Apprentissage non-supervisé

UQAC | Université du Québec à Chicoutimi

Apprentissage non-supervisé (clustering ou segmentation)

Objectif d'affaire

- Mieux cibler la publicité
- Proposer de meilleurs offres
- Développer de nouveaux produits
- Améliorer le service à la clientèle
- Identifier des groupes sous-représentés dans la clientèle

Objectif mathématique

- Séparer (partitionner, segmenter) les observations en un certain nombre de groupes homogènes. Il s'agit d'une méthode plutôt descriptive

UQAC

«Apprentissage» supervisé vs non-supervisé

Supervisé

- On connaît le groupe pour chacun de nos clients
- On connaît le nombre total de groupe
- L’objectif est associé à une variable cible en particulier

Non-supervisé

- On ne connaît pas d’avance le groupe auquel appartient un client
- On ne connaît pas le nombre de groupes

UQAC

Est-ce de l’apprentissage supervisé ou non ?

Une radio souhaite se repositionner dans le marché. Le contexte est très compétitif et elle souhaite revoir sa programmation musicale et sa publicité. Elle ne possède pas de données sur sa clientèle.

https://app.wooclap.com/ Code:DMHGCH

UQAC

Est-ce de l’apprentissage supervisé ou non ?

Un OBNL cherche à optimiser sa campagne de financement. Dans ce contexte, l’organisme souhaite n’appeler que les personnes les plus susceptibles de faire un don. Elle a en sa possession une liste de donateurs des années passées. Cette base de données contient le montant de leur don et certaines informations sur les donateurs.

https://app.wooclap.com/ Code:DMHGCH

UQAC

Est-ce de l’apprentissage supervisé ou non ?

Une entreprise souhaite prédire l’achalandage dans son centre d’appel pour mieux planifier les ressources humaines. L’entreprise possède l’historique de son achalandage depuis plusieurs années.

https://app.wooclap.com/ Code:DMHGCH

UQAC

Est-ce de l’apprentissage supervisé ou non ?

Une banque souhaite améliorer l’expérience de ses clients (ou de ses employés) en personnalisant davantage le service qu’elle leur offre. Elle constate que sa connaissance de sa clientèle est plutôt déficiente, bien qu’elle possède une vaste base de données.

https://app.wooclap.com/ Code:DMHGCH

UQAC

Les familles de techniques de segmentation

1. K-moyenne
2. Algorithme hiérarchique

UQAC

En pratique

Imaginez le contexte d'une station radio voulant repositionner son offre musicale. La station a fait écouter des extraits musicaux de type variés (jazz, classique, rock, pop, pop rock…) et les clients devaient noter leur appréciation de la musique sur une échelle de 1 à 10. L'ensemble des variables est décrit dans le tableau suivant :

Nom	Description	Type
id	Indentifiant unique	Caractère
m1 à m9	Appréciation des différents types de musique sur une échelle de 1 à 10	Ordinale traitée comme numérique
genre	Genre	Binaire
freq_radio	Fréquence à laquelle le répondant écoute la radio	Ordinale
musicien	Est-ce que le répondant est un musicien professionnel	Binaire
transport	Moyen de transport pour se rendre au travail	Nominal

UQAC

En pratique

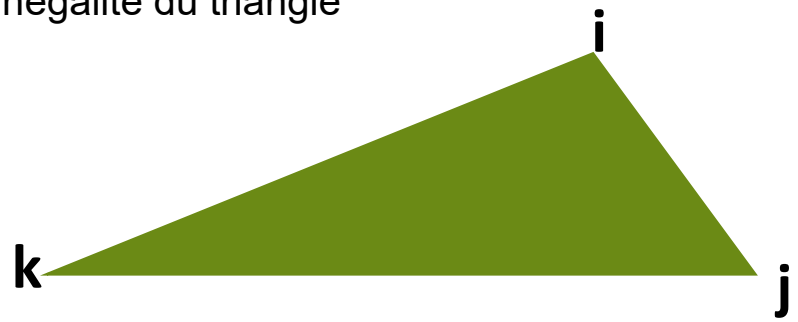
Voici un extrait des données:

id	m1	m2	m3	m4	m5	m6	m7	m8	m9	genre	freq_radio	musicien	transport
11	8	9	7	6	6	6	2	7	2	Femme	1_jamais	non	auto
13	8	7	3	4	2	4	8	8	8	Femme	1_jamais	non	auto
1	10	8	8	8	9	7	9	10	8	Homme	1_jamais	non	bus
10	8	10	9	6	5	6	4	5	2	Femme	1_jamais	non	bus
21	3	2	3	4	3	5	7	8	9	Homme	1_jamais	non	bus
22	2	5	5	5	4	7	8	9	6	Homme	1_jamais	non	marche

Les mesures de distance

Une mesure de distance doit:

1. d(i, j) ≥ 0 : C'est une valeur positive, une distance ne doit pas être négative
2. d(i, i) = 0 : La distance entre un point et lui même est nulle
3. d(i, j) = d(j, i): La distance entre le point A et le point B est la même qu'entre le point B et le point A
4. d(i, k) ≤ d(i, j) + d(j, k): C'est l'inégalité du triangle



UQAC

Similarité et dissemblance

Un indice de dissemblance est une valeur entre 0 et 1

Plus les individus se ressemblent, plus l'indice est faible

$$d^*(i,j) = 1 - s(i,j)$$

Un indice de similarité est une valeur entre 0 et 1

Plus les individus se ressemblent, plus l'indice est élevé

$$s(ij) = \frac{1}{1 + d(i,j)}$$

UQAC

Mesure de distance et type de variables

Le choix de la mesure de distance dépend du type de variables:

- Variables numériques
    - Il s’agit de variables dont la valeur numérique mesure quelque chose de quantifiable et dont la différence entre les valeurs reflète la différence entre les objets. On peut ainsi parler du revenu en dollars, de la masse, de l’âge, etc
- Variables nominales
- Variables ordinales

Mesure de distance: Variable Numérique

- Distance euclidienne

$$d_2(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}.$$

	Âge	Revenu
Individu 1	20	70
Individu 2	30	45
Individu 2	40	90

$$d_2(1,2) = \sqrt[2]{(20-30)^2+(70-45)^2} = 26.93$$

UQAC

Mesure de distance: Variable Numérique

- Distance euclidienne

Attention distance n'est PAS invariante à un changement d'échelle.

	l'individu 2	l'individu 3
Unités en k$	26.93	28.28
Unités en $	2.5×10^4	2×10^4

Pour régler ce problème, on peut travailler avec une distance standardisée entre les variables (on soustrait la moyenne et on divise par l'écart type)

UQAC

Mesure de distance: Variable Nominale (catégorielle)

Binaire (genre, fraude ou pas)

On commence par coder l'une modalité à 0 et l'autre à 1. Si on mesure p variables binaires pour chacun de deux individus i et j, on compte le nombre de variables pour lesquelles ces deux individus ont la même valeur pour une même variable

Individu	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
i	1	0	0	0	0	1	0	0	0	0
j	1	0	0	0	0	0	1	0	0	0

UQAC

Mesure de distance: Variable Nominale (catégorielle)

Binaire (genre, fraude ou pas)

Si une modalité est plus rare qu'une autre:

On assigne la modalité 1 à la valeur la plus rare (ou la plus importante) et la valeur 0 à l'autre modalité.

On peut ensuite utiliser **l'indice de Jaccard** défini par le nombre de variables pour lesquelles i et j prennent simultanément la valeur 1 sur le nombre de variables pour lesquelles au moins l'un de i ou j n'a pas la valeur 0

$$J(i,j) = \frac{\text{Nombre de variables où i et j} = 1}{\text{Nombre de variables où i ou j} = 1}$$

UQAC

Mesure de distance: Variable Nominale (catégorielle)

Polytomique (plusieurs catégories)

Si une variable est composée de plusieurs modalités, on peut la coder en utilisant $M - 1$ variables binaires (indicatrices).

Individu	Oui	Non	Je.ne.sais.pas
1	1	0	0
2	0	1	0

Mesure de distance: Variable Ordinale

Exemple

- Niveau de satisfaction (Satisfait, Moyennement satisfait, peu satisfait)
- Variables continues converties en catégories (groupes d’âges)

Solution:

- On leur assigne un rang et on les traite comme des variables numériques

Mesure de distance: Plusieurs types de variables

Procédure:

1. Recoder les variables nominales et ordinales
2. Déterminer le poids de chaque variable
3. Utiliser l'indice de Gower:
   - C'est un indice qui donne une sorte de moyenne pondérée de toutes les mesures de distance présentées précédemment.

Regroupement

On veut partitionner les n observations en K groupes de façon à ce que:

- les observations à l'intérieur d'un groupes soient le plus similaire possible
- les observations de deux groupes différents soient le moins similaires possible

Attention: Aucun algorithme ne garantit de trouver un optimum global!

Il faudrait faire tous les regroupements possibles. . . À titre d'exemple, il y a 8.5896253 × 1046 façons de partitionner 100 individus en 3 groupes

UQAC

Algorithme des k-moyennes

Conditions pour utiliser cette algorithme:

1. Variables quantitatives
2. Distances euclidienne

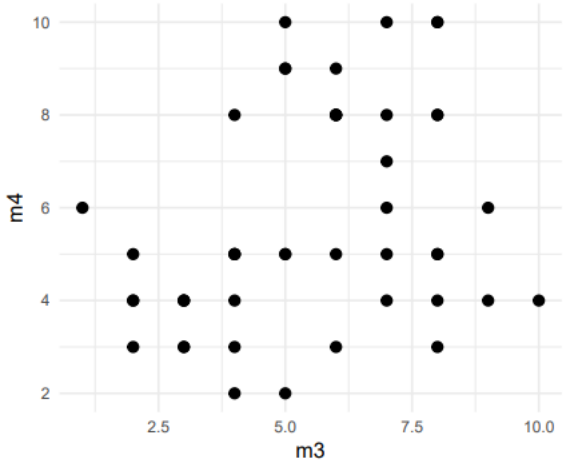
Algorithme des k-moyennes

Procédure:

1. On choisit le nombre de groupes K que l'on désire obtenir.
2. On partitionne aléatoirement les n observations en K groupes.
3. On calcule les coordonnées des centroïdes pour chacun des K groupes.
4. On calcule la distance entre chaque observation et chacun des K centroïde
5. On assigne chacune des n observations au groupe dont le centroïde est le plus près.
6. On répète les étapes 3 à 5 jusqu'à ce qu'aucune observation ne soit réassignée à un nouveau groupe.

Exemple

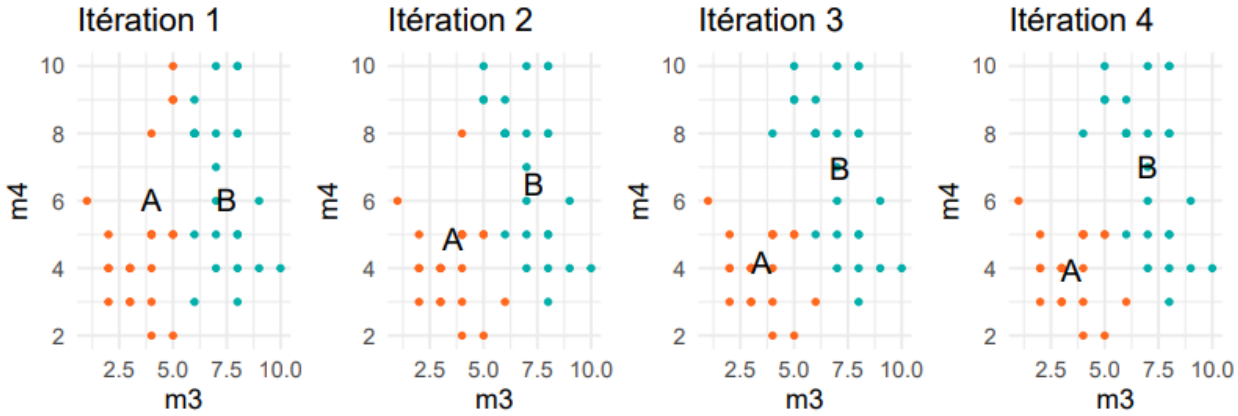
On veut créer **2 groupes** selon les variables **m4** et **m3**.



UQAC

Exemple

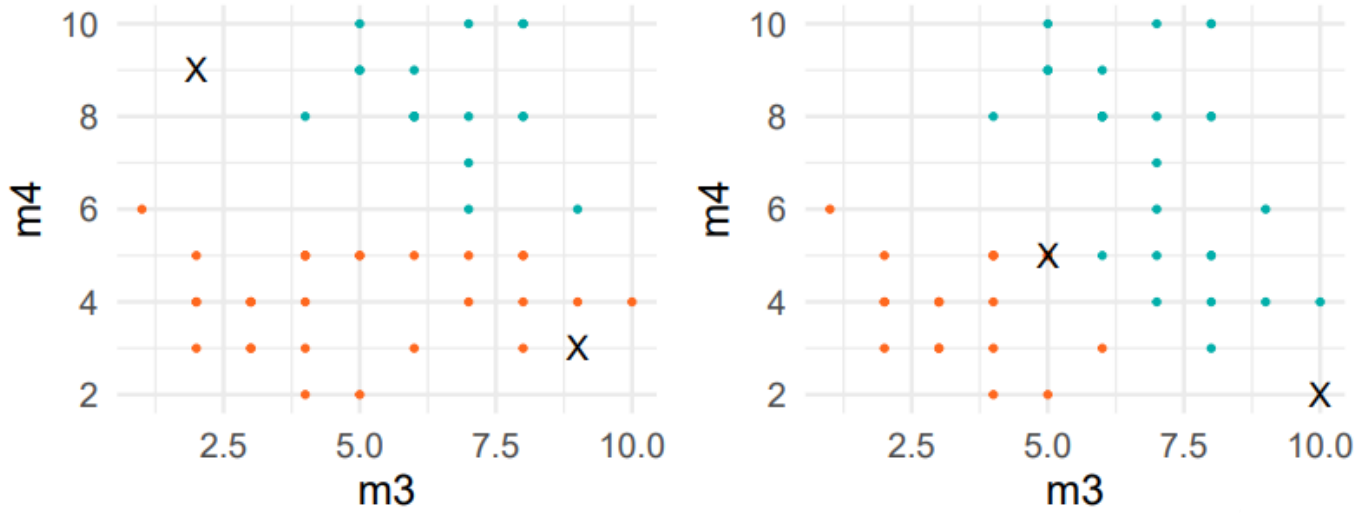
On veut créer **2 groupes** selon les variables **m4** et **m3**.



UQAC

Sensibilité au choix des centroïdes initiaux

En pratique, il est recommandé de tester plusieurs centroïdes initiaux.
Si tous les centroïdes initiaux mènent à des résultats similaires, les groupes sont probablement bien divisés et on peut accorder un bon niveau de confiance à la segmentation

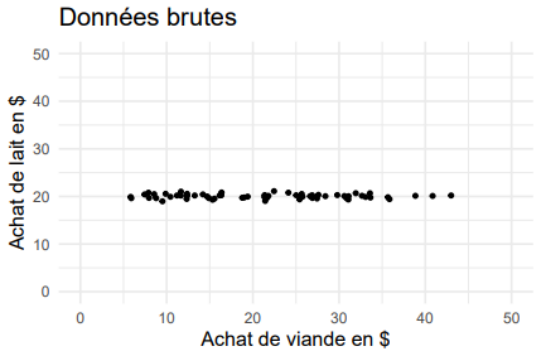


UQAC

L’effet de la standardisation

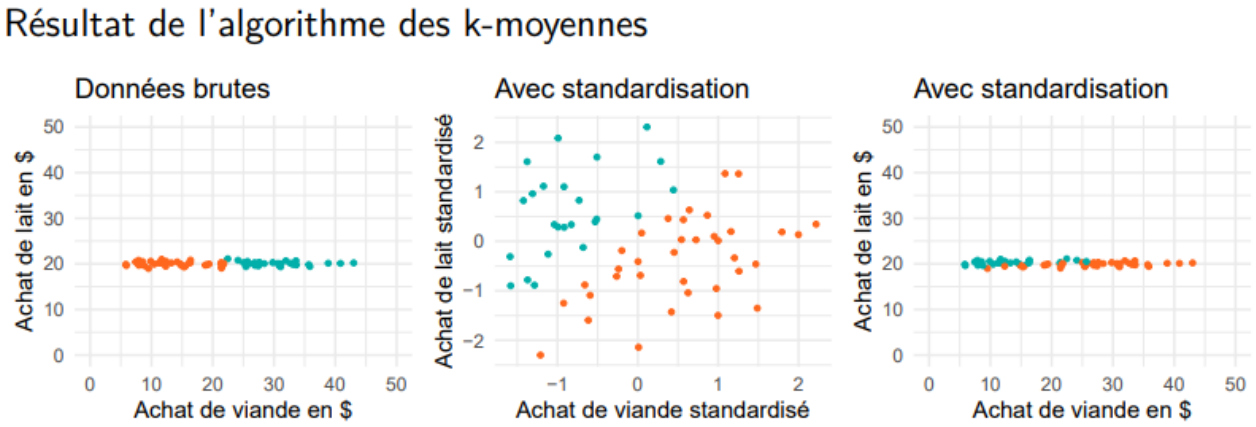
Bien qu’il soit généralement recommandé de standardiser les données avant de faire une segmentation, ce n’est pas toujours une bonne idée. Lorsque la segmentation implique des données de même nature et mesurées avec une même unité de mesure, ceci peut même s’avérer très contreproductif.

Exemple: Un épicier souhaite faire une segmentation de sa clientèle sur la base des achats effectués par les clients.



UQAC

L’effet de la standardisation



UQAC

Algorithme des k-medoïdes

L’algorithme des k-médoïdes est similaire à celui des k-moyennes, à la différence majeure qu’il permet d’inclure des variables de nature différente.

On n’utilise plus le centre, mais l’observations qui minimise les distances dans chaque groupe.

Points forts
- Permet d’intégrer des variables nominales
- Robuste (moins sensible aux valeurs extrêmes)
- Permet de bien spécifier la matrice de distance

Points faibles
- Il faut connaître le nombre de groupes
- Sensible au choix des centroïdes initiaux

Algorithmes hiérarchiques

La classification hiérarchique permet d'obtenir des partitions imbriquées les unes dans les autres, généralement présentées sous forme de dendrogramme.

Cette façon de segmenter est particulièrement utile dans les cas où on n'a aucune idée de nombre de segments ou l'on souhaite une segmentation à plusieurs niveaux.

Il existe deux grandes familles d'algorithmes hiérarchiques: les algorithmes ascendants (agglomératifs) et les algorithmes descendants. La première famille étant la plus populaire, c'est celle qui sera couverte ici.

Algorithmes hiérarchiques ascendants

Procédure:

1. Chaque observation est son propre groupe, c'est-à-dire qu'on démarre avec n groupes contenant chacun une seule observation
2. On fusionne les deux groupes les plus similaires.
3. On répète l'étape 2 jusqu'à ce qu'on obtienne un seul groupe contenant toutes les n observations.

Avec cet algorithme on doit travailler avec une dissemblance non pas entres des observations mais entre deux groupes. Nous allons voir quelques méthodes pour calculer cette dissemblance.

Plus proche voisin (single linkage)

La dissemblance entre deux groupes se définit comme étant la distance entre les deux observations les plus proches entre les deux groupes.

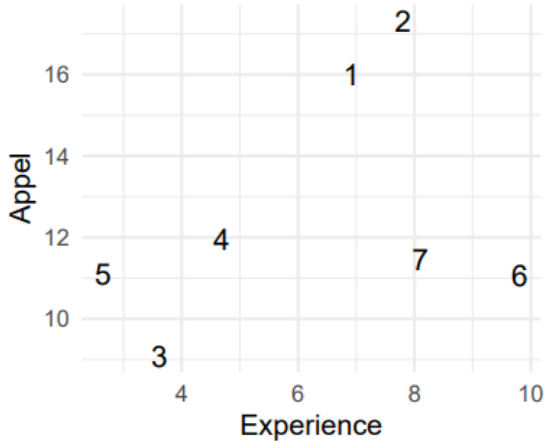
Principales caractéristiques:

- Cette méthode permet de créer des groupes dont la forme est très irrégulière.
- Elle est aussi très robuste aux valeurs extrêmes.
- Elle a tendance à former un grand groupe avec plusieurs petits groupes satellites, ce qui lui fait perdre un peu d'efficacité lorsque les groupes ont une forme plutôt régulière.

Plus proche voisin (single linkage)

Exemple

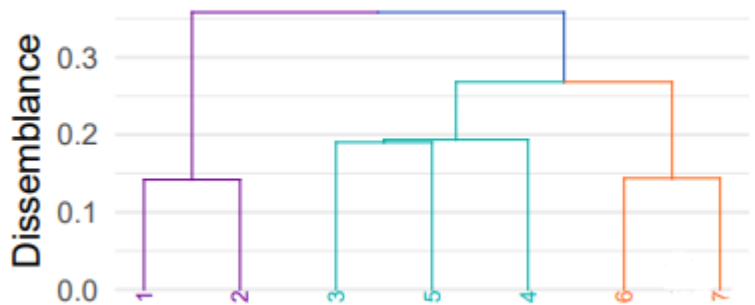
Données sur le nombre d'appels traitées par les employés d'un centre d'appel, selon leur expérience.



Matrice de distance

	1	2	3	4	5	6	7
1							
2	0.14						
3	0.65	0.79					
4	0.40	0.54	0.25				
5	0.60	0.74	0.19	0.19			
6	0.50	0.52	0.55	0.41	0.50		
7	0.36	0.38	0.46	0.27	0.40	0.14	

Dendrogramme



UQAC

Méthode du voisin le plus distant (complete linkage)

On utilise la dissemblance entre les observations les plus éloignées de chaque groupe.

Principales caractéristiques:

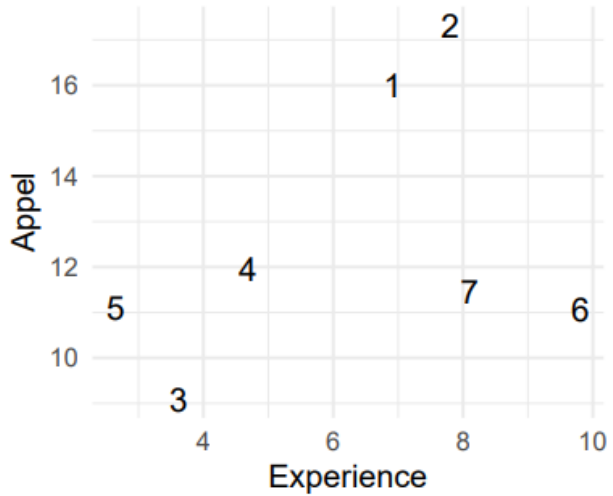
- Très sensible aux données aberrantes
- Tend à former des groupes de tailles égales

UQAC

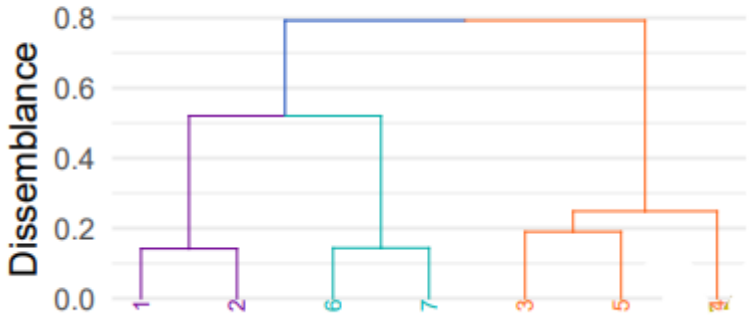
Méthode du voisin le plus distant (complete linkage)

Exemple - Voisin le plus distant

Nombre d'appels traitées par les employés d'un centre d'appel, selon leur expérience.



Dendrogramme



UQAC

Méthode de la moyenne (average linkage)

On utilise la dissemblance entre deux groupes se définit comme étant la moyenne des dissemblance entre chaque individu des deux groupes.

Principales caractéristiques:

Groupes de faible variance et de même variance.

Plus exigeant au niveau computationnel.

Méthode du centroïde

La distance entre deux groupes se définit comme étant la dissemblance entre les centroïdes des deux groupes.

Principales caractéristiques:

Robuste aux données aberrantes

Est peu efficace en l’absence de données aberrantes.

Méthode de la médiane

La dissemblance entre deux groupes se définit comme étant la dissemblance médiane entre toutes les observations des deux groupes.

Principales caractéristiques:

Robuste aux données aberrantes

Est peu efficace en l’absence de données aberrantes.

Algo. Hiérarchique: Avantages & inconvénients

Avantages:

- Facilite le choix du nombre de groupes
- Permet une hiérarchie des regroupements

Inconvénients:

- Une fois que deux observations sont regroupées, on ne peut plus les séparer.

Segmentation à 2 étapes

Il peut être intéressant de combiner les deux approches, par ce que l'on appelle parfois une classification à deux étapes.

Procédure:

1. Segmentation avec algorithme hiérarchique pour déterminer le nombre de groupes intéressants
2. Utiliser les centroïdes de ces groupes pour initialiser l'algorithme des k-moyennes ou des k-médoides.

Choix de la bonne méthode

Il n’y a ni règle claire ni étude sur le sujet.

Voici quelques cas où il est plus avantageux de choisir une méthode où l’autre:

K-moyenne

- Bonne idée du nombre de groupes
- Variables continues
- Si on a une idée préalable des centroïdes des groupes

Hiérarchique

- Aucune idée du nombre de groupes
- Besoin d’une segmentation à plusieurs niveaux
- Variables de nature différente

Choix du nombre de groupe

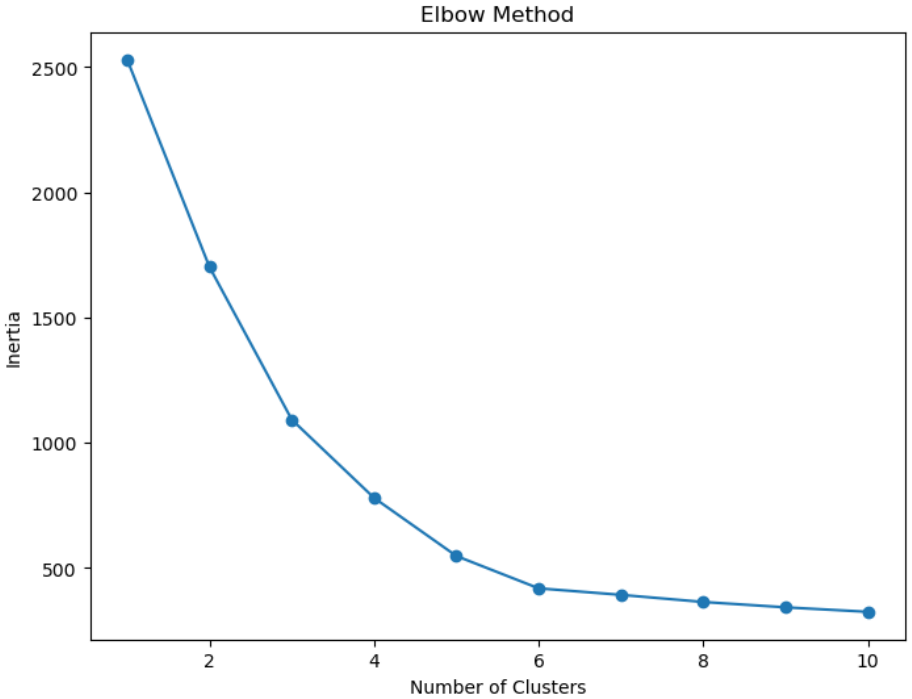
- Plusieurs des indicateurs reposent sur le concept d'inertie.
- L'inertie totale est la somme des distances au carré entre toutes les observations. L'inertie d'une segmentation se décompose en deux parties : l'inertie à l'intérieur des groupes et l'inertie entre les groupes.
- Ces indicateurs sont plus pertinents avec des variables continues
- Prendre garde au poids des variables et à la standardisation

Choix du nombre de groupe

Elbow Method (Méthode du coude)

Tracer la somme des distances au carrés intra-groupe (inertie) en fonction du nombre de groupe.

Le "coude" de la courbe représente un nombre optimal de groupe où l'ajout de plus de groupe ne réduit pas significativement l'inertie.



UQAC

Choix du nombre de groupe

- Privilégiez l’interprétabilité et l’utilité des groupes à un critère quelconque.
- Évitez les critères basés sur l’inertie ou la variance pour des groupes de taille et d’étendue inégales.

UQAC

Réduction de la dimensionnalité

- Très utile lorsque le nombre de variables est très élevé!
- On peut utiliser l’analyse en composante principale (ACP) ou l’analyse des correspondances.
- Permet de visualiser les groupes

UQAC