Introduction au Machine Learning

C6 - Partitionnement hiérarchique

LAMSADE - Université Paris-Dauphine
lucas.gnecco-heredia@dauphine.psl.eu - C602

10 novembre 2022

* *Remerciement spécial à Florian Yger*

Partitionnement et hiérarchie

Partitionnement

Jusqu'à présent, nous avons uniquement considéré des partitions disjointes.
Cependant, un cluster peut lui-même être constitué de sous-clusters ou lui-même faire partie d'un cluster plus grand.

- Notion de hiérarchie dans les clusters
- rejoint la façon dont les biologistes classent le vivant (taxonomie)

Philosophie de cette approche

Principe

On va appliquer un principe similaire et regrouper les points similaires en clusters et les clusters similaires en sur-clusters pour aboutir à une hiérarchie de clusters.

Questions

- comment définir un cluster pour qu'il admette une hiérarchie ?
- comment définir une notion de similarité entre clusters ?

Une première définition

Définition - niveau

On considère une suite de partitions de n observations dans c clusters.
La première de ces partitions est une partition en n clusters (où chaque cluster ne contient qu'une seule observation). La partition suivante est constituée de $(n-1)$ clusters et ainsi de suite, jusqu'à l'étape n où l'on a un unique cluster constitué de n observations.

Commentaires

- On est au niveau k de cette suite de partitions quand $c = n - k + 1$
- On peut vérifier que :
  - niveau $1 \rightarrow n$ clusters
  - niveau $n \rightarrow 1$ cluster

Une première définition

Définition - niveau

On considère une suite de partitions de n observations dans c clusters.
La première de ces partitions est une partition en n clusters (où chaque cluster ne contient qu'une seule observation). La partition suivante est constituée de $(n-1)$ clusters et ainsi de suite, jusqu'à l'étape n où l'on a un unique cluster constitué de n observations.

Définition - hiérarchie

Étant donné deux observations x et x', pour lesquelles il existe un niveau où x et x' appartiennent au même cluster. Si la suite de partitions est telle que toute paire d'observations dans le même cluster à un niveau k reste ensemble à tous les niveaux suivants, alors on a un clustering hiérarchique.

Représentations

Il existe plusieurs représentations (graphiques) possibles d'une suite de partitions :

- sous forme d'arbre
- diagramme de Venn
- en notations ensemblistes

Exemples

Approches possibles

Pour trouver une telle hiérarchie de partitionnements, deux approches sont possibles :

- la Classification Hiérarchique Ascendante (CHA)
  Une approche "agglomérante" qui commence au niveau 1 et fusionne des clusters similaires.
- la Classification Hiérarchique Descendante
  Une approche "divise" qui commence au niveau n et sépare les clusters les moins homogènes.

Approches possibles

Pour trouver une telle hiérarchie de partitionnements, deux approches sont possibles :

- la Classification Hiérarchique Ascendante (CHA)
  Une approche "agglomérante" qui commence au niveau 1 et fusionne des clusters similaires.
- la Classification Hiérarchique Descendante
  Une approche "divise" qui commence au niveau n et sépare les clusters les moins homogènes.

Remarque

En pratique, on utilise un nombre restreint de clusters et on privilégie la première approche.

Algorithme

Pour générer c clusters :

Algorithm 1 CHA

Require: c

$\hat{c} \leftarrow n$

$\forall i = 1, \cdots, n \; \mathcal{D}_i = \{x_i\}$

while $\hat{c} \geq c$ **do**

$\hat{c} \leftarrow \hat{c} - 1$

$(i, j) = \arg\min_{k,l} d(\mathcal{D}_k, \mathcal{D}_l)$ (paire de clusters la plus similaire)

fusion de D_i et D_j

end while

Algorithme

Pour générer c clusters :

Algorithm 2 CHA

Require: c

$\hat{c} \leftarrow n$

$\forall i = 1, \cdots, n \; \mathcal{D}_i = \{x_i\}$

while $\hat{c} \geq c$ **do**

$\hat{c} \leftarrow \hat{c} - 1$

$(i, j) = \arg\min_{k,l} d(\mathcal{D}_k, \mathcal{D}_l)$ (paire de clusters la plus similaire)

fusion de D_i et D_j

end while

Question

Comment définir la similarité/distance entre de deux clusters ? ?

Similarités

Pour trouver des clusters similaires, on va d'abord définir une distance entre clusters que l'on cherchera à minimiser. Le point-clé est donc de définir une distance entre cluster/ensemble d'observations.

Distance entre clusters

- plus proche voisin : $d_{\text{min}}(\mathcal{D}_i, \mathcal{D}_j) = \min_{\substack{x \in \mathcal{D}_i \\ x' \in \mathcal{D}_j}} ||x - x'||$
- diamètre maximum : $d_{\text{max}}(\mathcal{D}_i, \mathcal{D}_j) = \max_{\substack{x \in \mathcal{D}_i \\ x' \in \mathcal{D}_j}} ||x - x'||$
- distance moyenne : $d_{\text{moy}}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{n_i n_j} \sum_{x \in \mathcal{D}_i} \sum_{x' \in \mathcal{D}_j} ||x - x'||$
- distance de Ward : $d_{\text{ward}}(\mathcal{D}_i, \mathcal{D}_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} ||\bar{\mu}_i - \bar{\mu}_j||$

Similarités

Distance entre clusters

- plus proche voisin : $d_{\min}(\mathcal{D}_i, \mathcal{D}_j) = \min_{\substack{x \in \mathcal{D}_i \\ x' \in \mathcal{D}_j}} ||x - x'||$
- diamètre maximum : $d_{\max}(\mathcal{D}_i, \mathcal{D}_j) = \max_{\substack{x \in \mathcal{D}_i \\ x' \in \mathcal{D}_j}} ||x - x'||$
- distance moyenne : $d_{\text{moy}}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{n_i n_j} \sum_{x \in \mathcal{D}_i} \sum_{x' \in \mathcal{D}_j} ||x - x'||$
- distance de Ward : $d_{\text{ward}}(\mathcal{D}_i, \mathcal{D}_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} ||\bar{\mu}_i - \bar{\mu}_j||$

Remarque importante

Certaines de ces distances sont définies comme le résultat de problème d'optimisation (max ou min) mais dans tous les cas, on cherchera la paire $(\mathcal{D}_i, \mathcal{D}_j)$ qui minimise la distance choisie.

Illustrations

Remarques

Stratégie de regroupement

- saut minimal (*single linkage*) basé sur d_{min}
  - tendance à produire des clusters fins et allongés (par effet de chaînage) ou très généraux/vastes
  - lien avec l'arbre couvrant de poids minimal (Minimum Spanning Tree - Kruskal algo)
  - sensibilité au bruit, aux observations aberrantes
- saut maximal (*complete linkage*) basé sur d_{max}
  - tendance à produire des clusters compacts et de taille similaires, c'est-à-dire assez spécifiques
  - sensibilité au bruit, aux observations aberrantes
- saut moyen basé sur d_{moy}
  - tendance à produire des clusters de variance proche
- barycentre basé sur d_{ward}
  - bonne robustesse au bruit

Choix du nombre de clusters

- on n'a pas besoin de spécifier d'avance un nombre de cluster a priori
- on pourra le décider à posteriori :
  - en coupant à un niveau défini
  - en s'arrêtant lorsqu'on a le plus grand saut d'agrégation entre deux niveaux

Choix du nombre de clusters

- on n'a pas besoin de spécifier d'avance un nombre de cluster a priori
- on pourra le décider à posteriori :
  - en coupant à un niveau défini
  - en s'arrêtant lorsqu'on a le plus grand saut d'agrégation entre deux niveaux

illustrations/explications

Remarques générales

CHA

Le clustering Hiérarchique Ascendant est une méthode déterministe (pas de sensibilité à l'initialisation) mais le partitionnement produit dépend du choix de distance entre clusters.

Autre choix

D'autres méthodes de clustering existent à base de graphes (partitionnement spectral (*Spectral clustering*) ou de densité (DBSCAN)

Inertie(s)

- inertie intra-cluster
  $J_w = \sum_k \sum_{i \in C_k} d^2(x_i, \mu_k)$ inertie du cluster k

- inertie inter-cluster
  $J_b = \sum_k n_k d^2(\mu_k, \mu)$

Inertie(s)

- inertie intra-cluster
  $J_w = \sum_k \sum_{i \in C_k} d^2(x_i, \mu_k)$ inertie du cluster k
  $\sum_{i \in C_k} d^2(x_i, \mu_k)$: inertie de $C_k \leftrightarrow$ concentration des observations autour de μ_k
- inertie inter-cluster
  $J_b = \sum_k n_k d^2(\mu_k, \mu)$

Inertie(s)

- inertie intra-cluster
  $J_w = \sum_k \sum_{i \in C_k} d^2(x_i, \mu_k)$ inertie du cluster k
  $\sum_{i \in C_k} d^2(x_i, \mu_k)$: inertie de $C_k \leftrightarrow$ concentration des observations autour de μ_k
- inertie inter-cluster
  $J_b = \sum_k n_k d^2(\mu_k, \mu)$
  $d^2(\mu_k, \mu) \leftrightarrow$ éloignenemt des centres des clusters avec le centre du nuage de points

Critère

Dans l'idéal

On cherche à minimiser l'inertie intra cluster J_w et à maximiser l'inertie inter-cluster J_b

- on préfère des clusters denses/homogènes et bien séparés

Illustrations

TD

On considère un ensemble $E = \{w_1, \cdots, w_7\}$ de 7 observations pour lesquelles on a calculé toutes les distances paire à paire :

d	w_1	w_2	w_3	w_4	w_5	w_6	w_7
w_1	0	2	4.5	5.5	7.5	9.5	4
w_2	2	0	2.5	3.5	5.5	7.5	4
w_3	4.5	2.5	0	3	5	7	6.5
w_4	5.5	3.5	3	0	2	4	7.5
w_5	7.5	5.5	5	2	0	4	9.5
w_6	9.5	7.5	7	4	4	0	5.5
w_7	4	4	6.5	7.5	9.5	5.5	0

TD

d	w_1	w_2	w_3	w_4	w_5	w_6	w_7
w_1	0	2	4.5	5.5	7.5	9.5	4
w_2	2	0	2.5	3.5	5.5	7.5	4
w_3	4.5	2.5	0	3	5	7	6.5
w_4	5.5	3.5	3	0	2	4	7.5
w_5	7.5	5.5	5	2	0	4	9.5
w_6	9.5	7.5	7	4	4	0	5.5
w_7	4	4	6.5	7.5	9.5	5.5	0

- appliquer une CHA en utilisant l'agrégation de saut minimal
- appliquer une CHA en utilisant l'agrégation de saut maximal
- pour chaque cas, représenter le dendogramme associé et proposer un nombre de partitions.