

# Hierarchique

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

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

1. les algorithmes ascendants ;
2. les algorithmes 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 homogène en deux groupes.
- À la fin, après  $n$  étapes, chaque observation a 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 partition, mais  $n$  partitions, que l'on peut résumer à l'aide d'un graphique en forme d'arbre appelé dendrogramme. 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 qu'on démarre avec  $n$  groupes contenant 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 / similarité entre deux observations ;
- leur façon de mesurer les distances / similarité entre deux groupes.

## Distance entre groupes

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

Exemple : 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/similarité entre deux groupes d'observations est tout simplement la distance/similarité 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 régulière \* 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 nombre d'observations dans le groupe  $B$ .

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)

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

où

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

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

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

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  $x_1, \dots, x_n$  suivent des lois normales multivariées de  $K$  moyennes différentes mais toutes de même matrice de variance-covariance.

Basée sur les sommes des carrées

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

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

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

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

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

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 :

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

Méthode	$\alpha_A$	$\alpha_B$	$\beta$	$\gamma$
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  $\beta$ . Les auteurs suggèrent de poser  $\beta = -0.25$  sauf quand on soupçonne la présence de données aberrantes où l'on suggère  $\beta = -0.5$ .