Apprentissage automatique non supervisé

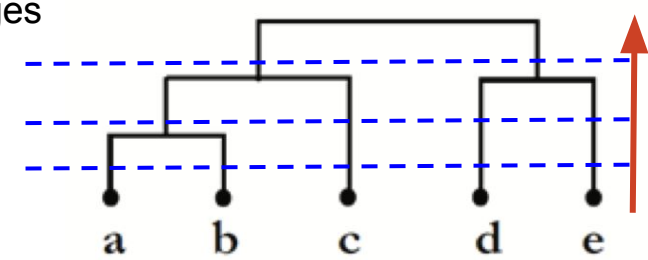
Amira Barhoumi

amira.barhoumi@univ-grenoble-alpes.fr

Année universitaire : 2025-2026

Classification Ascendante Hiérarchique CAH

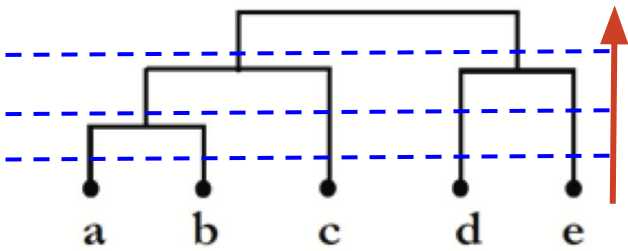
- Méthode d'apprentissage non supervisé (données non étiquetées).
- Le principe de la CAH consiste à rassembler des objets selon un critère de ressemblance traduit sous la forme d'un tableau des distances.
- Deux objets identiques ont une distance nulle.
- Plus les deux objets sont dissemblables, plus la distance est importante.
- La CAH rassemble les objets de manière itérative afin de produire un dendrogramme.
- La CAH est :
  - ascendante : partir des objets individuels
  - hiérarchique : produire des groupes de plus en plus larges



Classification Ascendante Hiérarchique CAH

- Résultat d'une CAH n'est pas une partition de l'ensemble des objets (par opposition à k-means)

- Résultat d'une CAH est une hiérarchie de classes telles que :
  - Toute classe est non vide
  - Tout objet appartient à une (et même plusieurs) classes
  - Deux classes distinctes sont disjointes ou vérifient une relation d'inclusion
  - Toute classe *C* est l'union des sous-classes qui sont incluses dans *C*



Classification Ascendante Hiérarchique CAH

- **Idée de CAH** : créer, à chaque étape, une partition de D = $\{\omega_1,...,\omega_n\}$ en regroupant les deux éléments les plus proches.

- Le terme "élément" désigne aussi bien un individu qu'un groupe d'individus.

- **Objectif de CAH**
  - mettre en relief les liens hiérarchiques entre les individus ou groupe d'individus
  - détecter les groupes d'individus qui se démarquent le plus.

- **Description de l'algorithme CAH**

Soient D = $\{\omega_1,...,\omega_n\}$ le jeu de données et E une distance

1- Construire le tableau des distances pour la partition initiale $P_0 = \{\{\omega_1\},...,\{\omega_n\}\}$ où chaque objet constitue un élément (sachant que "élément" désigne un objet ou un groupe d'objets).

2- Parcourir le tableau des distances et identifier le couple d'objets ayant la distance la plus petite.

Le regroupement de ces 2 objets forme le groupe (classe) A.

=> Une partition de n-1 éléments : A et les n-2 objets restants

3- Calculer le tableau des distances entre les n-1 éléments obtenus de l'étape 2- et identifier le couple d'éléments ayant la distance la plus petite et qui peut être composé de :

- 2 objets (parmi les n-2 objets restants)
- un objet parmi les n-2 objets restants et le groupe A

4- Répéter les étapes 2- et 3- jusqu'à ce qu'il ne reste que 2 éléments.

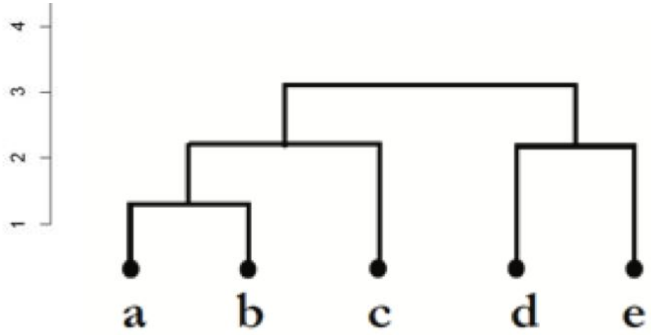
5- Regrouper les 2 éléments restants permettant d'obtenir une seule classe contenant tous les objets de D.

Classification Ascendante Hiérarchique CAH

- Résultat de l'algorithme CAH = Dendrogramme

Les partitions du jeu de données D = $\{\omega_1,...,\omega_n\}$ 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 distances correspondantes aux différents niveaux de regroupement. Cela se fait graphiquement par le biais de branches et de nœuds.



Classification Ascendante Hiérarchique CAH

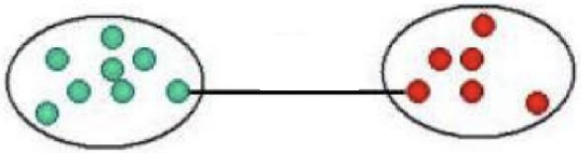
- Exemple 1 :

Soit le jeu de données D = {$\omega_1(2, 2)$, $\omega_2(7.5, 4)$, $\omega_3(3, 3)$, $\omega_4(0.5, 5)$, $\omega_5(6, 4)$}

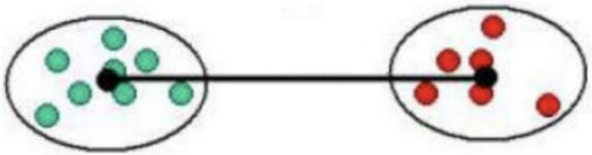
Appliquer l'algorithme de la CAH en utilisant

la méthode du voisin le plus éloigné (lien complet)

munie de la distance euclidienne



Lien simple



Lien moyen

Lien complet

- Exemple 1 (suite 1) :

Soit le jeu de données D = {$\omega_1(2, 2)$, $\omega_2(7.5, 4)$, $\omega_3(3, 3)$, $\omega_4(0.5, 5)$, $\omega_5(6, 4)$}

1- Partition initiale P_0 = {{ω_1},...,{ω_5}}

2- Tableau des distance associé à P_0 = {{ω_1},..., {ω_5}} est :

- Exemple 1 (suite 1) :

Soit le jeu de données D = {$\omega_1(2, 2)$, $\omega_2(7.5, 4)$, $\omega_3(3, 3)$, $\omega_4(0.5, 5)$, $\omega_5(6, 4)$}

1- Partition initiale P_0 = {{ω_1},...,{ω_5}}

2- Tableau des distance associé à P_0 = {{ω_1},..., {ω_5}} est :

	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

- Exemple 1 (suite 1) :

Soit le jeu de données D = {ω_1(2, 2), ω_2(7.5, 4), ω_3(3, 3), ω_4(0.5, 5), ω_5(6, 4)}

1- Partition initiale P_0 = {{ω_1},...,{ω_5}}

2- Tableau des distance associé à P_0 = {{ω_1},..., {ω_5}} est :

	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

Les éléments ω_1 et ω_3 ont la distance la plus petite => ω_1 et ω_3 sont les éléments les plus proches

⇒ On les rassemble pour former le groupe : A = {ω_1, ω_3}

⇒ On obtient une nouvelle partition de D : P_1 ={{ω_2},{ω_4},{ω_5},A}

- Exemple 1 (suite 2) :

3- Tableau des distance associé à P_1={{ω_2}, {ω_4}, {ω_5}, A} est :

- Exemple 1 (suite 2) :

3- Tableau des distance associé à P_1={{ω_2}, {ω_4}, {ω_5}, A} est :

	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

On a :

$d(\omega_2, A) = \max(d(\omega_2, \omega_1), d(\omega_2, \omega_3)) = \max(5.85, 4.60) = 5.85$

$d(\omega_4, A) = \max(d(\omega_4, \omega_1), d(\omega_4, \omega_3)) = \max(3.35, 3.20) = 3.35$

$d(\omega_5, A) = \max(d(\omega_5, \omega_1), d(\omega_5, \omega_3)) = \max(4.47, 3.16) = 4.47$

- Exemple 1 (suite 2) :

3- Tableau des distance associé à P_1={{ω_2}, {ω_4}, {ω_5}, A} est :

	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

On a :

$d(\omega_2, A) = \max(d(\omega_2, \omega_1), d(\omega_2, \omega_3)) = \max(5.85, 4.60) = 5.85$

$d(\omega_4, A) = \max(d(\omega_4, \omega_1), d(\omega_4, \omega_3)) = \max(3.35, 3.20) = 3.35$

$d(\omega_5, A) = \max(d(\omega_5, \omega_1), d(\omega_5, \omega_3)) = \max(4.47, 3.16) = 4.47$

Les éléments ω_2 et ω_5 ont la distance la plus petite => ω_2 et ω_5 sont les éléments les plus proches

⇒ On les rassemble pour former le groupe : B = {ω_2, ω_5}

⇒ On obtient une nouvelle partition de D : P_2 ={{ω_4}, A, B}

- Exemple 1 (suite 3) :

4- Tableau des distance associé à P_2={{ω_4}, A, B} est :

- Exemple 1 (suite 3) :

4- Tableau des distance associé à P_2={{ω_4}, A, B} est :

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

On a :

$d(\omega_4, B) = \max(d(\omega_4, \omega_2), d(\omega_4, \omega_5)) = \max(7.07, 5.59) = 7.07$

$d(\omega_4, A) = \max(d(\omega_4, \omega_1), d(\omega_4, \omega_3)) = \max(1.41, 3.35) = 3.35$

$d(B, A) = \max(d(\omega_2, A), d(\omega_5, A)) = \max(5.85, 4.47) = 5.85$

- Exemple 1 (suite 3) :

4- Tableau des distance associé à P_2={{ω_4}, A, B} est :

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

On a :

$d(\omega_4, B) = \max(d(\omega_4, \omega_2), d(\omega_4, \omega_5)) = \max(7.07, 5.59) = 7.07$

$d(\omega_4, A) = \max(d(\omega_4, \omega_1), d(\omega_4, \omega_3)) = \max(1.41, 3.35) = 3.35$

$d(B, A) = \max(d(\omega_2, A), d(\omega_5, A)) = \max(5.85, 4.47) = 5.85$

Les éléments ω_4 et A ont la distance la plus petite => ω_4 et A sont les éléments les plus proches

⇒ On les rassemble pour former le groupe : C = {ω_4, A} = {ω_1, ω_3, ω_4}

⇒ On obtient une nouvelle partition de D : P_3 ={B, C}

- Exemple 1 (suite 4) :

5- Tableau des distance associé à P_3 ={B, C} est :

- Exemple 1 (suite 4) :

5- Tableau des distance associé à P_3 ={B, C} est :

	B	C
B	0	**7.07**
C	**7.07**	0

On a :

$d(C, B) = \max(d(\omega_4, B), d(A, B)) = \max(7.07, 5.85) = 7.07$

- Exemple 1 (suite 4) :

5- Tableau des distance associé à P_3 ={B, C} est :

	B	C
B	0	**7.07**
C	**7.07**	0

On a :

d(C, B) = max(d(ω_4, B), d(A, B)) = max(7.07, 5.85) = 7.07

Fin de l'algorithme CAH (2 éléments finaux!)

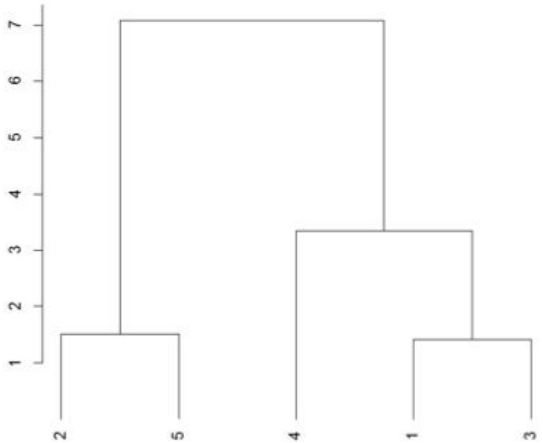
On rassemble les 2 éléments B et C

⇒ On obtient une nouvelle partition de D : P_4 = D

- Exemple 1 (suite 5) :

Résultat de l'algorithme CAH sur le jeu de données D :

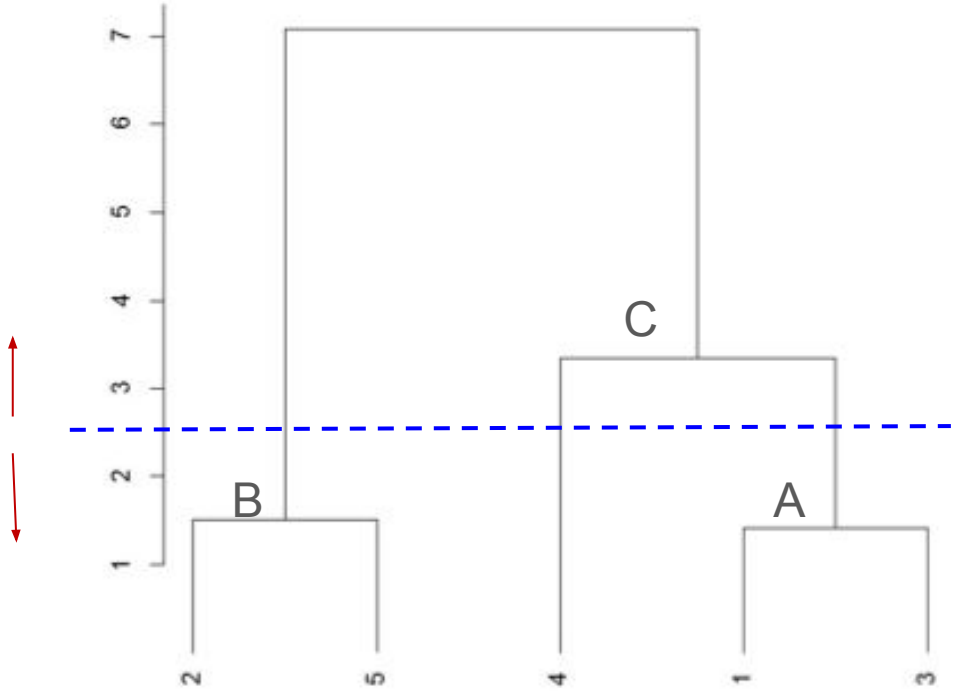
- Les éléments $\{\omega_1\}$ et $\{\omega_3\}$ ont été regroupés (en A) avec une distance de 1.41
- Les éléments $\{\omega_2\}$ et $\{\omega_5\}$ ont été regroupés (en B) avec une distance de 1.50
- Les éléments A = $\{\omega_1, \omega_3\}$ et $\{\omega_4\}$ ont été regroupés (en C= $\{\omega_4, A\}$) avec une distance de 3.35
- Les éléments C = $\{\omega_4, A\}$ et B = $\{\omega_2, \omega_5\}$ ont été regroupés avec une distance de 7.07



- Exemple 1 (suite 6) :

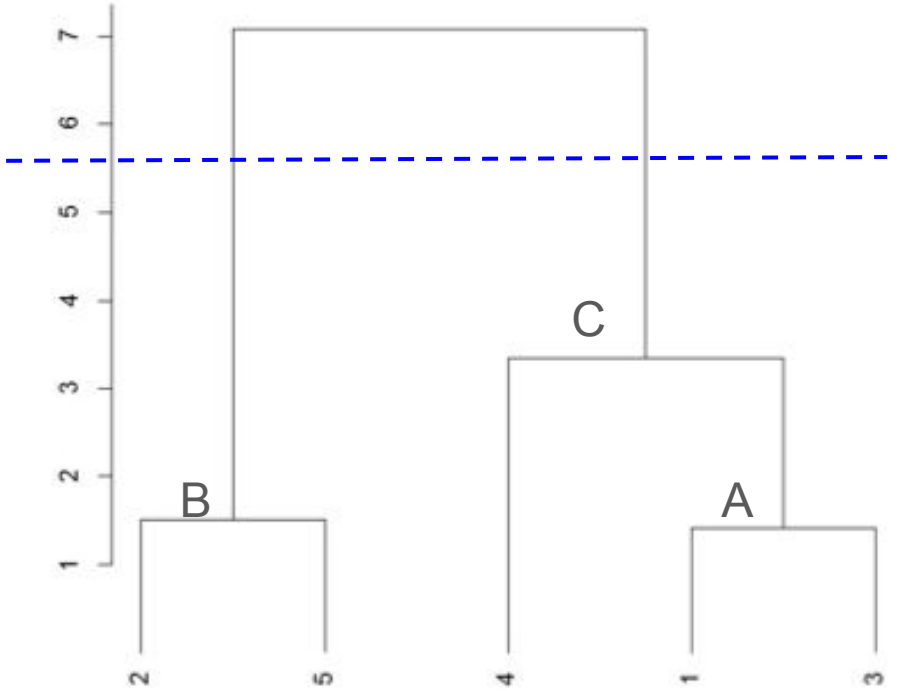
Résultat de l'algorithme CAH sur le jeu de données D :

Groupes ?



- Exemple 1 (suite 6) :

Résultat de l'algorithme CAH sur le jeu de données D :



Groupes ?

Le plus grand saut se situe entre les éléments

B et C (7.07 - 3.35 = 3.72)

⇒ Proposer les 2 groupes B et C

- Exemple 2 :

Soit D ={1, 2, 9, 12, 20}

Appliquer l'algorithme de classification ascendante hiérarchique en utilisant la méthode du voisin le plus proche (lien minimal) munie de la distance euclidienne.

Tracer le dendrogramme correspondant

Classification Ascendante Hiérarchique CAH

- Avantages
  - Obtenir une hiérarchie de classes
  - Choisir le nombre de classes optimal

- Inconvénients
  - Non adaptée au jeu de données volumineux
  - Valeurs d’attributs non numériques