 2 individus restants.

Algorithme CAH: description

3. On calcule le tableau des écarts entre les n − 1 éléments obtenus à l'étape précédente et on regroupe les deux éléments ayant l'écart le plus petit. On a donc une partition de Γde n − 2 éléments.
4. On itère la procédure précédente jusqu'à ce qu'il ne reste que deux éléments.
5. On regroupe les deux éléments restants. Il ne reste alors qu'un seul élément contenant tous les individus de Γ.

Algorithme CAH: dendrogramme

- **Dendrogramme : Les partitions de Γ faites à chaque étape de l'algorithme de la CAH** peuvent se visualiser via un arbre appelé dendrogramme. Sur un axe apparait les individus à regrouper et sur l'autre axe sont indiqués les écarts correspondants aux différents niveaux de regroupement. Cela se fait graphiquement par le biais de branches et de nœuds.

  Une partition naturelle se fait en coupant l'arbre au niveau du plus grand saut de nœuds.

Algorithme CAH: exemple

On considère la matrice de données Ω dans R2 définie par

Ω	X_1	X_2
ω_1	2	2
ω_2	7.5	4
ω_3	3	3
ω_4	0.5	5
ω_5	6	4

- On va regrouper les individus avec l'algorithme CAH et la méthode du voisin le plus éloigné munie de la distance euclidienne. Ainsi, la distance entre les points est donnée par la table suivante :

Algorithme CAH: exemple

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	5.85	**1.41**	3.35	4.47
ω_2	5.85	0	4.60	7.07	1.50
ω_3	**1.41**	4.60	0	3.20	3,16
ω_4	3.35	7.07	3.20	0	5.59
ω_5	4.47	1.50	3.16	5.59	0

$$P_0 = (\{\omega_1\}, \ldots, \{\omega_5\})$$

Algorithme CAH: exemple

- Les éléments (individus) *ω1 et ω3 ont l'écart le plus petit : ce sont les éléments les plus* proches. On les rassemble pour former le groupe : *A = {ω1, ω3}. On a une nouvelle partition* de Γ : P1 = ({*ω2}, {ω4}, {ω5}, A).*

- Le tableau des écarts associé à P1 est

Algorithme CAH: exemple

	ω_2	ω_4	ω_5	A
ω_2	0	7.07	**1.50**	5.85
ω_4	7.07	0	5.59	3.35
ω_5	**1.50**	5.59	0	4.47
A	5.85	3.35	4,47	0

Algorithme CAH: exemple

- Les éléments (individus) *ω2 et ω5 sont les plus proches. On les rassemble pour former le* groupe : *B = {ω2, ω5}.*

  On a une nouvelle partition de Γ : P2 = ({ω4}, A, B).

- Le tableau des écarts associé à P2 est

Algorithme CAH: exemple

	ω_4	A	B
ω_4	0	**3.35**	7.07
A	**3.35**	0	5.85
B	7.07	5.85	0

Algorithme CAH: exemple

- Les éléments ω4 et A sont les plus proches. On les rassemble pour former le groupe : C = {ω4, A} = {ω1, ω3, ω4}.

  On a une nouvelle partition de Γ : P3 = (B, C).

- Le tableau des écarts associé à P3 est

Algorithme CAH: exemple

	B	C
B	0	7.07
C	7.07	0

Algorithme CAH: exemple

- Il ne reste plus que 2 éléments, B et C ; on les regroupe. On obtient la partition

$$P5 = \{\omega 1, \ldots, \omega 5\} = \Gamma.$$

Algorithme CAH: exemple

Au final,

- les éléments {ω1} et {ω3} ont été regroupés avec un écart de 1.41,
- les éléments {ω2} et {ω5} ont été regroupés avec un écart de 1.50,
- les éléments A = {ω1, ω3} et {ω4} ont été regroupés avec un écart de 3.35,
- les éléments C = {ω4, A} et B = {ω2, ω5} ont été regroupés avec un écart de 7.07.

Algorithme CAH: exemple