Clustering Ascendant Hiérarchique

1 Clustering Ascendant Hiérarchique (CAH)

Le Clustering Ascendant Hiérarchique (CAH) est une méthode de classification non supervisée visant à regrouper des individus similaires en clusters, sans fixer à l'avance le nombre de groupes. Il s'agit d'une méthode **agglomérative** : chaque individu commence dans son propre cluster, puis les clusters sont fusionnés progressivement jusqu'à n'en former qu'un seul.

1.1 Principe général

Soit un ensemble de données :

$$X = \{x_1, x_2, \ldots, x_n\}, \quad x_i \in \mathbb{R}^p$$

Au départ, chaque point constitue un cluster :

$$C_i = \{x_i\}, \quad i = 1, \ldots, n$$

À chaque itération :

1. on calcule la distance entre tous les clusters,
2. on fusionne les deux clusters les plus proches,
3. on met à jour les distances entre le nouveau cluster et les autres,

jusqu'à obtenir un unique cluster contenant tous les individus.

L'ensemble des fusions successives forme une hiérarchie, représentée par un dendrogramme.

1.2 Distance entre individus

La distance la plus couramment utilisée est la distance euclidienne :

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$$

D'autres distances peuvent être utilisées selon la nature des données (Manhattan, cosinus, Hamming, etc.). En pratique, il est recommandé de standardiser les variables afin d'éviter qu'une variable à grande échelle ne domine la distance.

1.3 Distance entre clusters (linkage)

Soient deux clusters A et B. La distance entre clusters dépend du critère de liaison (*linkage*) choisi.

1.3.1 Distance minimale (Single Linkage)

La distance entre deux clusters est définie comme la plus petite distance entre deux points appartenant à des clusters différents :

$$D_{\text{single}}(A, B) = \min_{x \in A,\, y \in B} d(x, y)$$

Interprétation :

— favorise les effets de chaînage (clusters allongés),

— sensible aux points reliant artificiellement deux groupes,

— adapté pour détecter des formes non convexes.

1.3.2 Distance maximale (Complete Linkage)

La distance entre deux clusters est la plus grande distance entre deux points des clusters :

$$D_{\text{complete}}(A, B) = \max_{x \in A,\, y \in B} d(x, y)$$

Interprétation :

— produit des clusters compacts,

— réduit l'effet de chaînage,

— sensible aux valeurs aberrantes (outliers).

1.3.3 Distance moyenne (Average Linkage)

La distance est la moyenne de toutes les distances entre les points des deux clusters :

$$D_{\text{avg}}(A, B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$$

Interprétation :

— compromis entre single et complete linkage,

— moins sensible au chaînage,

— comportement généralement stable.

1.4 Méthode de Ward

La méthode de Ward ne repose pas directement sur une distance inter-cluster, mais sur la minimisation de la variance intra-cluster.

Pour un cluster C de centre μ_C :

$$\text{SSE}(C) = \sum_{x \in C} \|x - \mu_C\|^2$$

La fusion de deux clusters A et B est choisie de façon à minimiser l'augmentation d'inertie :

$$\Delta(A, B) = \mathrm{SSE}(A \cup B) - \mathrm{SSE}(A) - \mathrm{SSE}(B)$$

Dans le cas de la distance euclidienne, on obtient :

$$\Delta(A, B) = \frac{|A||B|}{|A| + |B|} \|\mu_A - \mu_B\|^2$$

Interprétation :

— favorise des clusters homogènes et compacts,

— très utilisé avec des données quantitatives standardisées,

— proche de la logique du *k-means*.

1.5 Interprétation du dendrogramme

Le dendrogramme représente l'ensemble des fusions :

— les feuilles correspondent aux individus,

— la hauteur d'une fusion représente la distance entre les clusters fusionnés.

Une fusion à grande hauteur indique que les clusters étaient très dissemblables. Le nombre de clusters est déterminé en coupant le dendrogramme à une hauteur donnée, généralement juste avant une augmentation brutale de la distance de fusion.

1.6 Conclusion

Le choix du critère de liaison influence fortement la structure finale des clusters :

— **single linkage** : connectivité,

— **complete linkage** : compacité stricte,

— **average linkage** : compromis,

— **Ward** : minimisation de la variance intra-cluster.

La CAH permet ainsi une analyse exploratoire riche, fournissant une vision hiérarchique de la structure des données.

2 Dendrogramme : définition mathématique

2.1 Structure hiérarchique

Soit $X = \{x_1, x_2, \ldots, x_n\}$ un ensemble de n individus. Un dendrogramme est une représentation graphique d'une suite de partitions hiérarchiques

$$\mathcal{P}_0 \succ \mathcal{P}_1 \succ \cdots \succ \mathcal{P}_{n-1},$$

où :

— $\mathcal{P}_0 = \big\{\{x_1\}, \ldots, \{x_n\}\big\}$ est la partition initiale,

— $\mathcal{P}_{n-1} = \{X\}$ est la partition finale,

— $\succ$ désigne une relation de raffinement (fusion de clusters).

À chaque étape t, la partition $\mathcal{P}_t$ contient $n - t$ clusters.

2.2 Fonction de hauteur (niveau de fusion)

À l'étape t, deux clusters A_t et B_t sont fusionnés. On associe à cette fusion une valeur réelle positive appelée *hauteur de fusion* :

$$h_t = D(A_t, B_t),$$

où $D(\cdot, \cdot)$ est la dissimilarité inter-clusters définie par le critère de liaison (single, complete, average, Ward). [1.2,1.4]

La suite $(h_t)_{t=1}^{n-1}$ est croissante :

$$0 \leq h_1 \leq h_2 \leq \cdots \leq h_{n-1}.$$

2.3 Dendrogramme comme arbre métrique

Le dendrogramme peut être vu comme un arbre binaire enraciné :

— les feuilles correspondent aux individus x_i,

— chaque nœud interne représente la fusion de deux clusters,

— la hauteur du nœud est égale à h_t.

Cet arbre induit une *ultramétrique* d_u sur X, définie par :

$$d_u(x_i, x_j) = \min \big\{ h_t \mid x_i \text{ et } x_j \text{ sont dans le même cluster à l'étape } t \big\}.$$

La fonction d_u vérifie la propriété d'ultramétrie :

$$d_u(x_i, x_j) \leq \max \big(d_u(x_i, x_k), d_u(x_k, x_j) \big), \quad \forall x_i, x_j, x_k \in X.$$

2.4 Coupe du dendrogramme

Soit $\lambda > 0$ un seuil de hauteur. La coupe horizontale du dendrogramme à la hauteur λ définit une partition :

$$\mathcal{P}_\lambda = \{ C \subset X \mid \forall x_i, x_j \in C,\ d_u(x_i, x_j) \leq \lambda \}.$$

Le nombre de clusters est égal au nombre de composantes connexes après la coupe.

2.5 Interprétation mathématique

— Une grande valeur de h_t indique une fusion entre clusters très dissemblables.

— Un saut important entre h_t et h_{t+1} suggère une séparation naturelle des données.

— La hauteur de coupe λ contrôle le compromis entre granularité et homogénéité.

2.6 Résumé

Le dendrogramme est une représentation hiérarchique équivalente :

— à une suite de partitions emboîtées,

— à un arbre binaire pondéré,

— à une ultramétrique définie sur l'ensemble des individus.