Thématique 6 - Le clustering

Data Sciences Revision

Mention Mathématiques et Informatique

Parcours HPDA

8/10/2024

Rédigé par :
PAULY Alexandre
alexandre.pauly@cy-tech.fr

L'objectif des méthodes de clustering est de distinguer des groupes homogènes (classes, segments, clusters) au sein d'un grand volume de données. Le clustering est la technique non supervisée la plus répandue en datamining. De part leur constitution, ces groupes peuvent apporter une information pertinente sur les données, notamment s'ils sont représentés graphiquement à l'aide d'une ACP. Mais ils peuvent aussi servir à découper une étude en sous-parties, chacune pouvant bénéficier de traitements particuliers.

L'objectif des méthodes est à la fois, de regrouper les observations ayant des caractéristiques similaires au sein d'une même classe, et à la fois de construire des classes les plus dissemblables possibles.

1 Métriques sur les classes

Pour trouver des similarités entre les observations il faut définir une métrique sur les observations.

Names	Equations
Euclidean distance	$\|x_i - x_j\|_2 = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2}$
Manhattan distance	$\|x_i - x_j\|_1 = \|a_i - a_j\| + \|b_i - b_j\|$
Maximum distance	$\|x_i - x_j\|_\infty = \max\{\|a_i - a_j\|, \|b_i - b_j\|\}$
Mahalanobis distance	$\sqrt{(x_i - x_j)^T S^{-1}(x_i - x_j)}$, where S is the covariance matrix and x_i and x_j are variable vectors of x_i and x_j

x_i and x_j are ith and jth observations, where i and j are indices. a and b are feature variables.

Figure 1: Distances sur observations

Pour construire des classes dissemblables il faut définir une métrique sur les classes. Pour cela, nous disposons de plusieurs méthodes (distances).

Distance Minimale (Single Linkage) : fusionne les clusters en minimisant la distance entre les points les plus proches de deux clusters :

$$d_{\min}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

Distance Maximale (Complete Linkage) : fusionne les clusters en maximisant la distance entre les points les plus éloignés de deux clusters :

$$d_{\max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

Distance Moyenne (Average Linkage) : fusionne les clusters en utilisant la moyenne des distances entre tous les points de deux clusters :

$$d_{\text{moy}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

Méthode de Ward (Ward's Method) : fusionne les clusters en minimisant la variance totale au sein des clusters. Elle cherche à minimiser l'augmentation de la somme des carrés des distances des points au centre de leur cluster respectif :

$$d_{\text{Ward}}(C_i, C_j) = \frac{|C_i||C_j|}{|C_i| + |C_j|} \|\mu_i - \mu_j\|^2$$

où μ_i et μ_j sont les centroïdes des clusters C_i et C_j.

2 Inertie inter et intra classes

L'inertie cherche la partition qui minimise l'inertie intra-classes (homogénéité des observations dans les classes) et la partition qui maximise l'inertie inter-classes (dissimilarité des classes entre elles).

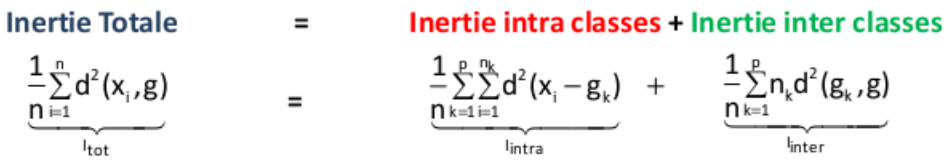

Figure 2: Formule de l'inertie

Le coefficient

$$R^2 = \frac{I_{inter}}{I_{intra}}$$

est le pourcentage d'inertie du nuage expliquée par les classes. L'objectif est d'obtenir un

$$R^2$$

proche de 1 avec un minimum de classes. Il peut servir pour :

- Comparer deux partitionnements ayant le même nombre de classes
- Sélectionner le nombre de classes

3 Méthodes de Clustering

3.1 Méthodes des K-Means

K-Means partitionne les données en kk clusters en minimisant la somme des distances des points à leur centroïde. Il fonctionne en plusieurs étapes :

1. Initialisation de kk centroïdes (aléatoire ou par une méthode comme K-Means++).
2. Assignation de chaque point au centroïde le plus proche.

3. Recalcul des centroïdes en prenant la moyenne des points assignés.

4. Répétition jusqu'à convergence.

Le problème c'est que cela nécessite de choisir k à l'avance, sensible aux outliers, il fonctionne mal avec des clusters non sphériques.

Pour savoir quel k choisir, il est utile de déterminer la valeur du coude. Le "coude" dans le graphe de K-Means est utilisé pour déterminer le nombre optimal de clusters k. En traçant la somme des distances intra-cluster (inertie) en fonction de k, on observe généralement que l'inertie diminue à mesure que k augmente.

Le point où la courbe commence à "se plier", formant un coude, indique un compromis entre le nombre de clusters et la réduction de l'inertie. Ce point marque souvent un choix raisonnable pour k, car augmenter davantage le nombre de clusters ne réduit plus significativement l'inertie, évitant un surajustement des données.

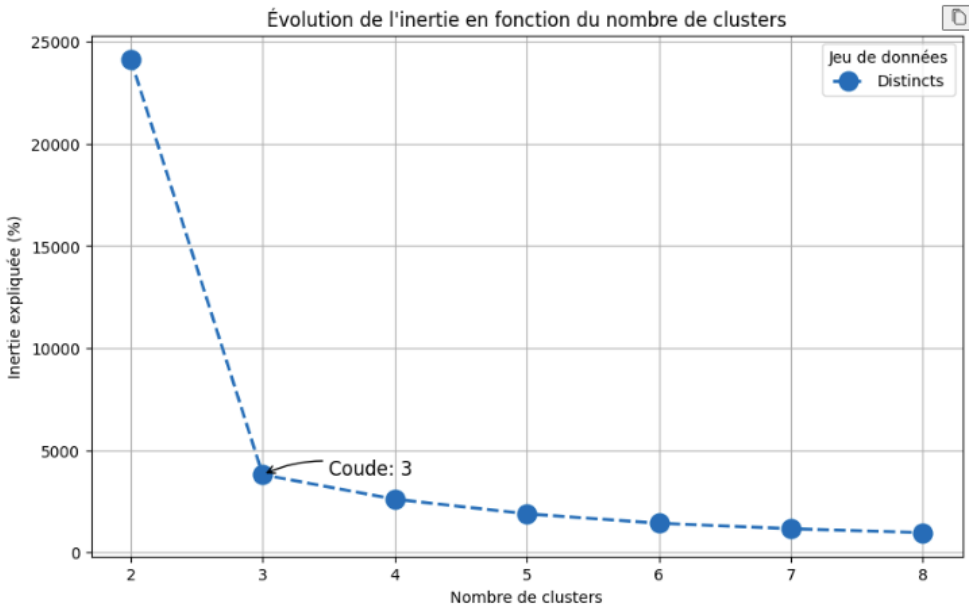

Figure 3: Coude du graphe des K-Means

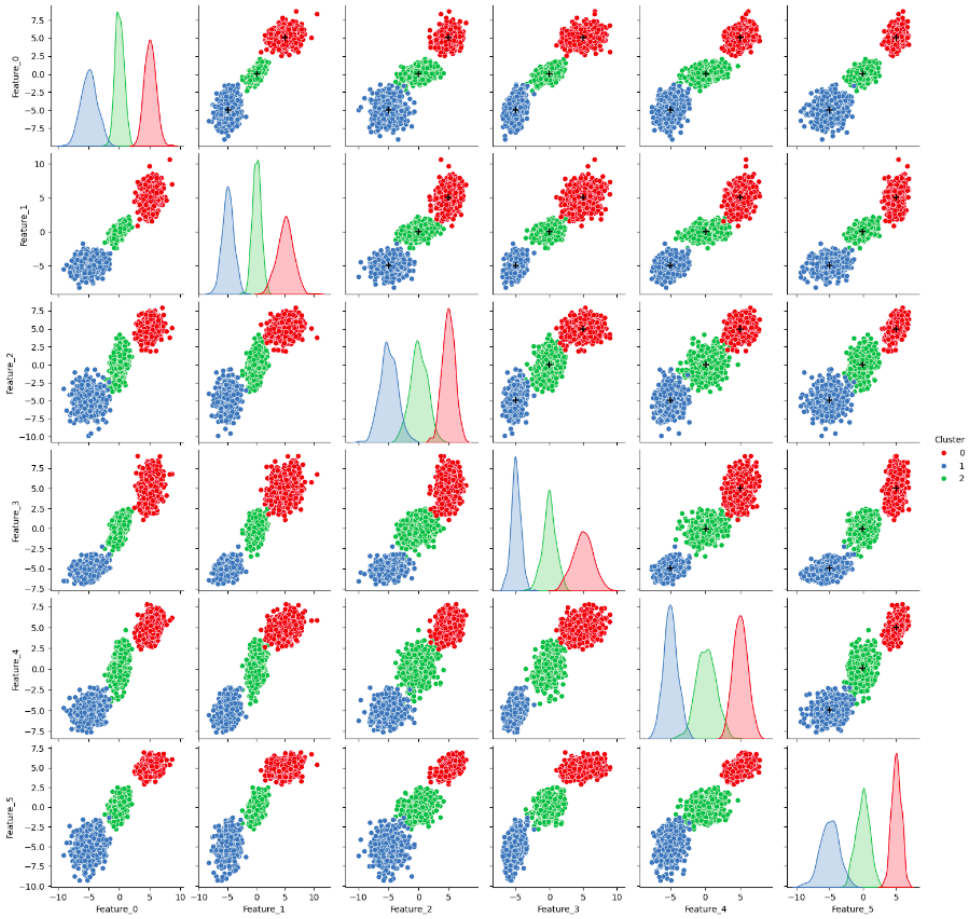

Figure 4: Pairplot des clusters et leurs centres

3.2 Classification Hiérarchique Ascendante

La CAH fusionne progressivement les points en clusters hiérarchisés, formant un dendrogramme. L'algorithme fonctionne en plusieurs étapes :

1. Chaque point commence comme son propre cluster.
2. À chaque étape, les deux clusters les plus proches sont fusionnés.
3. Le processus se poursuit jusqu'à ce qu'il ne reste qu'un seul cluster.

Le problème c'est que cela nécessite de choisir k à l'avance, sensible à la méthode de distance, coûteux en temps pour de grands datasets.

Le résultat de la CAH est résumé par un dendogramme sur lequel il faut suffit de regarder où couper le dendrogramme. Par défaut, il faut garder les clusters minimisant la distance entre eux.

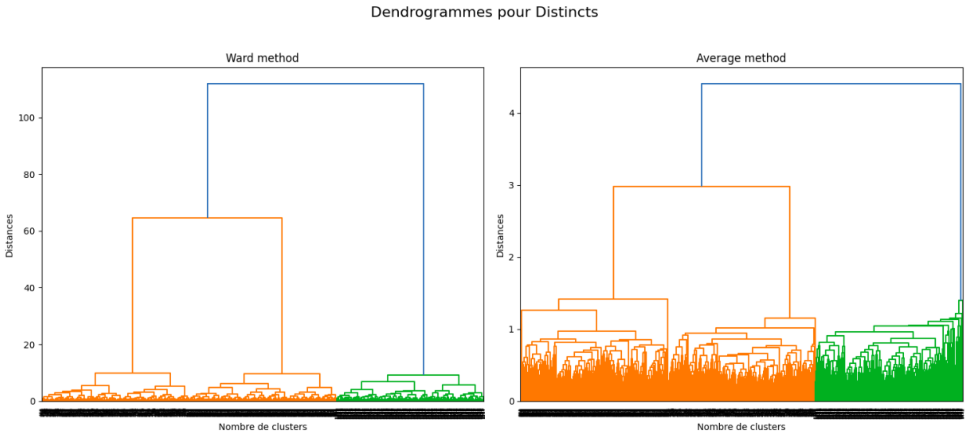

Figure 5: Dendogramme de la CAH

3.3 DBSCAN

DBSCAN identifie des clusters denses de points en explorant les voisins proches tout en marquant les points isolés comme du bruit. Il fonctionne en plusieurs étapes :

1. Définir deux paramètres : ϵ (rayon de voisinage) et MinPts (nombre minimal de points dans un voisinage pour former un cluster).
2. Un point est un noyau s'il a au moins MinPts voisins dans un rayon .
3. Les clusters se forment en reliant les points denses.
4. Les points hors de ces zones sont considérés comme du bruit.

Il a pour avantage de détecter les formes irrégulières, gèrer le bruit, et ne nécessite pas de définir k. Par contre, il est sensible au choix de ϵ et du nombre minimal de points.