DIFIQ
Niveau 2
Statistiques
Questions de révision: CAH et K-means

Jean-François Berger-Lefébure

21 Novembre 2024

Contents

1 Questions de révision

1.1 Classification Ascendante Hiérarchique (CAH)

1. **Quel est l'objectif principal de la CAH ?**
   Réponse : Regrouper les individus en formant une hiérarchie de clusters.
   Explication : La CAH construit une hiérarchie en regroupant successivement les clusters selon des critères de proximité.

2. **Quels types de distances peuvent être utilisées dans la CAH ?**
   Réponse : Euclidienne, Manhattan, Chebyshev, Mahalanobis.
   Explication : La CAH peut s'adapter à divers contextes en utilisant des distances adaptées à la nature des données.

3. **Quels types de liens sont utilisés pour définir la distance entre clusters dans la CAH ?**
   Réponse : Single Linkage, Complete Linkage, Average Linkage, Ward.
   Explication : Les stratégies de lien déterminent comment la distance entre deux clusters est mesurée.

4. **Quel est le rôle du dendrogramme dans la CAH ?**
   Réponse : Visualiser les fusions successives entre clusters.
   Explication : Le dendrogramme permet d'observer à quelle étape les clusters sont regroupés et de décider du nombre de classes.

5. **Pourquoi la méthode de Ward est-elle particulière ?**
   Réponse : Elle minimise l'inertie intra-classe et utilise uniquement la distance euclidienne.
   Explication : Contrairement aux autres méthodes, Ward a une base mathématique qui garantit une minimisation de l'inertie intra-classe.

1.2 K-means

1. **Quel est le critère que K-means cherche à minimiser ?**
   Réponse : L'inertie intra-classe.
   Explication : L'objectif de K-means est de minimiser la dispersion des points autour de leurs centroïdes.

2. **Quelle distance est généralement utilisée dans K-means ?**
   Réponse : La distance euclidienne.
   Explication : K-means fonctionne principalement avec la distance euclidienne pour mesurer la proximité entre un point et un centroïde.

3. **Quels sont les principaux inconvénients de K-means ?**
   Réponse : Nécessite de fixer K à l'avance et sensible à l'initialisation des centres.
   Explication : Fixer K à l'avance peut être difficile, et une mauvaise initialisation des centres peut conduire à un résultat sous-optimal.

4. **Qu'est-ce que l'algorithme K-means++ améliore ?**
   Réponse : L'initialisation des centres pour garantir une meilleure répartition initiale.
   Explication : K-means++ sélectionne les centres initiaux de manière stratégique pour éviter des clusters mal formés.

1.3 Méthodes hybrides (CAH et K-means)

1. **Pourquoi utilise-t-on une CAH avant un K-means dans une méthode hybride ?**
   Réponse : Pour déterminer un bon nombre de clusters K.
   Explication : La CAH aide à identifier un découpage approprié avant d'optimiser avec K-means.

2. **Quel avantage K-means a-t-il sur la CAH pour de grands jeux de données ?**
   Réponse : Il est moins coûteux en calcul.
   Explication : La CAH est très coûteuse car elle calcule toutes les distances possibles entre individus, alors que K-means est scalable.

1.4 Distances et liens dans CAH

1. **Quel lien utilise la distance minimale entre deux clusters ?**
   Réponse : Single Linkage.
   Explication : Single Linkage mesure la plus petite distance entre un point de chaque cluster.

2. **Quel lien utilise la distance maximale entre deux clusters ?**
   Réponse : Complete Linkage.
   Explication : Complete Linkage mesure la plus grande distance entre un point de chaque cluster.

3. **Quel lien utilise la moyenne des distances entre tous les points des clusters ?**
   Réponse : Average Linkage.
   Explication : Average Linkage mesure la distance moyenne entre tous les points des deux clusters.

1.5 Concepts mathématiques et interprétation des formules

1. **Que représente la formule suivante dans K-means ?**

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

   Réponse : L'inertie intra-classe.
   Explication : Cette formule mesure la compacité globale des clusters en additionnant les distances pondérées entre les points et leurs centroïdes.

2. **Que représente la formule suivante dans CAH (méthode de Ward) ?**

$$\Delta(C_1, C_2) = \frac{\mu_{C_1} \mu_{C_2}}{\mu_{C_1} + \mu_{C_2}} d^2(g_{C_1}, g_{C_2})$$

   Réponse : La perte d'inertie intraclasse due à la fusion de deux clusters.
   Explication : La méthode de Ward calcule la distance entre clusters en termes de perte d'inertie intraclasse.

2 Questions de révision ++

2.1 Classification Ascendante Hiérarchique (CAH)

1. **Quelle est la formule de la distance euclidienne entre deux individus x et y ?**
   Réponse :

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

   Explication : La distance euclidienne mesure la racine carrée de la somme des carrés des différences entre les coordonnées des deux individus.

2. **Quelle est la formule du Single Linkage entre deux clusters C_1 et C_2 ?**
   Réponse :

$$d(C_1, C_2) = \min_{i \in C_1, j \in C_2} d(i, j)$$

   Explication : Le Single Linkage mesure la plus petite distance entre un point de chaque cluster.

3. **Quelle est la formule du Complete Linkage entre deux clusters C_1 et C_2 ?**
   Réponse :

$$d(C_1, C_2) = \max_{i \in C_1, j \in C_2} d(i, j)$$

   Explication : Le Complete Linkage mesure la plus grande distance entre un point de chaque cluster.

4. **Quelle est la formule de l'Average Linkage entre deux clusters C_1 et C_2 ?**
   Réponse :
$$d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{i \in C_1} \sum_{j \in C_2} d(i, j)$$

   Explication : L'Average Linkage calcule la moyenne des distances entre tous les points des deux clusters.

5. **Quelle est la formule de la méthode de Ward ?**
   Réponse :
$$\Delta(C_1, C_2) = \frac{\mu_{C_1} \mu_{C_2}}{\mu_{C_1} + \mu_{C_2}} d^2(g_{C_1}, g_{C_2})$$

   Explication : La méthode de Ward mesure l'augmentation de l'inertie intra-classe due à la fusion de deux clusters.

2.2 K-means

1. **Quelle est la formule de l'inertie intra-classe dans K-means ?**
   Réponse :
$$\sum_{k=1}^{K} \sum_{i \in C_k} \omega_i d^2(i, g_k)$$

   Explication : Cette formule mesure la compacité des clusters en sommant les distances pondérées au carré entre chaque point et son centroïde.

2. **Quels sont les principaux inconvénients de K-means ?**
   Réponse : Nécessite de fixer K à l'avance et sensible à l'initialisation des centres.
   Explication : Fixer K à l'avance peut être difficile, et une mauvaise initialisation des centres peut conduire à un résultat sous-optimal.

3. **Qu'est-ce que l'algorithme K-means++ améliore ?**
   Réponse : L'initialisation des centres pour garantir une meilleure répartition initiale.
   Explication : K-means++ sélectionne les centres initiaux de manière stratégique pour éviter des clusters mal formés.

2.3 Concepts mathématiques et interprétation des formules

1. **Comment se lit cette formule ?**
$$\sum_{k=1}^{K} \sum_{i \in C_k} \omega_i d^2(i, g_k)$$

   Réponse : Pour chaque cluster C_k, on calcule la somme des distances quadratiques entre les individus et leur centroïde g_k, pondérées par leurs poids ω_i, et on additionne ces valeurs pour tous les clusters.
   Explication : Cette formule exprime la compacité globale des clusters, que l'algorithme K-means cherche à minimiser.