Hierarchique

Certains problèmes de l'algorithme k-means peuvent être résolu avec des algorithmes de classification hiérarchique. Par exemple, lorque l'on a à dispostion qu'une matrice de similiarité/distance entre les observations.

La classification hiérarchique permet d'obtenir des partitions toutes imbriquées les unes dans les autres. Il existe deux types d'algortihmes pour effectuer de la classification hiérarchique :

1. les algorithmes ascendants ;
2. les algortihmes descendants.

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

Algorithme descendant

Un algorithme descendant fonctionne ainsi :

— Au départ, toutes les observations sont dans un seul et même groupe de n observations ;
— À chaque étape, on divise le groupe le moins homogéne en deux groupes.
— À la fin, après n étapes, chaque observation à son propre groupe, c'est-à-dire qu'on obtient n groupes contenant une seule observation.

L'exécution de cet algorithme ne donne pas une seule parition, mais n partitions, que l'on peut résumer à l'aide d'un graphique en forme d'arbre appelé dendogramme. Certains critères peuvent aider à choisir l'une parmi les n partitions proposées par l'algorithme. Ils demandent beaucoup de temps de calcul.

Algorithme ascendant

Un algorithme ascendant fonctionne ainsi :

— Au départ, chaque observation est son propre groupe, c'est-à-dire uq'on démarre avec n groupes content chacun une seule observation.
— À chaque étape, on fusionne les deux groupes les plus similaires.
— À la fin, après n étapes, on obtient un seul groupe contenant toutes les n observations.

Différences entre les algorithmes :

— caractère ascendant ou descendant ;
— leur façon de mesurer les distances / similiarité entre deux observations ;
— leur façon de mesurer les distances / similiarité entre deux groupes.

Distance entre groupes

Pour mettre en oeuvre les algorithmes précédent, on doit définir la distance entre deux groupes d'observations A et B, $d(A, B)$.

Example : On sait comment mesurer la distance entre les trois paires possibles de $\{1\}$, $\{2\}$ et $\{3\}$, mais comment mesurer la distance entre $\{1, 2\}$ et $\{3\}$?

Il existe plusieurs façons de calculer une telle distance entre deux groupes.

— Méthode du plus proche voisin (single linkage)

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

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

La distance/similiarité entre deux groupes d'observations est tout simplement la distance/similiarité entre les points de chaque groupe qui sont les plus rapprochés/similaires

Avantages : * Donne de bons résultats lorsque les variables sont de types différents * Possède d'excellentes propriétés théoriques * Permet de créer des groupes dont la forme est très irrégulière * Est robuste aux données aberrantes

Désavantages : * Tend à former un grand groupe avec plusieurs petits groupes satellites * Perd de l'efficacité si les vrais groupes sont de forme réegulière * Possède des propriétés théoriques ne semblant pas se réaliser en pratique dans certaines études

— Méthode du voisin le plus distant (complete linkage)

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

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

La distance/similiarité entre deux groupes d'observations est tout simplement la distance/similiarité entre les points de chaque groupe qui sont les plus éloignés/dissimilaires.

Avantages : * Donne de bons résultats lorsque les variables sont de types différents * Tend à former des groupes de taille égale

Désavantages : * Est extrêmement sensible aux données aberrantes * Tend à former des groupes de taille égale * Est très peu utilisée en pratique

— Méthode de la moyenne (average linkage)

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

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

On doit calculer les $n_A \times n_B$ distances/similiarités possibles entre les points des deux groupes, ensuite on prend la moyenne de ces distances/similiarités comme étant celle qui sépare les groupes.

Avantages : * Tend à former des groupes de faible variance

Désavantages : * Tend à former des groupes de même variance

— Méthode du centroïde (centroid method)

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

où

$$\overline{x}_A = \frac{1}{n_A} \sum_{i \in A} x_i, \quad \text{et} \quad \overline{x}_B = \frac{1}{n_B} \sum_{j \in B} x_j$$

La moyenne $\overline{x}_{AB}$ du nouveau groupe résultant de la fusion des groupes A et B se calcule comme suit :

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

Avantages : * Est très robuste aux données aberrantes

Désavantages : * Est peu efficace en l'absence de données aberrantes

— Méthode de la médiane (median method)

À une étape donnée, nous avons toujours à notre disposition la distance entre les groupes déjà formés. On fusionne les deux groupes les moins distants/les plus similaires, disons A et B pour obtenir un groupe AB. On met à jour la matrice des distances : la distance entre le nouveau groupe AB et tout autre groupe C est donnée par

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

Avantages : * Est encore plus robuste en présence de données aberrantes

Désavantages : * Est très peu efficace en l'absence de données aberrante

— Méthode de Ward (Ward's method)

Variante de la méthode du centroïde pour tenir compte de la taille des groupes. Elle a été conçue pour être optimale si les n vecteurs $x_1, \ldots, x_n$ suivent des lois normales multivariées de K moyennes différentes mais toutes de même matrice de variance-covariance.

Basée sur les sommes des carrées

$$SC_A = \sum_{i \in A} (x_i - \overline{x}_A)^\top (x_i - \overline{x}_A).$$

$$SC_B = \sum_{j \in B} (x_j - \overline{x}_B)^\top (x_j - \overline{x}_B).$$

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

où $\overline{x}_A$, $\overline{x}_B$ et $\overline{x}_{AB}$ sont calculées comme dans la méthode du centroïde. On regroupe les classes A et B pour lesquelles

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

est minimale.

Avantages : * Est optimale si les observations sont approximativement distribuées selon une loi normale multidimensionnelle de même matrice de variances-covariances

Désavantages : * Est sensible aux données aberrantes * Tend à former des groupes de petite taille * Tend à former des groupes de même taille

— Méthode flexible

Les auteurs de cette méthode ont remarqué que pour plusieurs méthodes connues, on a les relations suivantes :

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

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

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

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

Ainsi,

$$\alpha_A = \alpha_B = \frac{1-\beta}{2}.$$

Et il ne reste qu'à choisir β. Les autheurs suggèrent de poser $\beta = -0.25$ sauf quand on soupçonne la présence de données aberrantes où l'on suggère $\beta = -0.5$.

Pratique

L'exécution d'un algorithme nous donne une séquence de n partitions ayant de n à 1 groupes.

Quelle partition de cette séquence devrions-nous choisir ?

L'une des n partitions est-elle particulièrement interprétable ? L'une des n partitions a-t-elle a un sens pratique ? Visions-nous séparer la population en un nombre K de groupes ?

S'il n'y a pas de réponse claire à ces questions, des critères peuvent nous guider …

Il y a plusieurs indications pour nous aider dans le choix du nombre de classe (surtout si les variables sont continues). La librairie **NbClust** en contient une trentaine : https ://www.rdocumentation.org/packages/NbClust/versions/3.0.1/topics/NbClust

— Les indicateurs basées sur l'inertie

$$I_{tot} = I_{intra-groupe} + I_{inter-groupe}$$

Ces indicateurs sont plus pertinents avec des variables continues. Prendre garde au poids des variables et à la standardisation.

— Pseudo- R^2

$$Pseudo - R^2 = \frac{I_{inter-groupe}}{I_{tot}}$$

— Statistique de Caliliski-Harabasz (CH) :

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