DIFIQ
Niveau 2
Statistiques
K-means
Articulations entre CAH et K-means

Jean-François Berger-Lefébure

21 Novembre 2024

Contents

1 K-means : Généralités et principe

1.1 Généralités

- Le principe des K-means est de partitionner les données en K classes (ou clusters).
- Le nombre de clusters K doit être défini à l'avance.
- L'algorithme fonctionne de manière itérative pour ajuster les clusters et minimiser l'inertie intra-classe.
- Il existe plusieurs variantes de cet algorithme pour s'adapter à différents cas.
- La qualité des résultats dépend fortement du choix initial des centres et du nombre de clusters K.

1.2 Critère à minimiser

L'objectif de K-means est de minimiser l'inertie intra-classe, définie comme suit :

$$\sum_{k=1}^{K} \sum_{i \in C_k} \omega_i d^2(i, g_k)$$

où :

- K : nombre de clusters.
- C_k : cluster k.
- ω_i : poids de l'individu i.
- g_k : centre de gravité du cluster C_k.
- $d^2(i, g_k)$: distance quadratique entre i et le centre g_k.

1.3 Caractéristiques des clusters

- **Poids d'un cluster** C_k :
$$\mu_k = \sum_{i \in C_k} \omega_i$$
- **Centre de gravité d'un cluster** C_k :
$$g_k = \frac{1}{\mu_k} \sum_{i \in C_k} \omega_i x_i$$
- Il n'existe pas de solution analytique directe pour minimiser ce critère. L'algorithme procède donc de manière itérative.

1.4 Résumé des étapes du K-means

1. **Choisir** K :
   - Fixer le nombre de clusters K à l'avance.
   - Choisir aléatoirement les centres initiaux ou selon une méthode spécifique.
2. **Étape d'affectation** :
   - Assigner chaque individu au cluster dont le centre est le plus proche (selon la distance choisie).
3. **Étape de mise à jour** :
   - Recalculer les centres de chaque cluster comme les centres de gravité des individus assignés.
4. **Répétition** :
   - Répéter les étapes d'affectation et de mise à jour jusqu'à convergence.

Conclusion

K-means est une méthode simple mais puissante pour partitionner les données. Cependant, elle nécessite de fixer le nombre de clusters K à l'avance, et sa performance peut être influencée par l'initialisation des centres.

Algorithme des K-means (Lloyd/Forgy)

Principe général

L'objectif de l'algorithme K-means est de partitionner n individus en K clusters en répétant deux étapes clés jusqu'à convergence :

1. Affecter chaque individu au cluster correspondant au centre de classe le plus proche.
2. Recalculer les centres de classes en fonction des clusters formés (calcul des centroïdes).

Étapes détaillées de l'algorithme

1. **Initialisation des centres de classe** :
   - Tirer K centres de classe aléatoirement dans l'espace des individus.
   - Ces K centres représentent les clusters initiaux.
2. **Itérations successives** :
   a. **Affectation des points** :
      - Chaque individu est affecté au cluster correspondant au centre de classe le plus proche (selon une distance, généralement euclidienne).
   b. **Recalcul des centres de classe** :
      - Les nouveaux centres de classe sont calculés comme les **centroïdes** des clusters formés, c'est-à-dire les centres de gravité.
3. **Critères de convergence** :
   - Répéter les étapes 2.a et 2.b jusqu'à convergence.
   - Les critères d'arrêt incluent :
     – Un nombre maximal d'itérations.
     – Une stabilisation des centres (les centres ne bougent plus ou très peu).

Particularités de l'algorithme

- **Initialisation** :
  – L'efficacité de l'algorithme dépend souvent des centres initiaux choisis aléatoirement.
  – Une mauvaise initialisation peut entraîner un découpage sous-optimal.
- **Convergence** :
  – La convergence n'est pas toujours garantie. Des ajustements mineurs peuvent persister, d'où l'importance des critères d'arrêt.
- **Nom** :
  – L'algorithme est connu sous le nom de **Lloyd/Forgy**.
  – Il est couramment appelé **K-means**, bien que ce ne soit pas son appellation d'origine.

Avantages et inconvénients

Avantages :

- Facile à comprendre et à implémenter.
- Convergence rapide pour des jeux de données de taille moyenne.

Inconvénients :

- Nécessite de fixer K à l'avance.
- Sensible à l'initialisation des centres.
- Peut converger vers des solutions sous-optimales.

2 Lecture et interprétation de la formule de l'inertie intra-classe

La formule utilisée dans l'algorithme K-means est la suivante :

$$\sum_{k=1}^{K} \sum_{i \in C_k} \omega_i d^2(i, g_k)$$

Lecture de la formule

De gauche à droite :

- $\sum_{k=1}^{K}$: On effectue une somme sur tous les clusters k, de $k = 1$ à K.
- $\sum_{i \in C_k}$: Pour chaque cluster C_k, on effectue une somme sur tous les individus i appartenant à ce cluster.
- ω_i : C'est le poids de l'individu i, qui peut refléter son importance. Si tous les individus sont également importants, $\omega_i = 1$.
- $d^2(i, g_k)$: La distance au carré entre l'individu i et le centre de gravité g_k du cluster C_k.

Interprétation

Cette formule mesure la **compacité globale des clusters**. Voici les étapes pour l'interpréter :

1. Pour chaque cluster C_k, on calcule la somme des distances quadratiques ($d^2(i, g_k)$) entre les individus du cluster et le centre g_k.
2. On pondère chaque distance par le poids ω_i.
3. On répète cette opération pour tous les clusters C_k, puis on additionne les résultats.

L'objectif de l'algorithme K-means est de **minimiser cette valeur**, afin de former des clusters compacts et homogènes.

But de la formule

La formule donne une mesure globale de la **compacité intra-classe**. Plus cette valeur est faible, plus les points sont proches de leurs centroïdes, et plus les clusters sont homogènes.

Phrase explicative de gauche à droite

La formule se lit comme suit : *Pour chaque cluster C_k, on calcule la somme des distances quadratiques entre les individus i du cluster et leur centre g_k, pondérées par ω_i, et on additionne ces valeurs pour tous les clusters.*

3 Comprendre les itérations de l'algorithme K-means

L'algorithme K-means fonctionne en itérations, alternant entre deux étapes principales jusqu'à convergence :

1. **Étape d'affectation** : Chaque point est affecté au cluster dont le centre est le plus proche.
2. **Étape de mise à jour** : Les centres des clusters (centroïdes) sont recalculés comme les centres de gravité des points qui leur sont associés.

Exemple concret

Nous avons les points suivants dans un plan 2D :

$$\{(1,1),(2,1),(4,3),(5,4)\}$$

et nous souhaitons les regrouper en $K = 2$ clusters.

Étape 1 : Initialisation

On choisit au hasard deux centres initiaux :

$$g_1 = (1,1) \quad \text{et} \quad g_2 = (5,4).$$

Itération 1

1. Étape d'affectation : On calcule la distance de chaque point à g_1 et g_2, puis on affecte chaque point au cluster correspondant au centre le plus proche :

$$\begin{aligned}
&d((1,1),g_1) = 0 \quad \text{et} \quad d((1,1),g_2) = 5, \quad \text{donc } (1,1) \in C_1,\\
&d((2,1),g_1) = 1 \quad \text{et} \quad d((2,1),g_2) = 4.24, \quad \text{donc } (2,1) \in C_1,\\
&d((4,3),g_1) = 3.61 \quad \text{et} \quad d((4,3),g_2) = 1, \quad \text{donc } (4,3) \in C_2,\\
&d((5,4),g_1) = 5 \quad \text{et} \quad d((5,4),g_2) = 0, \quad \text{donc } (5,4) \in C_2.
\end{aligned}$$

Résultat :

$$C_1 = \{(1,1),(2,1)\}, \quad C_2 = \{(4,3),(5,4)\}.$$

2. Étape de mise à jour : On recalcule les nouveaux centres (centroïdes) pour C_1 et C_2 :

$$g_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5,1), \quad g_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5,3.5).$$

Itération 2

1. Étape d'affectation : On répète le processus avec les nouveaux centres $g_1 = (1.5,1)$ et $g_2 = (4.5,3.5)$. Les calculs montrent que chaque point reste dans son cluster assigné.

$$C_1 = \{(1,1),(2,1)\}, \quad C_2 = \{(4,3),(5,4)\}.$$

2. Étape de mise à jour : Les centres g_1 et g_2 ne changent pas, donc l'algorithme converge.

Conclusion

- L'algorithme K-means alterne entre l'affectation des points aux clusters et le recalcul des centres.
- L'algorithme s'arrête lorsque les centres ne changent plus ou très peu (convergence).
- Chaque itération améliore la compacité des clusters en minimisant l'inertie intra-classe.

—

4 Articulations entre CAH et K-means selon les contextes (synthèse)

4.1 Quand utiliser CAH et K-means ?

- **Petits jeux de données** :
  - Utilisation possible de la CAH ou des K-means.
  - La CAH peut être utilisée pour déterminer le nombre de classes avant d'appliquer les K-means.
- **Gros jeux de données** :
  - Les K-means sont plus adaptés car la CAH devient coûteuse en calcul.
  - Méthode hybride souvent utilisée:
    1. Lancer un K-means avec $K_1 \ll n$ classes.
    2. Appliquer une CAH sur les K_1 centroïdes obtenus.
    3. Lancer un nouveau K-means sur K_2 classes déterminées.

4.2 Choix dans les K-means

Nombre de classes (K) :

- Connaissance a priori.
- Résultat d'une CAH.
- Utilisation de la règle du coude en testant différents K.

Centres de classes :

- **K-means** : Utilise les centres de gravité (*centroïdes*) des clusters.
- **K-medoids** :
  - Utilise les **médioïdes**, éléments centraux des clusters :

$$\text{médioïde} = \arg\min_{i \in C_k} \sum_{j \in C_k, j \neq i} \omega_j d^2(i,j).$$

  - Plus robuste aux outliers.

4.3 Initialisation des centres

- Centres choisis aléatoirement, avec plusieurs essais pour trouver la meilleure partition.
- Méthode K-means++ pour des centres éloignés et bien répartis.
- Centres définis par une CAH préliminaire.

4.4 Points clés

- La **CAH** et les **K-means** sont complémentaires.
- Une méthode hybride (CAH pour déterminer K, puis K-means pour optimiser) combine leurs avantages.
- Le choix des centres initiaux et du nombre de classes K est crucial pour la performance des K-means.

—

5 Articulations entre CAH et K-means selon les contextes

L'articulation entre la CAH et K-means dépend de deux facteurs principaux :

1. La **taille du jeu de données** (petit ou grand).
2. L'objectif de l'analyse (réduction de dimension, identification des clusters, optimisation des calculs).

5.1 Cas 1 : Petit jeu de données

- **Méthodes utilisées : CAH ou K-means.**
- **Pourquoi ?**
  - Dans les petits jeux de données, le coût calculatoire de la CAH n'est pas prohibitif, donc elle peut être utilisée directement pour construire une hiérarchie et déterminer un nombre de clusters.
  - K-means peut également être utilisé si le nombre de clusters est fixé a priori, mais il ne donne pas une hiérarchie comme la CAH.
- **Utilisation optimale :**
  - **CAH :** Détermine les clusters et construit la hiérarchie.
  - **K-means :** Part directement sur les données en ajustant les clusters pour minimiser l'inertie intra-classe.

5.2 Cas 2 : Gros jeu de données

- **Méthodes utilisées : Méthode hybride (K-means et CAH).**
- **Pourquoi ?**
  - La CAH est coûteuse en calcul sur des jeux de données volumineux, car elle nécessite de calculer toutes les distances entre individus.
  - K-means est plus adapté pour réduire la dimensionnalité et gérer de gros jeux de données rapidement.
- **Stratégie optimale :**
  1. **Étape 1 : K-means avec un petit K_1**
     Lancer un K-means sur les n individus pour obtenir un nombre réduit de K_1 centroïdes ($K_1 \ll n$). Cela réduit la dimension du problème.
  2. **Étape 2 : CAH sur les K_1 centroïdes**
     Utiliser la CAH pour analyser la hiérarchie entre ces centroïdes et déterminer le nombre optimal de clusters K_2. Cette étape permet de choisir un découpage plus pertinent.
  3. **Étape 3 : K-means global avec K_2**
     Repartir des individus initiaux pour exécuter un nouveau K-means en utilisant le nombre de clusters K_2 déterminé à l'étape précédente.

5.3 Points clés à connaître pour l'examen

1. **Pourquoi articuler CAH et K-means ?**
   - La combinaison des deux méthodes permet d'exploiter leurs avantages respectifs :
     - **CAH :** Identifier un nombre optimal de clusters et visualiser les hiérarchies.
     - **K-means :** Optimiser les regroupements pour de grands ensembles de données.
   - Cette approche est **flexible et scalable**, notamment pour de gros jeux de données.
2. **Les étapes à connaître par cœur :**
   - **K-means réduit la dimension en formant des centroïdes (K_1).**
   - **CAH analyse la hiérarchie entre ces centroïdes pour fixer un nombre de classes K_2.**
   - **K-means final regroupe les individus initiaux en K_2 clusters.**
3. **Particularité importante à retenir :**

- Lors du K-means final, on repart toujours des **individus initiaux** et non des clusters ou centroïdes obtenus précédemment.

4. **Avantages et limites :**

   - **Avantages :**
     - Méthode efficace pour de grands jeux de données.
     - Combinaison des forces : réduction de dimension (K-means) et hiérarchie optimale (CAH).
   - **Limites :**
     - Peut être coûteux pour des bases extrêmement volumineuses si K_1 n'est pas bien choisi.

5.4 Synthèse des articulations pour révision rapide

Taille des données	**Méthode**	**Objectif**
Petit jeu	CAH ou K-means	Classer les individus directement (CAH hiérarchiqu
Gros jeu	Méthode hybride (K-means + CAH + K-means)	Réduire la dimension, identifier K_2, optimiser les c

Conclusion

Cette articulation entre la CAH et K-means est essentielle à connaître pour tout examen. Vous devez maîtriser les étapes d'application selon la taille des données, les critères de choix entre CAH et K-means, et les spécificités de la méthode hybride pour maximiser les performances des deux approches.