Intelligence Artificielle et Apprentissage

Cours 3 : clustering hiérarchique

Adrien Revault d'Allonnes

ara@up8.edu

Université Paris 8 – Vincennes à Saint-Denis

IAA – S2 – 2024

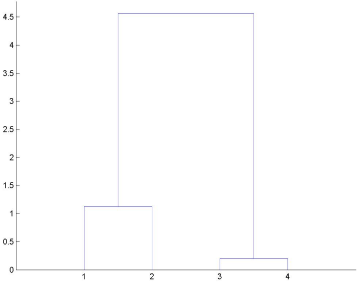
Clustering hiérarchique : un aperçu*

- Révéler des structures dans un arbre de clusters
- Deux variantes :
  - descendante (V.O. : *divisive*)
  - ascendante (V.O. : *agglomerative*)
- Représentation usuelle : le dendrogramme

**** encore très inspiré d'Eamonn Keogh***

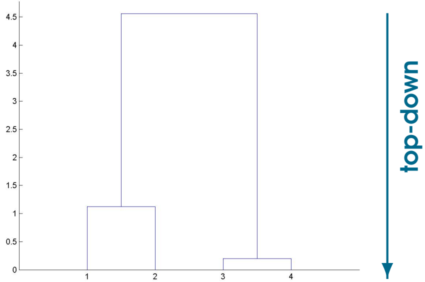
Clustering hiérarchique : un aperçu

- Révéler des structures dans un arbre de clusters
- Deux variantes :
  - descendante (V.O. : *divisive*)
  - ascendante (V.O. : *agglomerative*)
- Représentation usuelle : **le dendrogramme**



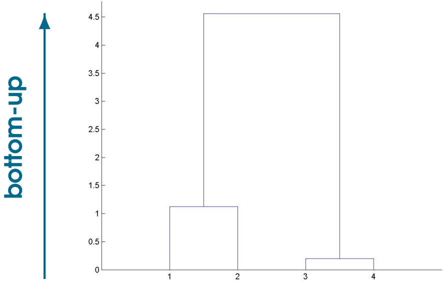
Clustering hiérarchique : un aperçu

- Révéler des structures dans un arbre de clusters
- Deux variantes :
  - **descendante (V.O. : *divisive*)**
  - ascendante (V.O. : *agglomerative*)
- Représentation usuelle : le dendrogramme



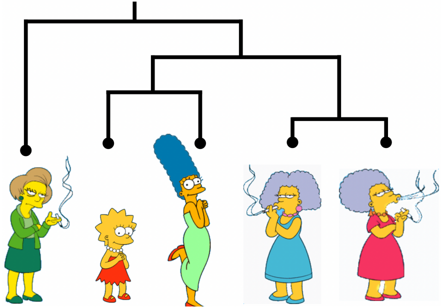
Clustering hiérarchique : un aperçu

- Révéler des structures dans un arbre de clusters
- Deux variantes :
  - descendante (V.O. : *divisive*)
  - **ascendante (V.O. : *agglomerative*)**
- Représentation usuelle : le dendrogramme



Clustering hiérarchique : un aperçu

- Révéler des structures dans un arbre de clusters
- Deux variantes :
  - descendante (V.O. : *divisive*)
  - **ascendante (V.O. : *agglomerative*)**
- Représentation usuelle : le dendrogramme



Comment on fait ?

- Nombre de dendrogrammes pour n feuilles :

$$\frac{(2n-3)!}{2^{n-2} \times (n-2)!}$$

n	# dendro.
2	1
3	3
4	15
5	105
$\vdots$	$\vdots$
10	$34\,459\,425$

Comment on fait ?

- Nombre de dendrogrammes pour n feuilles :

$$\frac{(2n-3)!}{2^{n-2} \times (n-2)!}$$

n	# dendro.
2	1
3	3
4	15
5	105
$\vdots$	$\vdots$
10	$34\,459\,425$

- Impossible de lister tous les dendrogrammes

Comment on fait ?

- Nombre de dendrogrammes pour n feuilles :

$$\frac{(2n-3)!}{2^{n-2} \times (n-2)!}$$

n	# dendro.
2	1
3	3
4	15
5	105
⋮	⋮
10	$34\,459\,425$

- Impossible de lister tous les dendrogrammes

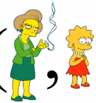
⇒ Construction heuristique

Construction ascendante

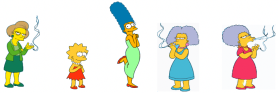
- Au départ, chaque point dans son propre cluster
- Choisir la meilleure paire à fusionner
- Répéter jusqu'à ce que tous les clusters soient fusionnés

Construction ascendante

- On commence avec la matrice des distances

D() = 8

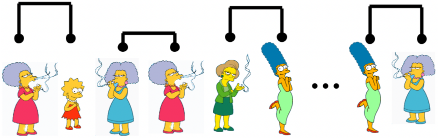
D() = 1



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

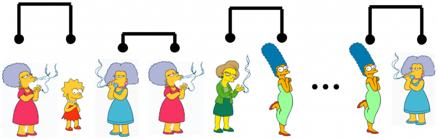
Construction ascendante

de toutes
les fusions

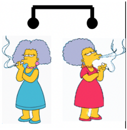


Construction ascendante

de toutes
les fusions

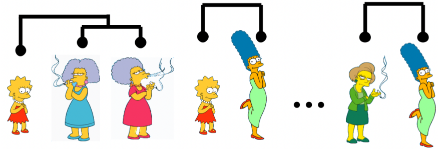


choisir la
meilleure

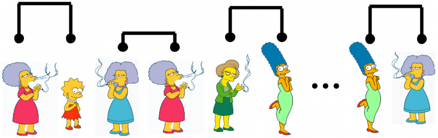


Construction ascendante

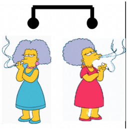
de toutes
les fusions



de toutes
les fusions

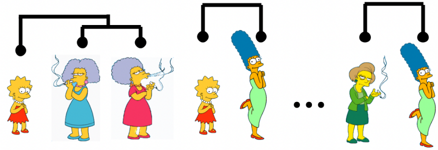


choisir la
meilleure

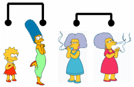


Construction ascendante

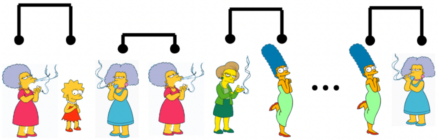
de toutes
les fusions



choisir la
meilleure



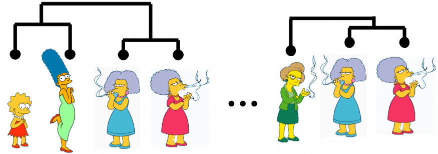
de toutes
les fusions



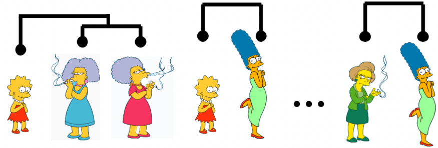
choisir la
meilleure

Construction ascendante

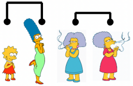
de toutes
les fusions



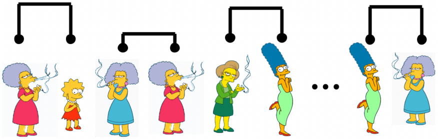
de toutes
les fusions



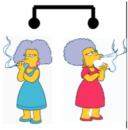
choisir la
meilleure



de toutes
les fusions

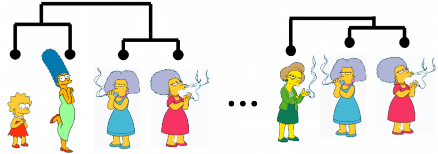


choisir la
meilleure

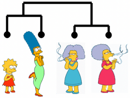


Construction ascendante

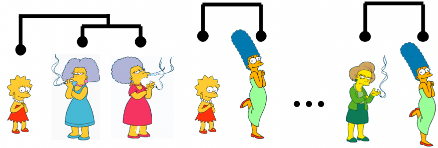
de toutes
les fusions



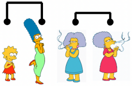
choisir la
meilleure



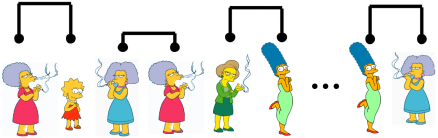
de toutes
les fusions



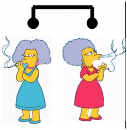
choisir la
meilleure



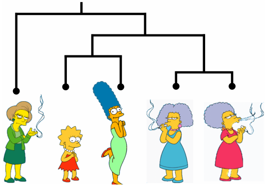
de toutes
les fusions



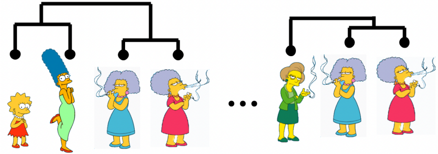
choisir la
meilleure



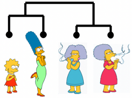
Construction ascendante



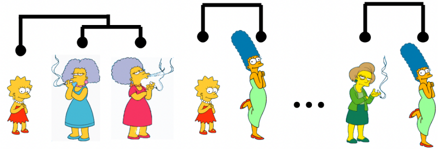
de toutes
les fusions



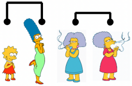
choisir la
meilleure



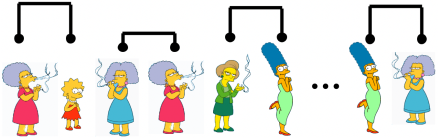
de toutes
les fusions



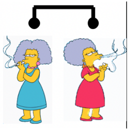
choisir la
meilleure



de toutes
les fusions



choisir la
meilleure



Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?

Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?
- Différentes méthodes de liaison (V.O. : *linkage*) :

Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?
- Différentes méthodes de liaison (V.O. : *linkage*) :
  - simple linkage : $\min_{a \in A, b \in B} d(a, b)$

    distance entre les deux plus proches voisins, un dans chaque cluster

Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?
- Différentes méthodes de liaison (V.O. : *linkage*) :
  - simple linkage : $\min_{a \in A, b \in B} d(a,b)$

    distance entre les deux plus proches voisins, un dans chaque cluster
  - complete linkage : $\max_{a \in A, b \in B} d(a,b)$

    distance entre les deux points les plus éloignés des deux clusters

Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?
- Différentes méthodes de liaison (V.O. : *linkage*) :
  - simple linkage : $\min_{a \in A, b \in B} d(a, b)$

    distance entre les deux plus proches voisins, un dans chaque cluster
  - complete linkage : $\max_{a \in A, b \in B} d(a, b)$

    distance entre les deux points les plus éloignés des deux clusters
  - group average linkage : $\frac{1}{|A| \times |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$

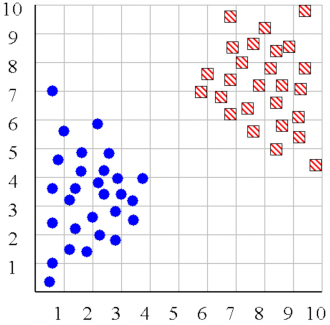
    moyenne des distances de toutes les paires de points entre clusters

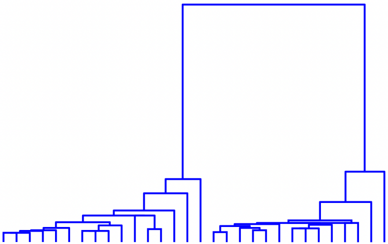
Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?
- Différentes méthodes de liaison (V.O. : *linkage*) :
  - simple linkage : $\min_{a \in A, b \in B} d(a, b)$
    distance entre les deux plus proches voisins, un dans chaque cluster
  - complete linkage : $\max_{a \in A, b \in B} d(a, b)$
    distance entre les deux points les plus éloignés des deux clusters
  - group average linkage : $\dfrac{1}{|A| \times |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$
    moyenne des distances de toutes les paires de points entre clusters
  - Ward linkage : $\sum_{x \in A \cup B} ||x - \mu_{A \cup B}||^2 - \sum_{x \in A} ||x - \mu_A||^2 - \sum_{x \in B} ||x - \mu_B||^2$
    minimisation de la variance des clusters fusionnés

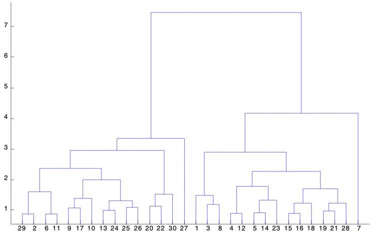
Méthodes de Liaison

- On sait comparer deux données
- ¿ Comment comparer une donnée et un cluster ?
- ¿ Comment comparer deux clusters ?
- Différentes méthodes de liaison (V.O. : *linkage*) :
  - simple linkage : $\min_{a \in A, b \in B} d(a, b)$
    distance entre les deux plus proches voisins, un dans chaque cluster
  - complete linkage : $\max_{a \in A, b \in B} d(a, b)$
    distance entre les deux points les plus éloignés des deux clusters
  - group average linkage : $\dfrac{1}{|A| \times |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$
    moyenne des distances de toutes les paires de points entre clusters
  - Ward linkage : $\sum_{x \in A \cup B} ||x - \mu_{A \cup B}||^2 - \sum_{x \in A} ||x - \mu_A||^2 - \sum_{x \in B} ||x - \mu_B||^2$
    minimisation de la variance des clusters fusionnés
- plein d'autres `https://en.wikipedia.org/wiki/Hierarchical_clustering`

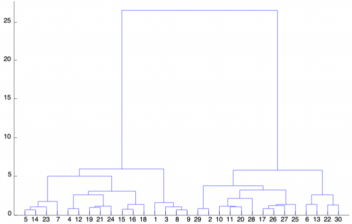




Single linkage



Average linkage



Wards linkage

Conclusion

- **Points forts :**
  - pas besoin de spécifier nombre de clusters
  - côté hiérarchique intuitif dans certains domaines

- **Points faibles :**
  - pas de passage à l'échelle :
    complexité au moins $O(n^2)$ pour n objets
  - optima locaux courants avec méthodes heuristiques
  - interprétation pas toujours évidente