DIFIQ
Niveau 2
Statistiques
Les Inerties et le Théorème de Huygens

Jean-François Berger-Lefébure

21 Novembre 2024

Contents

1 Lien entre les Métriques/Distances, Centre de Gravité et Inertie

Les **distances** permettent de mesurer la **proximité** entre les individus et sont essentielles pour la création des **clusters**. Dans des méthodes comme le **K-means** ou la **Classification Ascendante Hiérarchique (CAH)**, ces distances sont utilisées pour regrouper les individus en fonction de leur similarité.

Centre de Gravité

Une fois les individus regroupés en **clusters**, le **centre de gravité** de chaque cluster est calculé pour définir un point central autour duquel les individus du cluster sont concentrés. Ce centre de gravité est obtenu en prenant la **moyenne pondérée des coordonnées** des individus, pondérée par leurs poids.
Le centre de gravité reflète la **cohésion interne** du cluster : plus les individus sont proches les uns des autres, plus le centre de gravité sera proche de leur position.

Inertie

L'**inertie** mesure la qualité du clustering en quantifiant à quel point les individus au sein d'un même cluster sont proches de leur centre de gravité.
Elle est calculée comme la somme des carrés des distances entre chaque individu et le centre de gravité du cluster:
Une faible inertie indique que les individus sont bien regroupés autour du centre de gravité.
L'**inertie globale** mesure l'efficacité du clustering sur l'ensemble des clusters : elle est faible si les clusters sont cohésifs et bien séparés.

Lien entre Distances, Centre de Gravité et Inertie

- Les **distances** entre individus déterminent la manière dont les **clusters** sont formés.
- Le **centre de gravité** est calculé à partir des distances entre les individus du cluster.
- **L'inertie** mesure la qualité des regroupements en fonction des distances : elle quantifie si les individus sont proches du centre de gravité et si les clusters sont bien séparés.

En résumé, les distances sont utilisées pour former les clusters, le centre de gravité représente la position centrale de chaque cluster, et l'inertie quantifie la cohésion et la séparation des clusters.

—

2 Centre de gravité

Le **centre de gravité** est un point central qui représente la position moyenne pondérée d'un ensemble d'individus.
Formule :

$$g = \sum_{i=1}^{n} \omega_i x_i$$

Lecture: Le centre de gravité g est obtenu en faisant la somme pondérée (ω_i) des positions (x_i) de tous les individus, de $i = 1$ à n.
Interprétation : Chaque individu x_i est associé à un poids ω_i, qui reflète son importance. Si tous les poids sont égaux, cela revient à calculer une moyenne classique.

2.1 Cas des partitions en classes

On suppose que les individus sont regroupés en K classes notées $C_1, C_2, \ldots, C_K$. Chaque classe C_k possède :

Poids total d'une classe :

$$\mu_k = \sum_{i \in C_k} \omega_i$$

Lecture : Le poids total μ_k d'une classe C_k est la somme des poids ω_i des individus appartenant à cette classe.

Centre de gravité d'une classe :

$$g_k = \frac{1}{\mu_k} \sum_{i \in C_k} \omega_i x_i$$

Lecture : Le centre de gravité g_k d'une classe C_k est obtenu en calculant la moyenne pondérée des positions (x_i) des individus dans cette classe. Les poids sont normalisés par le poids total μ_k.

2.2 Exemple

Considérons deux classes :

- C_1 avec 3 individus de positions $x_1 = 2$, $x_2 = 4$, $x_3 = 6$ et poids égaux ($\omega_i = 1$).
- C_2 avec 2 individus de positions $x_4 = 8$ et $x_5 = 10$ et poids égaux ($\omega_i = 1$).

Classe C_1 :

$$\mu_1 = 1 + 1 + 1 = 3$$

$$g_1 = \frac{1}{3}(2 \cdot 1 + 4 \cdot 1 + 6 \cdot 1) = \frac{12}{3} = 4$$

Classe C_2 :

$$\mu_2 = 1 + 1 = 2$$

$$g_2 = \frac{1}{2}(8 \cdot 1 + 10 \cdot 1) = \frac{18}{2} = 9$$

Centre de gravité global :

$$g = \frac{3 \cdot 4 + 2 \cdot 9}{3 + 2} = \frac{12 + 18}{5} = 6$$

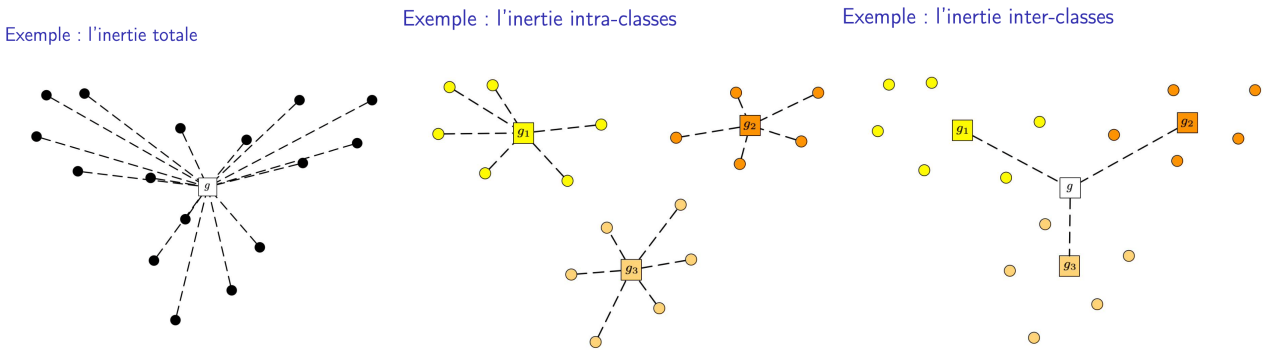
2.3 Questions pour réviser

Question 1 : Que représente le centre de gravité dans un nuage de points ?
Réponse : Il représente la position moyenne pondérée des points dans l'espace.
Question 2 : Comment calcule-t-on le poids total d'une classe C_k ?
Réponse : En additionnant les poids des individus présents dans cette classe : $\mu_k = \sum_{i \in C_k} \omega_i$.
Question 3 : Comment calcule-t-on le centre de gravité d'une classe C_k ?
Réponse : En prenant la moyenne pondérée des positions des individus dans la classe : $g_k = \frac{1}{\mu_k} \sum_{i \in C_k} \omega_i x_i$

—

3 Les Inerties

L'**inertie** mesure la qualité du clustering en quantifiant à quel point les individus au sein d'un même cluster sont proches de leur centre de gravité.



3.1 Inertie totale :

L'inertie totale mesure la dispersion globale des individus par rapport au centre de gravité g.
Formule :

$$I_{tot} = \sum_{i=1}^{n} \omega_i d^2(i, g)$$

Lecture : L'inertie totale I_{tot} est la somme pondérée (ω_i) des distances au carré (d^2) entre chaque individu i et le centre de gravité global g.
Interprétation : Elle reflète l'étalement général des points dans l'espace.

3.2 Inertie intra-classes :

L'inertie intra-classes mesure la concentration des points au sein de chaque classe.
Formule :

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

Lecture : L'inertie intra-classes I_{intra} est la somme des distances au carré entre chaque individu i et le centre de gravité de sa classe g_k, pondérée par ω_i.
Interprétation : Elle mesure la compacité des points dans chaque classe. Une inertie intra faible signifie que les points sont proches les uns des autres au sein des classes.

3.3 Inertie inter-classes :

L'inertie inter-classes mesure l'éloignement entre les centres de gravité des classes et le centre de gravité global.
Formule :

$$I_{inter} = \sum_{k=1}^{K} \mu_k d^2(g_k, g)$$

Lecture : L'inertie inter-classes I_{inter} est la somme des distances au carré entre chaque centre de gravité de classe g_k et le centre de gravité global g, pondérée par μ_k (poids total de la classe).
Interprétation : Elle mesure la séparation entre les classes. Une inertie inter élevée indique des classes bien séparées.

3.4 Exemple de calcul de l'Inertie Totale (sans le Théorème de Huygens)

Soit un ensemble de 5 individus répartis en 2 classes :

- Classe 1 : $x_1 = 1, x_2 = 2$.
- Classe 2 : $x_3 = 8, x_4 = 9, x_5 = 10$.

Étape 1 : Calcul du centre global

$$g = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{1 + 2 + 8 + 9 + 10}{5} = \frac{30}{5} = 6$$

Étape 2 : Calcul de l'inertie totale

$$I_{\text{totale}} = (x_1 - g)^2 + (x_2 - g)^2 + (x_3 - g)^2 + (x_4 - g)^2 + (x_5 - g)^2$$

$$I_{\text{totale}} = (1 - 6)^2 + (2 - 6)^2 + (8 - 6)^2 + (9 - 6)^2 + (10 - 6)^2$$

$$I_{\text{totale}} = (-5)^2 + (-4)^2 + 2^2 + 3^2 + 4^2$$

$$I_{\text{totale}} = 25 + 16 + 4 + 9 + 16 = 70$$

—

4 Théorème de Huygens

Le théorème de Huygens établit la relation fondamentale suivante :

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

Lecture : L'inertie totale est égale à la somme de l'inertie intra-classes et de l'inertie inter-classes.
Interprétation : Ce résultat montre que la dispersion totale peut être décomposée en deux parties :

- La dispersion interne aux classes (intra).
- La dispersion entre les classes (inter).

4.1 Exemple

Soit un ensemble de 5 individus répartis en 2 classes :

- Classe 1 : $x_1 = 1, x_2 = 2$.
- Classe 2 : $x_3 = 8, x_4 = 9, x_5 = 10$.

Étape 1 : Calcul des centres de gravité

$$g_1 = \frac{1+2}{2} = 1.5, \quad g_2 = \frac{8+9+10}{3} = 9$$

Étape 2 : Calcul de I_{intra}

$$I_{intra} = (1-1.5)^2 + (2-1.5)^2 + (8-9)^2 + (9-9)^2 + (10-9)^2$$

$$I_{intra} = 0.5 + 0.5 + 1 + 0 + 1 = 3$$

Étape 3 : Calcul de I_{inter} Centre global :

$$g = \frac{1+2+8+9+10}{5} = 6$$

$$I_{inter} = 2(1.5-6)^2 + 3(9-6)^2$$

$$I_{inter} = 2(20.25) + 3(9) = 40.5 + 27 = 67.5$$

Étape 4 : Vérification de Huygens

$$I_{tot} = I_{intra} + I_{inter} = 3 + 67.5 = 70.5$$

4.2 Questions pour réviser

Question 1 : Comment interpréter un faible I_{intra} associé à un I_{inter} élevé dans un clustering ?
Réponse : Cela indique que les classes sont bien compactes et bien séparées, suggérant un bon regroupement.
Question 2 : Pourquoi le théorème de Huygens est-il utile dans l'analyse des inerties ?
Réponse : Il permet de décomposer l'inertie totale en deux composantes pour évaluer la qualité de la classification en comparant homogénéité et séparation.
Question 3 : Que signifie un I_{inter} nul dans une partition ?
Réponse : Cela signifie que toutes les classes ont leur centre de gravité identique au centre global, donc aucune séparation entre elles.
Question 4 : Comment peut-on minimiser I_{intra} tout en maximisant I_{inter} dans un algorithme de clustering ?
Réponse : En optimisant les positions des centres de gravité et en ajustant les frontières des classes pour maximiser leur éloignement.

Conclusion

L'étude des inerties permet d'analyser la structure des données en termes de dispersion et de regroupement. Le théorème de Huygens fournit une décomposition simple pour évaluer la qualité des partitions dans un clustering. Un bon modèle vise à minimiser I_{intra} tout en maximisant I_{inter}.

5 Résumé et transition vers les Algorithmes

L'inertie permet de mesurer la dispersion des points à l'intérieur des clusters (inertie intra-cluster) et entre les clusters (inertie inter-cluster). L'objectif est de minimiser l'inertie intra-cluster tout en maximisant l'inertie inter-cluster, afin d'obtenir un clustering à la fois cohérent et bien séparé.

- **Minimiser l'inertie intra-cluster** : L'objectif est de rendre les points d'un même cluster aussi proches que possible du centre du cluster. Cela permet d'obtenir des clusters homogènes et de minimiser l'inertie intra-cluster.
- **Maximiser l'inertie inter-cluster** : L'objectif est de maximiser la distance entre les centres des différents clusters. Cela assure que les clusters sont bien séparés les uns des autres.

Qualité du clustering : La qualité d'un clustering peut être évaluée par le ratio

$$\frac{I_{\text{inter}}}{I_{\text{tot}}}$$

où I_{inter} est l'inertie inter-cluster et I_{tot} est l'inertie totale. Ce ratio mesure la proportion de l'inertie totale qui est expliquée par la séparation entre les clusters.
CAH vs K-means : La Classification Ascendante Hiérarchique (CAH) n'est pas conçue pour minimiser l'inertie intra-cluster de manière optimale, contrairement au K-means. Ce dernier cherche spécifiquement à minimiser cette inertie, mais ce n'est pas l'objectif principal de la CAH, qui se concentre plutôt sur la hiérarchisation des clusters sans nécessairement optimiser leur compacité.