Méthodes hiérarchiques

L'algorithme k-means présente plusieurs limitations. Par exemple, celles-ci peuvent être problématique lorsque l'on ne dispose que d'une matrice de similiarité ou de distance entre les observations et que l'on n'a pas accès aux données originales. Dans un tel contexte, les méthodes de classification hiérarchique sont pertinentes.

La classification hiérarchique permet d'obtenir une série de partitions imbriquées, allant de la partition la plus fine (chaque observation dans son propre groupe) à la plus grossière (toutes les observations dans un seul groupe). Cette approche ne fournit donc pas une seule parition, mais une hiérarchie de partitions. Cette hiérarchie peut être représenté à l'aide d'un **dendogramme**, un arbre résumant comment ces partitions sont imbriquées. Il existe deux types d'algorithmes pour effectuer une classification hiérarchique :

1. les algorithmes ascendants, qui partent des observations individuelles et procèdent par fusions successives ;
2. les algorithmes descendants, qui partent d'un groupe contenant toutes les observations et procèdent par divisions successives.

Dans les deux cas, on obtient n partitions hiérarchiques constituées de 1 à n groupes.

1 Algorithmes

1.1 Algorithmes descendants

Les algorithmes descendants commencent avec l'ensemble des n observations réunis dans un seul groupe. À chaque étape, le groupe jugé le moins homogéne est divisé en deux sous-groupes, en cherchant à maximiser la dissimilarité entre les deux sous-groupes. On continue ainsi jusqu'à ce que chaque observation soit isolé dans son propre groupe.

Ce type d'algorithme est coûteux en temps de calcul, car il faut évaluer, à chaque étape, toutes les manières possibles de diviser un groupe en deux. On l'utilise donc rarement en pratique.

1.2 Algorithme ascendant

À l'inverse, les algorithmes ascendants débutent avec n groupes distincts, chacun contenant une seule observation. À chaque étape, on fusionne les deux groupes les plus similaires, i.e. ceux dont la dissimilarité est la plus faible selon un critère choisi. L'algorithme continue jusqu'à ce qu'il ne reste plus qu'un seul groupe contenant toutes les observations.

2 Distance entre groupes

Pour mettre en oeuvre les algorithmes précédent, on doit définir la distance entre deux groupes d'observations A et B, notée $d(A, B)$. Si l'on sait généralement mesurer distance entre deux individus, on doit définir une distance entre deux groupes contenant un nombre différents d'éléments. Il existe plusieurs façon de calculer une telle distance entre deux groupes.

2.1 Méthode du plus proche voisin (*single linkage*)

Dans cette approche, la distance entre deux groupes est définie comme la plus petite distance entre un individu de A et un individu de B :

$$d(A, B) = \min\{d_{ij} : i \in A, j \in B\}.$$

Dit autrement, deux groupes A et B sont considérés comme proche si un élément de A est proche d'un élément de B. Cette méthode présente plusieurs avantages. Elle donne de bons résultats lorsque les variables sont de nature différente (e.g., quantitatives et qualitatives) et permet de construire des groupes aux formes irrégulière. De plus, elle est relativement robuste aux données aberrantes. Enfin, ses propriétés mathématiques théoriques sont intéressantes.

Cependant, cette méthode tend à créer des groupes déséquilibrés : un grand groupe central entouré de plusieurs petits groupes satellites. Elle est moins performante lorsque les groupes naturels sont de forme régulière. Bien qu'elle est de bonnes propriétés mathématiques, celles-ci ne se vérifie pas toujours empiriquement.

2.2 Méthode du voisin le plus distant (*complete linkage*)

À l'inverse de la méthode précédente, la distance entre deux groupes est définie comme la plus grande distance entre un individu de A et un individu de B :

$$d(A, B) = \max\{d_{ij} : i \in A, j \in B\}.$$

Dit autrement, deux groupes sont considérés proches si tous les éléments de A sont proches de tous les éléments de B. Cette méthode a tendence à produire des groupes réguliers de taille homogène. Comme la méthode du plus proche voisin, elle est bien adaptée aux variables de différents types. Cependant, elle est **extrêmement** sensible aux données aberrantes. En effet, un seul individu peut augmenter artificiellement la distance entre deux groupes. De plus, elle a tendance à forcer la formation de groupes de même taille, ce qui n'est pas toujours justifié en pratique.

2.3 Méthode de la moyenne (*average linkage*)

Ici, la distance entre deux groupes est définie comme la moyenne des distances entre toutes les paires d'individus, issu de A et de B :

$$d(A,B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d(X_i, X_j).$$

où n_A est le nombre d'observations dans le groupe A et n_B est le nombre d'observations dans le groupe B.

Cette méthode consiste à considérer toutes les interactions possibles entre les éléments des deux groupes, puis à en faire la moyenne. Elle tend à produire des groupes dont la variance interne est faible, i.e. relativement homogène. Toutefois, cette méthode privilégie la formation de groupes de variance similaire, ce qui n'est pas toujours justifié en pratique.

2.4 Méthode du centroïde (*centroid method*)

Pour cette méthode, la distance entre deux groupes est définie comme la distance entre leurs centroïdes, i.e. les moyennes des observations de chaque groupe :

$$d(A,B) = d(\overline{X}_A, \overline{X}_B).$$

où

$$\overline{X}_A = \frac{1}{n_A} \sum_{i \in A} X_i, \quad \text{et} \quad \overline{X}_B = \frac{1}{n_B} \sum_{j \in B} X_j$$

Après la fusion de A et de B, le nouveau centroïde $\overline{X}_{AB}$ est donné par la moyenne pondérée :

$$\overline{x}_{AB} = \frac{n_A \overline{x}_A + n_B \overline{x}_B}{n_A + n_B}.$$

Cette approche est assez robuste aux données aberrantes, mais elle est généralement peu performante lorsqu'il n'y en a pas.

2.5 Méthode de la médiane (*median method*)

La méthode de la médiane repose sur une mise à jour des distances de façon récursive. Lorsqu'on fusionne deux groupes A et B, on définit la distance entre le nouveau groupe AB et autre groupe C par la formule :

$$d(AB, C) = \frac{d(A, C) + d(B, C)}{2} - \frac{d(A, B)}{4}.$$

Cette méthode est particulièrement robuste aux données aberrantes (davantage que la méthode du centroïde). Elle est cependant très peu efficace lorsque de telles valeurs extrêmes sont absentes.

2.6 Méthode de Ward (*Ward's method*)

La méthode de Ward est une variante de la méthode du centroïde. Elle est optimale dans le cas où les observations suivent des lois normales multivariées, de même matrice de variance-covariance mais de moyennes différentes. Elle est basée sur une mesure de l'inertie intra-groupe. Pour chaque groupe A, groupe B et groupe $A \cup B$, noté AB, on définit

$$SC_A = \sum_{i \in A} (X_i - \overline{X}_A)^\top (X_i - \overline{X}_A),$$

$$SC_B = \sum_{j \in B} (X_j - \overline{X}_B)^\top (X_j - \overline{X}_B),$$

$$SC_AB = \sum_{k \in A \cup B} (X_k - \overline{X}_{AB})^\top (X_k - \overline{X}_{AB}).$$

où $\overline{X}_A$, $\overline{X}_B$ et $\overline{X}_{AB}$ sont calculées comme dans la méthode du centroïde. On regroupe les groupes A et B qui minimisent l'augmentation de l'inertie :

$$I_{AB} = SC_{AB} - SC_A - SC_B = \frac{d^2(\overline{X}_A, \overline{X}_B)}{\frac{1}{n_A} + \frac{1}{n_B}}.$$

Cette méthode est très efficace lorsque les groupes sont homogènes, de taille comparable et que les hypothèses gaussiennes sont raisonnablement satisfaites. En revanche, elle est sensible aux données aberrantes et tend à former des regroupements de même taille.

2.7 Méthode flexible (*flexible clustering*)

La méthode flexible repose sur une formule générale permettant de représenter plusieurs méthodes de mise à jour des distances. Si l'on fusionne deux groupes A et B pour former AB, et que l'on souhaite calculer la distance entre AB et autre groupe C, on peut utiliser la relation

$$d(C, AB) = \alpha_A d(C, A) + \alpha_B d(C, B) + \beta d(A, B) + \gamma \left| d(C, A) - d(C, B) \right|.$$

Selon les valeurs choisies pour les coefficients α_A, α_B, β et γ, on peut retrouver les formules de mise à jour correspondant aux différentes méthodes précédentes. Le Table 1 présente les valeurs des différents coefficients à choisir pour retrouver les différentes méthodes.

TABLE 1 – Coefficients pour retrouver les méthodes précédentes.

Méthode	α_A	α_B	β	γ
Plus proche	$1/2$	$1/2$	0	$-1/2$
Plus distant	$1/2$	$1/2$	0	$1/2$
Médiane	$1/2$	$1/2$	$-1/4$	0
Moyenne	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	0	0
Centroïde	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	$-\frac{n_A n_B}{n_A+n_B}$	0
Ward	$\frac{n_A+n_C}{n_A+n_B+n_C}$	$\frac{n_B+n_C}{n_A+n_B+n_C}$	$-\frac{n_C}{n_A+n_B+n_C}$	0

Pour la méthode flexible, on impose arbitrairement les contraintes suivantes :

$$\alpha_A + \alpha_B + \beta = 1, \quad \alpha_A = \alpha_B, \quad \gamma = 0.$$

Ainsi, on a $\alpha_A = \alpha_B = \frac{1-\beta}{2}$. Il ne reste qu'un seul paramètre à fixer. Généralement, on choisit $\beta = -0.25$. Si l'on soupçonne la présence de données aberrantes, on peut opter pour $\beta = -0.5$ afin d'accroître la robustesse de l'algortihme.

3 Quelle partition choisir ?

Un algorithme de classification hierarchique, ascendant ou descendant, produit une séquence de n partitions imbriquées, allant de n groupes (où chaque observation est isolée) à un seul groupe contenant toutes les observations. En pratique, on se pose donc la question suivante : quelle partition faut-il considérer ?

Plusieurs pistes peuvent guider ce choix :

— Une des partitions est-elle particulièrement interprétable d'un point de vue scientifique ou métier ?
— Une des partitions est-elle pertinente sur le plan opérationnel, par exemple pour orienter des décisions ?
— Cherche-t-on explicitement à segmenter la population en un nombre de groupes K, déterminé à l'avance ?

Lorsque ces considérations pratiques ne suffisent pas à trancher, on peut s'appuyer sur des critères statistiques pour évaluer la qualité des partitions et suggérer un nombre "optimal" de groupes. Ces critères sont surtout adaptés lorsque les variables sont continues.

La librairie R `NbClust` en propose une trentaine et la librairie Python `sklearn` une dizaine. On présente ensuite les critères les plus fréquemment utilisés.

3.1 Critères basés sur l'inertie

L'inertie mesure la dispersion des observations appartenant à un groupe par rapport à son centroïde. L'inertie totale I_{tot} de l'ensemble des n observations est

$$I_{\text{tot}} = \frac{1}{n} \sum_{i=1}^{n} d(X_i, G),$$

où la fonction d est une distance et G le centre de gravité de l'ensemble des observations. L'inertie totale I_{tot} se décompose en une inertie **intra-groupe** et une inertie **inter-groupe** de la façon suivante :

$$I_{\text{tot}} = I_{\text{intra-groupe}} + I_{\text{inter-groupe}}.$$

L'inertie inter-groupe $I_{\text{inter-groupe}}$ est l'inertie des centres de gravité des groupes pondérés par le nombre d'observations dans le groupe. Elle mesure la séparation entre les groupes. Elle est donné par

$$I_{\text{inter-groupe}} = \frac{1}{n} \sum_{k=1}^{K} n_k d(G_k, G),$$

où G_k est le centre de gravité du groupe k et n_k le nombre d'observations dans ce groupe. L'inertie intra-groupe $I_{\text{intra-groupe}}$ est la somme des inerties des groupes. Elle mesure donc l'hétérogénéité des groupes. Elle est donnée par

$$I_{\text{intra-groupe}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} d(X_i, G_k),$$

où l'ensemble C_k contient les observations du groupe k.

Plus les groupes sont compacts, plus l'inertie intra-groupe est faible, ce qui signifie que l'inertie inter-groupe est élevé. Une partition de bonne qualité maximise donc l'inertie inter-groupe. Ces critères supposent généralement que les variables sont continues et ont été standardisées pour

éviter qu'une variable à grande échelle ne domine les autres. Parmi les critères se basant sur l'inertie, on peut citer : le pseudo-R^2 et la statistique de Calinski-Harabasz (CH).

Le pseudo-R^2 mesure la proportion d'inertie expliquée par la partition :

$$\text{pseudo-}R^2 = \frac{I_{\text{inter-groupe}}}{I_{\text{tot}}}.$$

Un pseudo-R^2 élevé indique que la parition capture une grande part de la structure des données. La statistique CH est une variation du pseudo-R^2 normalisé par le nombre de groupes :

$$\text{CH} = \frac{I_{\text{inter-groupe}}/(K-1)}{I_{\text{intra-groupe}}/(n-K)}.$$

Un score élevé de la statistique CH suggère un bon équilibre entre groupes compactes et groupes séparés.

3.2 Critères basés sur la distance

On peut aussi utiliser des critères se basant sur la distance entre les observations pour mesurer la qualité d'une partition. Parmi les critères se basant sur la distance, on peut citer : l'indice de Dunn et l'indice de silhouette.

L'indice de Dunn cherche à maximiser la distance minimale entre deux groupes, tout en minimisant la distance maximale entre les observations à l'intérieur d'un groupe :

$$D = \frac{\text{Distance minimale entre 2 groupes}}{\text{Distance maximale dans un groupe}}.$$

Ce critère favorise des groupes denses et bien séparés. En ce qui concerne le critère de silhouette, il mesure la qualité d'affection d'une observation X_i à son groupe :

$$S(X_i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

où a_i est la distance moyenne entre l'observation X_i et les autres observations de son groupe et b_i est la distance moyenne entre l'observation X_i et les observations du groupe le plus proche de X_i. On souhaite maximiser la silhouette moyenne des observations. La silhouette moyenne sur toutes les observations est donc un bon indicateur de la cohérence globale de la partition.

En résumé, plusieurs critères peuvent nous guider dans le choix du nombre de groupes, mais aucun n'est parfait. Il est souvent recommendé de croiser connaissance métier, visualisation des partitions et indicateurs statistiques pour prendre une décision éclairée.