Fondation ubs Université Bretagne Sud

CHAIRE D-CC "DÉCISIONNEL-CONNAISSANCE CLIENT" - IUT de Vannes - 8 rue Montaigne - 56017 Vannes Cx / 02 97 62 64 64 / chaire.d-cc@univ-ubs.fr

Classification non-supervisée : méthodes géométriques

M. Gousseff-Fondation UBS

19 janvier 2017

Rappel sur le partitionnement
Classification non supervisée
Méthodes de partitionnement
Hypothèse implicite
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels

Rappel sur le partitionnement

Classification non supervisée

Rappel sur le partitionnement
Classification non supervisée
Méthodes de partitionnement
Hypothèse implicite
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels

Soit n individus sur lesquels on mesure p variables. On note ce jeu de données $X = \{x_i \in \mathbb{R}^p\}_{i=1,\ldots,n}$

On fait l'hypothèse que cette population n'est pas homogène. Il

existe des classes sous-jacentes qu'on va essayer de retrouver : **classification**

On ne dispose même pas d'un échantillon sur lequel on connaîtrait ces classes :
non supervisée
Afin d'éviter toute confusion avec la classification supervisée (=classement)on parle de **partitionnement** ou en anglais de **clustering**

Méthodes de partitionnement

Rappel sur le partitionnement
Classification non supervisée
Méthodes de partitionnement
Hypothèse implicite
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels

- ✔ = Approches géométriques
  - ✘ Clustering Hiérarchique (Classification Ascendante Hiérarchique)
  - ✘ Clustering par affectation dynamique = K-Means
- ✔ = Approche modélisation
  - ✘ Clustering à l'aide de modèle de mélange : le moins utilisé, un des plus performants !

Hypothèse implicite

Rappel sur le partitionnement
Classification non supervisée
Méthodes de partitionnement
Hypothèse implicite
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels

Ranger des individus (des observations) en classes en s'aidant des caractéristiques (traduites par les valeurs de variables observées sur ces individus) de sorte que :

- ✔ 2 objets d'un même groupe se ressemblent le plus possible
- ✔ 2 objets de groupes distincts diffèrent le plus possible

Source : cours Arlette Antoni

Hypothèse implicite

Rappel sur le partitionnement
Classification non supervisée
Méthodes de partitionnement
Hypothèse implicite
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels

Ranger des individus (des observations) en classes en s'aidant des caractéristiques (traduites par les valeurs de variables observées sur ces individus) de sorte que :

- ✔ 2 objets d'un même groupe se ressemblent le plus possible
- ✔ 2 objets de groupes distincts diffèrent le plus possible

Source : cours Arlette Antoni
Si, entre les "vrais" groupes sous-jacents, il n'existe pas de différences pour les variables qu'on mesure, les algorithmes qu'on s'apprête à étudier donneront des résultats peu pertinents.

Hypothèse implicite

Rappel sur le partitionnement
Classification non supervisée
Méthodes de partitionnement
Hypothèse implicite
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels

Ranger des individus (des observations) en classes en s'aidant des caractéristiques (traduites par les valeurs de variables observées sur ces individus) de sorte que :

- ✔ 2 objets d'un même groupe se ressemblent le plus possible
- ✔ 2 objets de groupes distincts diffèrent le plus possible

Source : cours Arlette Antoni
Si, entre les "vrais" groupes sous-jacents, il n'existe pas de différences pour les variables qu'on mesure, les algorithmes qu'on s'apprête à étudier donneront des résultats peu pertinents.

Comment Quantifier "se ressemblent" ou "diffèrent" ?

Rappel sur le partitionnement
Distances
Notations distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

Distances

Notations

Rappel sur le partitionnement
Distances
Notations
distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

On décrit n individus en mesurant p variables aléatoires sur un espace E.

$$X_{ij} = (X_{i1}, \cdots, X_{ip})$$

avec $1 \leq i \leq n$

Si les variables sont quantitatives $E = R$.
Si les variables sont qualitatives l'espace est l'ensemble des modalités possibles.

distances et dissimilarités

Rappel sur le partitionnement
Distances
Notations
distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

Soit Xi et Xj deux individus.
Une fonction $d : E \times E \to \mathbb{R}^+$ qui vérifie :

distances et dissimilarités

Rappel sur le partitionnement
Distances
Notations
distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

Soit Xi et Xj deux individus.
Une fonction $d : E \times E \to \mathbb{R}^+$ qui vérifie :

1. $d(X_i, X_j) \geq 0$
2. $d(X_i, X_j) = d(X_j, X_i)$
3. $d(X_i, X_j) = 0 \Leftrightarrow X_j = X_i$ dans notre cas, les individus ont les mêmes valeurs sur toutes les variables mesurées.
4. $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ quelque soit $X_k \in E$

est appelée une distance. Si d vérifie 1, 2 et 3, on parle de dissimilarité.

Distances usuelles : variables quantitatives

Rappel sur le partitionnement
Distances
Notations distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

On raisonne le plus souvent en distances au carré (cf. Lien avec l'inertie).
Pour des variables quantitatives l'espace dans lequel les variables évoluent est $E = \mathbb{R}$
Les distances les plus fréquemment utilisées s'écrivent sous la forme :

$$d^2(X_i, X_j) = (X_i - X_j)^t M (X_i - X_j)$$

- ✔ Euclidienne $M = I$: $d^2(X_i, X_j) = (\sum_{l=1}^{p} (X_{il} - X_{jl})^2)$
- ✔ Euclidienne sur variables centrées $M = diag(\frac{1}{\sigma_1} \cdots \frac{1}{\sigma_p})$
- ✔ Mahalanobis $M = V^{-1}$, ou V est la matrice de covariance
- ✔ ... Il y en a beaucoup d'autres (Minkovski)

Distances usuelles : variables qualitatives

Rappel sur le partitionnement
Distances
Notations distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

Cas de variables binaires :

- ✔ a_{ij}=nombre de caractères communs aux deux individus
- ✔ b_{ij}=nombre de caractères possédés par i mais pas par j
- ✔ c_{ij}=nombre de caractères possédés par j mais pas par i
- ✔ d_{ij}=nombre de caractères possédés ni par j ni par i
- ✔ avec $a_{ij} + b_{ij} + c_{ij} + d_{ij} = p$.

On construit des indices de similarité courants puis on prend le complément à 1 : Concordance $= \frac{a_{ij}+d_{ij}}{p}$, Jaccard$=\frac{a_{ij}}{a_{ij}+b_{ij}+c_{ij}}$,p...

Distance euclidienne du χ^2, $d^2_{\chi^2} = \frac{n}{p}\sum_{j=1}^{p}\sum_{l=1}^{m_j}\delta_{ik}^{jl}\frac{1}{n_l^j}$

avec δ_{ik}^{jl}=1 si ind. i et k présentent une discordance sur la l-ième modalité de X^p.

Pour préparer la suite

Rappel sur le partitionnement
Distances
Notations distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

Ces distances (ou dissimilarités) nous permettent de quantifier si deux points sont proches / similaires, ou éloignés / différents.
Il sera nécessaire par la suite de calculer deux autres types de "distances".

- ✔ Des distances entre groupes de points
  - ✘ Pour agréger des groupes : classification ascendante hiérarchique
  - ✘ Pour décomposer l'inertie (choix du nombre de classes)
- ✔ Des distances entre partitions-classifications (indice de Rand)

Formalisation du partitionnement

Rappel sur le partitionnement
Distances
Notations distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

$P = (C_1, \cdots, C_k, \cdots, C_K)$ une partition du jeu de données.
On cherche $P_K \in \mathcal{P}_\mathcal{K}$ la partition de l'ensemble de toutes les partitions possibles qui optimise un critère $W : \mathcal{P}_\mathcal{K} \to \mathcal{R}^+$
En pratique on cherche en général à minimiser l'inertie intra-classe ou, de manière strictement équivalente,

Formalisation du partitionnement

Rappel sur le partitionnement
Distances
Notations distances et dissimilarités
Distances usuelles : variables quantitatives
Distances usuelles : variables qualitatives
Pour préparer la suite
Formalisation du partitionnement
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus

$P = (C_1, \cdots, C_k, \cdots, C_K)$ une partition du jeu de données.
On cherche $P_K \in \mathcal{P}_\mathcal{K}$ la partition de l'ensemble de toutes les partitions possibles qui optimise un critère $W : \mathcal{P}_\mathcal{K} \to \mathcal{R}^+$
En pratique on cherche en général à minimiser l'inertie intra-classe ou, de manière strictement équivalente, à maximiser l'inertie inter-classes.

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

Centres mobiles

Principes

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

Différents algorithmes ont été définis autour du même principe de ré-allocation dynamique des individus à des centres de classes, eux-mêmes recalculés à chaque itération.

Algorithmes les plus courants :

- ✔ Moyennes mobiles, ou K-means.
- ✔ Une méthode similaire, plus robuste et implémentée dans R : Partitionnement autour de médioïdes.

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

1. Initialisation : tirages au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

1. Initialisation : tirages au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

1. Initialisation : tirages au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$
3. Recalcul : les barycentres (moyennes) des groupes sont recalculés et servent de centres à l'itération suivante.
   $\mu_g = \frac{1}{N_k} \sum_{i \in C_k} x_i$

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

1. Initialisation : tirages au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$
3. Recalcul : les barycentres (moyennes) des groupes sont recalculés et servent de centres à l'itération suivante.
   $\mu_g = \frac{1}{N_k} \sum_{i \in C_k} x_i$

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

1. Initialisation : tirages au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$
3. Recalcul : les barycentres (moyennes) des groupes sont recalculés et servent de centres à l'itération suivante.
   $\mu_g = \frac{1}{N_k} \sum_{i \in C_k} x_i$

Ça ne vous rappelle rien ?

Algorithme des Kmeans

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

On mesure p variables sur n individus. $X_i \in E$, typiquement $E = \mathcal{R}^p, i = \{1, \cdots, n\}$ muni d'une distance.

1. Initialisation : tirages au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$
3. Recalcul : les barycentres (moyennes) des groupes sont recalculés et servent de centres à l'itération suivante.
   $\mu_g = \frac{1}{N_k} \sum_{i \in C_k} x_i$

Ça ne vous rappelle rien ? Analogie avec algorithme EM.

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

Convergence

Deux façons intuitives de voir la convergence.

"Les centres ne bougent plus" : $||\mu_k|| < \epsilon$

Convergence

Rappel sur le partitionnement
Distances
Centres mobiles
Principes
Algorithme des Kmeans
Convergence
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées

Deux façons intuitives de voir la convergence.

"Les centres ne bougent plus" : $||\mu_k|| < \epsilon$
"L'inertie interclasse ne croît plus" $\equiv$ "L'inertie intra classe ne décroît plus"

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Décomposition de l'inertie du nuage de points

Inertie Totale

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie totale T du nuage de points s'écrit : $T = \sum_{i=1}^{n} d^2(x_i, g)$ où g est le centre d'inertie du nuage (le barycentre).

Inertie Totale

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie totale T du nuage de points s'écrit : $T = \sum_{i=1}^{n} d^2(x_i, g)$ où g est le centre d'inertie du nuage (le barycentre).
T ne dépend pas des classes.

Inertie Totale

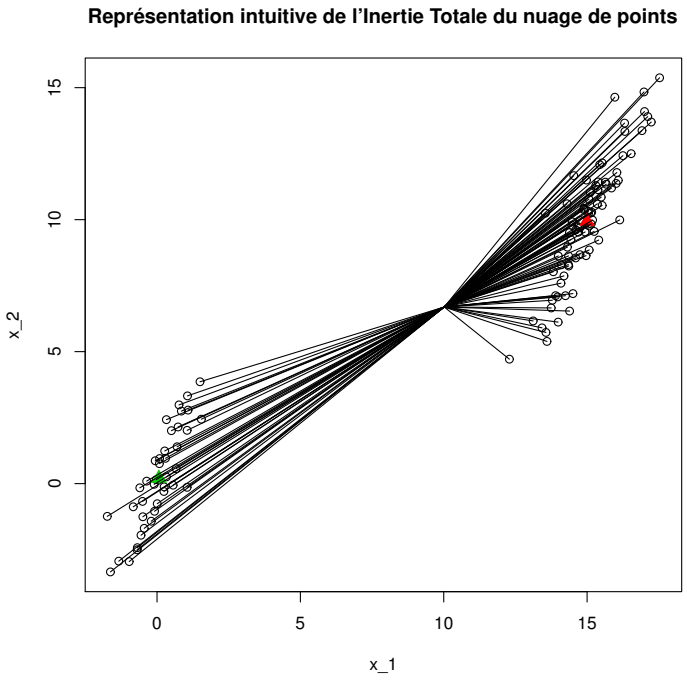
Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie totale T du nuage de points s'écrit : $T = \sum_{i=1}^{n} d^2(x_i, g)$ où g est le centre d'inertie du nuage (le barycentre).
T ne dépend pas des classes.
Si on considère que tous les points ont le même poids, alors g a pour coordonnées les moyennes de chaque variable.

Inertie Totale

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie totale T du nuage de points s'écrit : $T = \sum_{i=1}^{n} d^2(x_i, g)$ où g est le centre d'inertie du nuage (le barycentre).
T ne dépend pas des classes.
Si on considère que tous les points ont le même poids, alors g a pour coordonnées les moyennes de chaque variable.



Inertie Inter classes

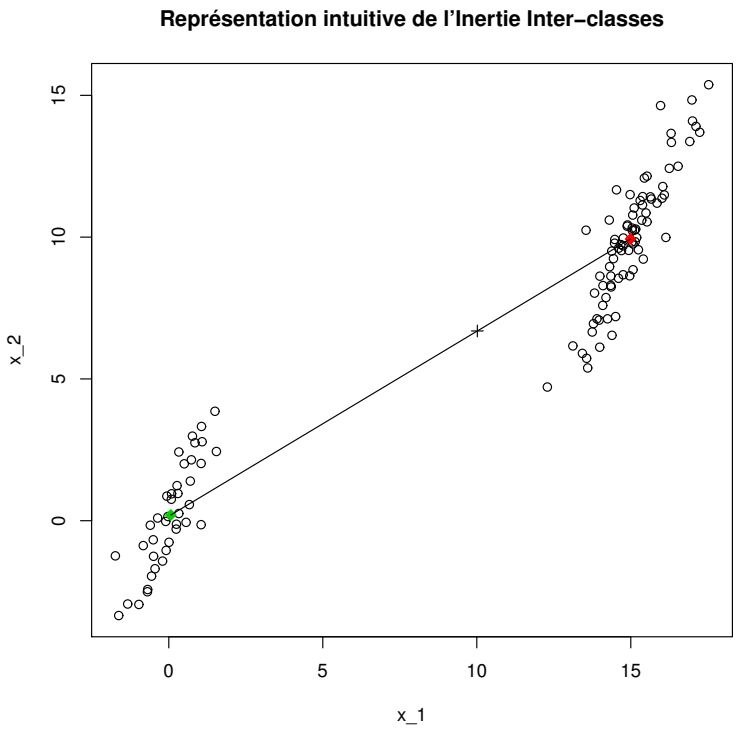
Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie inter-classes I s'écrit : $B = \sum_{k=1}^{K} d^2(g_k, g)$ ou g_k est le centre d'inertie de la classe k.

Inertie Inter classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie inter-classes I s'écrit : $B = \sum_{k=1}^{K} d^2(g_k, g)$ ou g_k est le centre d'inertie de la classe k.



Inerties Intra classe

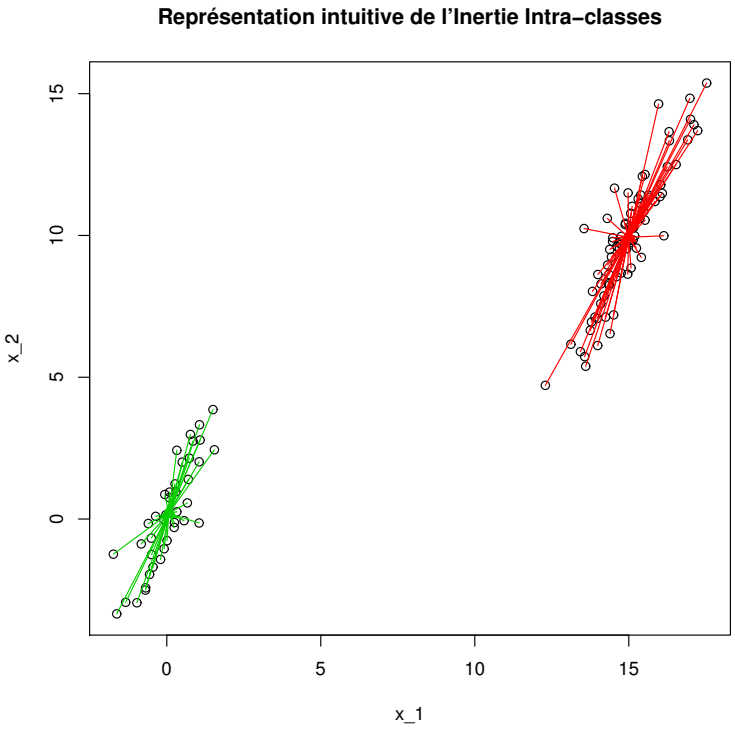
Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie intra-classe s'écrit : $W = \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(x_i, g_k)$

Inerties Intra classe

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'inertie intra-classe s'écrit : $W = \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(x_i, g_k)$



Décomposition de l'inertie d'un nuage de points

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

On a la relation fondamentale :

$$T = B + W$$

$$\sum_{i=1}^{n} d^2(x_i, g) = \sum_{k=1}^{K} d^2(g_k, g) + \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(x_i, g_k)$$

(qui se déduit du Théorème de Huygens).
Maximiser B revient à minimiser W.

Un indicateur de qualité d'une partition

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Le rapport entre l'inertie intra et l'inertie totale est un indicateur de qualité d'une partition

$$0 \leq \frac{B}{T} \leq 1$$

$\frac{B}{T} = 0 \Rightarrow \forall k, \bar{x}_{x \in C_k} = \bar{x}$: les centres de gravité de toutes les classes sont confondus

Un indicateur de qualité d'une partition

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Le rapport entre l'inertie intra et l'inertie totale est un indicateur de qualité d'une partition

$$0 \leq \frac{B}{T} \leq 1$$

$\frac{B}{T} = 0 \Rightarrow \forall k, \bar{x}_{x \in C_k} = \bar{x}$: les centres de gravité de toutes les classes sont confondus : impossible de classifier avec une méthode géométrique (éventuellement, penser aux modèles de mélange)

Un indicateur de qualité d'une partition

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Le rapport entre l'inertie intra et l'inertie totale est un indicateur de qualité d'une partition

$$0 \leq \frac{B}{T} \leq 1$$

$\frac{B}{T} = 0 \Rightarrow \forall k, \bar{x}_{x \in C_k} = \bar{x}$: les centres de gravité de toutes les classes sont confondus : impossible de classifier avec une méthode géométrique (éventuellement, penser aux modèles de mélange)

$\frac{B}{T} = 1 \Rightarrow \forall k, \forall x \in C_k, x_i = \bar{x}_k$: tous les individus d'une classe sont identiques, un individu par classe. Parfait en termes de critère, mais totalement inutile d'un point de vue pratique.

Un indicateur de qualité d'une partition

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Le rapport entre l'inertie intra et l'inertie totale est un indicateur de qualité d'une partition

$$0 \leq \frac{B}{T} \leq 1$$

$\frac{B}{T} = 0 \Rightarrow \forall k, \bar{x}_{x \in C_k} = \bar{x}$: les centres de gravité de toutes les classes sont confondus : impossible de classifier avec une méthode géométrique (éventuellement, penser aux modèles de mélange)

$\frac{B}{T} = 1 \Rightarrow \forall k, \forall x \in C_k, x_i = \bar{x}_k$: tous les individus d'une classe sont identiques, un individu par classe. Parfait en termes de critère, mais totalement inutile d'un point de vue pratique.

✔ Pb : B/T dépend du nombre de classes, du nombre d'individus, et sans modèle sous-jacent, pas possible d'entrer dans une logique de "tables" reposant sur une distribution de B/W

Utilisation de l'inertie comme critère de convergence

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Rappel de l'algorithme des K-Means :

1. Initialisation : tirage au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$
3. Recalcul : les barycentres (moyennes) des groupes sont recalculés et servent de centres à l'itération suivante.
   $\mu_g = \frac{1}{N_k} \sum_{i \in C_k} x_i$

Utilisation de l'inertie comme critère de convergence

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Rappel de l'algorithme des K-Means :

1. Initialisation : tirage au hasard de K centre μ_k parmi les n observations.
   Tant que la partition n'est pas stabilisée :
2. Affectation : chaque observation est affectée au centre le plus proche
   $C(x_i) = \min_k d(x_i, \mu_k)$
3. Recalcul : les barycentres (moyennes) des groupes sont recalculés et servent de centres à l'itération suivante.
   $\mu_g = \frac{1}{N_k} \sum_{i \in C_k} x_i$
4. Condition de convergence $\Delta_{(r-1;r)}(W) \leqslant ?\epsilon$
   Où $\Delta_{(r;r+1)}(W)$ est la variation de W entre l'itération r et l'itération r.

Propriétés de convergences K-means

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences K-means
K-means encore un paramètre
exemple de choix

- ✔ On peut montrer que l'algorithme K-means converge
- ✔ La solution obtenue peut dépendre de l'initialisation (maxima locaux)

Propriétés de convergences K-means

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences K-means
K-means encore un paramètre
exemple de choix

- ✔ On peut montrer que l'algorithme K-means converge
- ✔ La solution obtenue peut dépendre de l'initialisation (maxima locaux)
- ✔ En pratique : on fait tourner plusieurs fois l'algorithme avec des initialisations différentes et on conserve la partition dont l'inertie intra W est minimale (T ne dépendant pas du partitionnement, minimiser W/T revient à minimiser W).

K-means encore un paramètre

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

La minimisation de W se fait...

K-means encore un paramètre

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

La minimisation de W se fait... à K constant !

K-means encore un paramètre

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

La minimisation de W se fait... à K constant !

✔ Problème : $W(P(K))$ décroit de façon monotone en K. A l'extrême, un individu par classe, inertie Intra W=0 !

K-means encore un paramètre

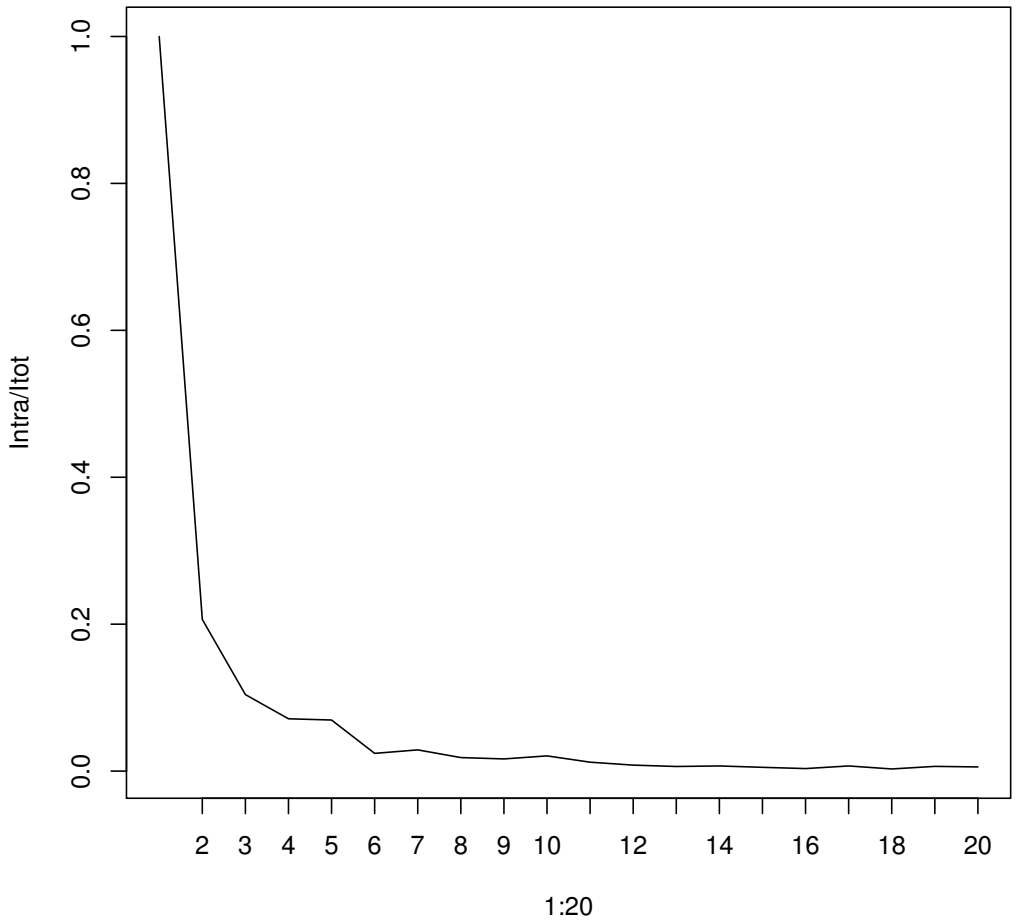
Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

La minimisation de W se fait... à K constant !

- ✔ Problème : $W(P(K))$ décroit de façon monotone en K. A l'extrême, un individu par classe, inertie Intra W=0 !
- ✔ Choix de K : on peut répéter l'algorithme en faisant varier K et repérer une décroissance plus forte de W (coude)

exemple de choix de classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix de classes



Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Quelle est la "quantité de calculs" à effectuer ?

Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Quelle est la "quantité de calculs" à effectuer ?

Combien de distances à calculer ?

Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Quelle est la "quantité de calculs" à effectuer ?

Combien de distances à calculer ?

$K \times n$ distances à calculer

Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Quelle est la "quantité de calculs" à effectuer ?

Combien de distances à calculer ?

$K \times n$ distances à calculer en dimension p

Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Quelle est la "quantité de calculs" à effectuer ?

Combien de distances à calculer ?

$K \times n$ distances à calculer en dimension p à chaque itération.

$K \times n \times p \times I$ calculs.

Un petit pas vers la data-science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

L'algorithme des K-means est-il résistant à la taille du jeu de données ?

Quelle est la "quantité de calculs" à effectuer ?

Combien de distances à calculer ?

$K \times n$ distances à calculer en dimension p à chaque itération.

$K \times n \times p \times I$ calculs.

À comparer avec la classification hiérarchique...

Limites de K-means

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Rechercher les points qui sont les plus proches d'un centre à partir de la distance euclidienne $\equiv$ rechercher des "sphères".

Limites de K-means

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Rechercher les points qui sont les plus proches d'un centre à partir de la distance euclidienne $\equiv$ rechercher des "sphères". L'absence de modèle ne permet pas de prendre en compte la forme éventuelle des clusters.

Limites de K-means

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Inertie Totale
Inertie Inter classes
Inerties Intra classe
Décomposition de l'inertie d'un nuage de points
Un indicateur de qualité d'une partition
Utilisation de l'inertie comme critère de convergence
Propriétés de convergences
K-means
K-means encore un paramètre
exemple de choix

Rechercher les points qui sont les plus proches d'un centre à partir de la distance euclidienne $\equiv$ rechercher des "sphères". L'absence de modèle ne permet pas de prendre en compte la forme éventuelle des clusters.

Les itérations de l'algorithme de K-means font converger l'inertie intra, mais éventuellement vers un minimum local+initialisation aléatoire $\Rightarrow$ plusieurs initialisations et meilleure partition.

Le calcul de la moyenne peut pénaliser la robustesse aux valeurs atypiques. On peut utiliser des centres mobiles basés sur la médiane (médioids, algorithme pam dans R)

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Classification ascendante hiérarchique

Algorithme général de la CAH

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

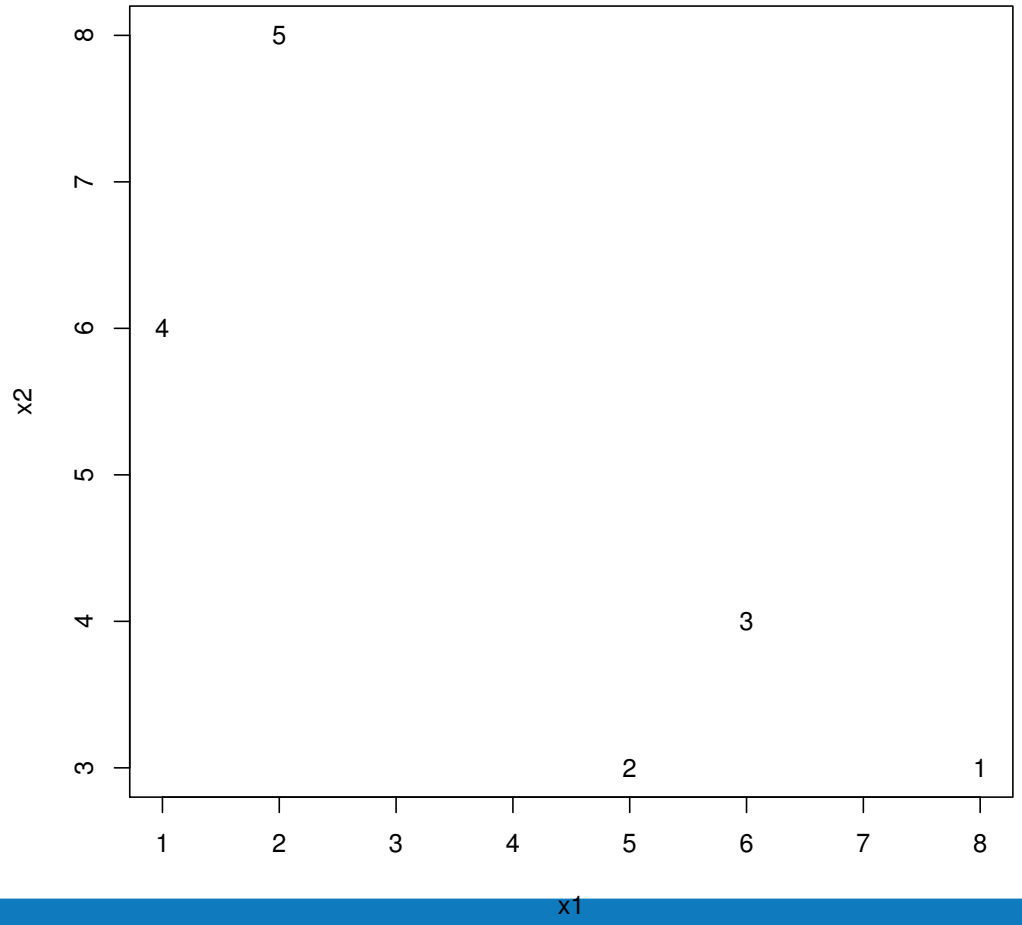
Une fois qu'on a défini une distance d^2 entre deux classes.

1. Initialiser : chaque individu (observation) constitue une classe
   Tant qu'il y a plus d'une classe
2. Calculer les "distances" entre toutes les paires de classes
3. Regrouper les classes les plus proches

Un exemple très simple

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Deux dimensions, 5 points.



Distances entre classes : critère d'agrégation

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La distance entre deux classes est le critère d'agrégation à l'étape suivante.
Il en existe plusieurs, dont les propriétés sont différentes.
Les plus fréquemment utilisés sont :

- ✔ Le saut minimum, ou single linkage
- ✔ Le saut maximum, ou complete linkage
- ✔ La distance moyenne
- ✔ la distance de Ward

Single Linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Le single linkage est la distance minimale entre deux points n'appartenant pas au même groupe. La distance entre voisins les plus proches.

$$d(C_1, C_2) = min(d(xi, x_{i'}))$$

avec $x_i \in C_1, x_{i'} \in C_2$.

Exemple avec $C_1 = (1, 2, 3)$:

	1	2	3	4	5
1					
2	3				
3	2.24	1.41			
4	7.61	5.00	5.39		
5	7.81	5.83	5.66	2.24	

Single Linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Le single linkage est la distance minimale entre deux points n'appartenant pas au même groupe. La distance entre voisins les plus proches.

$$d(C_1, C_2) = min(d(xi, x_{i'}))$$

avec $x_i \in C_1, x_{i'} \in C_2$.

Exemple avec $C_1 = (1, 2, 3)$:

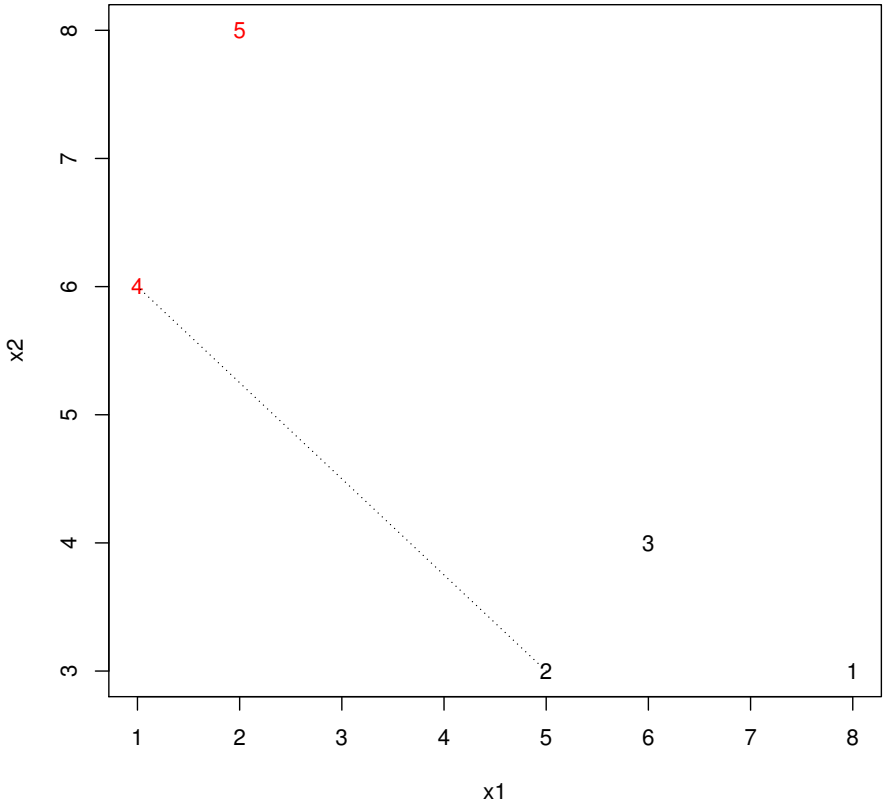
	1	2	3	4	5
1					
2	3				
3	2.24	1.41			
4	7.61	5.00	5.39		
5	7.81	5.83	5.66	2.24	

$d(C_1, C_2) = d(i_4, i_2) = 5$

Single Linkage sur notre exemple

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

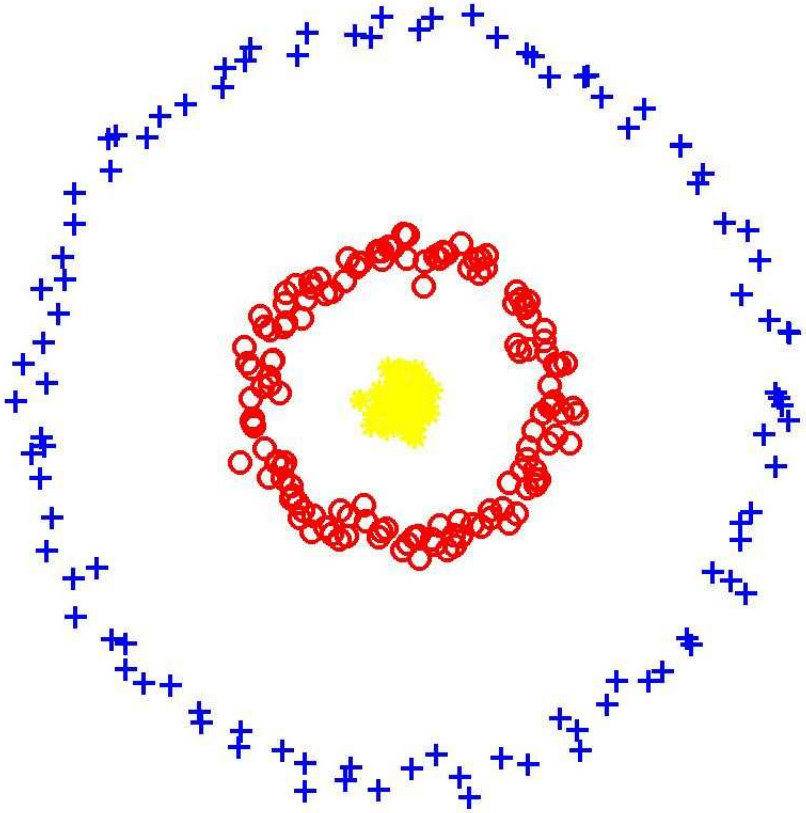
Sur notre exemple :



Bon usage du single linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Si on soupçonne un effet de "chaînage", le single linkage est susceptible de le capturer.



Complete Linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Le complete linkage est la distance maximale entre deux points n'appartenant pas au même groupe. Le complete linkage est la distance du voisin le plus éloigné.

Complete Linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Le complete linkage est la distance maximale entre deux points n'appartenant pas au même groupe. Le complete linkage est la distance du voisin le plus éloigné.

$$d_{cmplt}(C_1, C_2) = max(d(xi, x_{i'}))$$

avec $x_i \in C_1, x_{i'} \in C_2$.

Exemple :

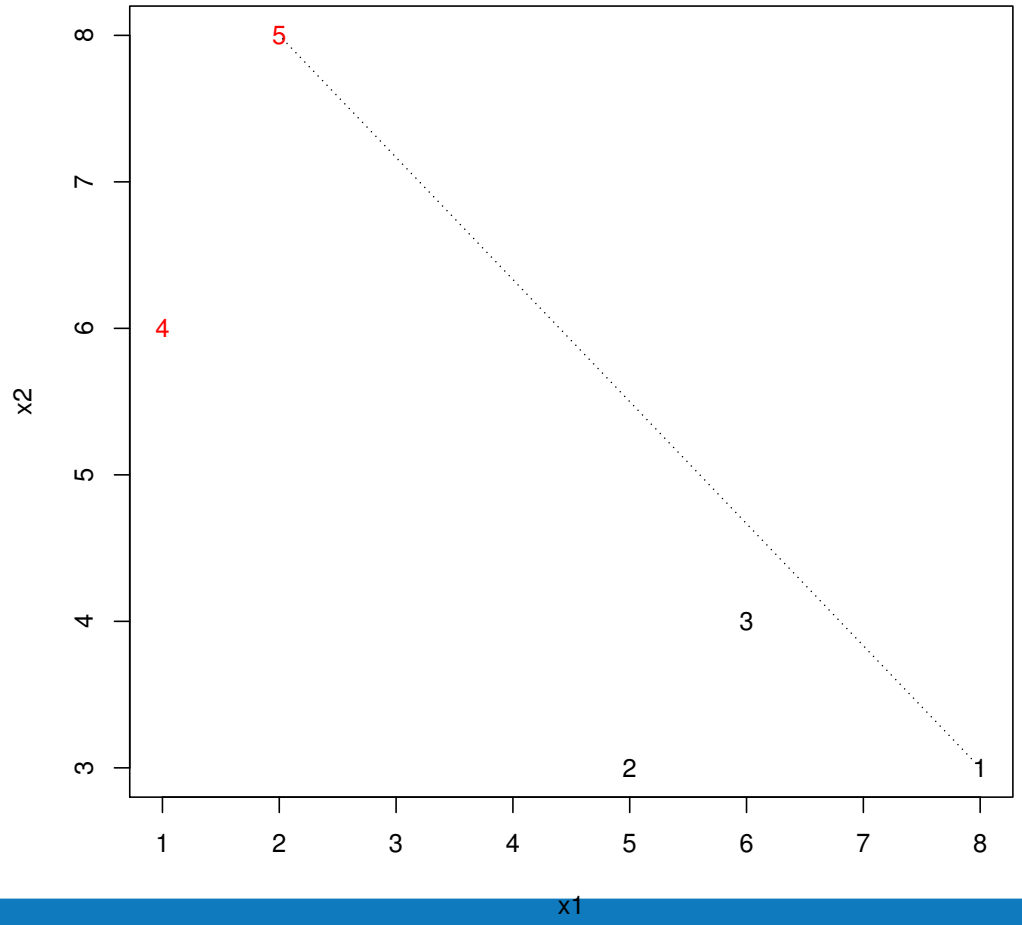
	1	2	3	4	5
1					
2	3				
3	2.24	1.41			
4	7.61	5.00	5.39		
5	7.81	5.83	5.66	2.24	

$d_{max}(C_1, C_2) = d(i_1, i_5) = 7.81$

Complete Linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Sur notre exemple :



Au delà du complete linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La distance du voisin le plus éloigné n'entraîne pas d'effet de chaînage, et mène en général à des groupes plutôt compact.

Complete/single linkage se basent sur un seul individu par groupe $\rightarrow$ sensibles aux points atypiques.

Au delà du complete linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La distance du voisin le plus éloigné n'entraîne pas d'effet de chaînage, et mène en général à des groupes plutôt compact.

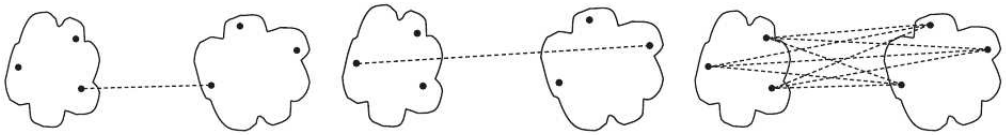
Complete/single linkage se basent sur un seul individu par groupe $\rightarrow$ sensibles aux points atypiques. Comment prendre en compte tous points des groupes ?

Au delà du complete linkage

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La distance du voisin le plus éloigné n'entraîne pas d'effet de chaînage, et mène en général à des groupes plutôt compact.

Complete/single linkage se basent sur un seul individu par groupe → sensibles aux points atypiques. Comment prendre en compte tous points des groupes ?



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Figure 8.14. Graph-based definitions of cluster proximity

Figure 1: Emprunté à[6]

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire la moyenne des distances deux à deux entre individus des deux groupes candidats à la fusion.

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire la moyenne des distances deux à deux entre individus des deux groupes candidats à la fusion.
- ✔ Pour éviter l'influence d'individus atypiques, on peut également avoir recours à

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire la moyenne des distances deux à deux entre individus des deux groupes candidats à la fusion.
- ✔ Pour éviter l'influence d'individus atypiques, on peut également avoir recours à la médiane des distances deux à deux entre individus de groupes candidats à la fusion.

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire la moyenne des distances deux à deux entre individus des deux groupes candidats à la fusion.
- ✔ Pour éviter l'influence d'individus atypiques, on peut également avoir recours à la médiane des distances deux à deux entre individus de groupes candidats à la fusion.
- ✔ La distance dite centroïde prend comme distance

Différentes façons de moyenner des distances

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire la moyenne des distances deux à deux entre individus des deux groupes candidats à la fusion.
- ✔ Pour éviter l'influence d'individus atypiques, on peut également avoir recours à la médiane des distances deux à deux entre individus de groupes candidats à la fusion.
- ✔ La distance dite centroïde prend comme distance la distance entre les points moyens des groupes candidats à la fusion.

Différentes façons de moyenner des distances

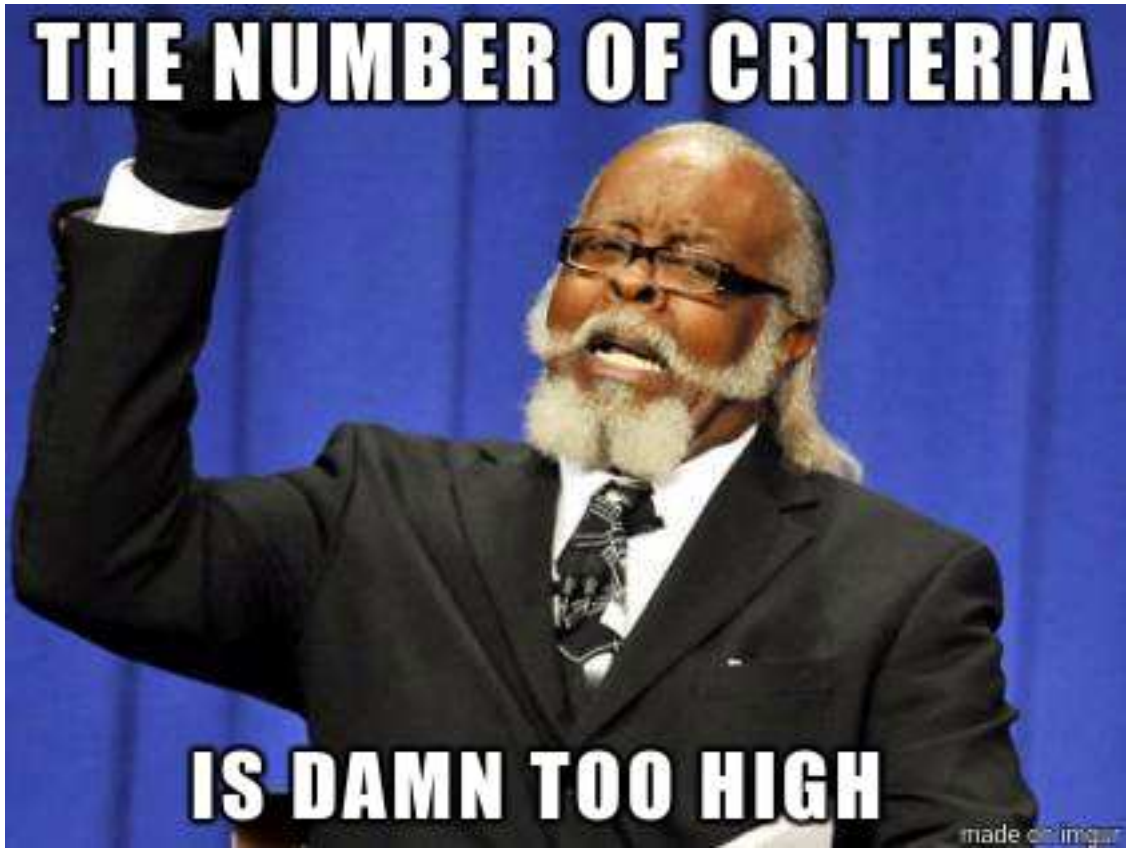
Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage
Différentes façons de moyenner des distances

Si on prend en compte tous les points des classes candidates à la fusion, on peut intuitivement proposer plusieurs façon de moyenner les distances :

- ✔ Comme son nom l'indique, le critère d'average linkage revient à faire la moyenne des distances deux à deux entre individus des deux groupes candidats à la fusion.
- ✔ Pour éviter l'influence d'individus atypiques, on peut également avoir recours à la médiane des distances deux à deux entre individus de groupes candidats à la fusion.
- ✔ La distance dite centroïde prend comme distance la distance entre les points moyens des groupes candidats à la fusion.
- ✔ Enfin, on peut prendre en compte les effectifs des groupes candidats en pondérant par le effectifs des classes.

Trop de critères !

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage



Une ébauche d'approche modèle : la distance de Ward

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

A l'étape r, on dispose de $k_(r)$ classes. On cherche à savoir quelles sont les deux classes qu'on doit fusionner.
Clustering $\equiv$ expliquer l'inertie totale du nuage de points à l'aide d'une variable classe (qu'on ne connaît pas).

Une ébauche d'approche modèle : la distance de Ward

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

A l'étape r, on dispose de $k_(r)$ classes. On cherche à savoir quelles sont les deux classes qu'on doit fusionner.
Clustering $\equiv$ expliquer l'inertie totale du nuage de points à l'aide d'une variable classe (qu'on ne connaît pas).

Inertie : variable quantitative
Classe : variable qualitative

Une ébauche d'approche modèle : la distance de Ward

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

A l'étape r, on dispose de $k_(r)$ classes. On cherche à savoir quelles sont les deux classes qu'on doit fusionner.
Clustering $\equiv$ expliquer l'inertie totale du nuage de points à l'aide d'une variable classe (qu'on ne connaît pas).

Inertie : variable quantitative
Classe : variable qualitative

Quel modèle permet d'expliquer une variable quantitative à l'aide d'une variable qualitative ?

Une ébauche d'approche modèle : la distance de Ward

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

A l'étape r, on dispose de $k_(r)$ classes. On cherche à savoir quelles sont les deux classes qu'on doit fusionner.
Clustering $\equiv$ expliquer l'inertie totale du nuage de points à l'aide d'une variable classe (qu'on ne connaît pas).

Inertie : variable quantitative
Classe : variable qualitative

Quel modèle permet d'expliquer une variable quantitative à l'aide d'une variable qualitative ?

Analogie avec l'analyse de la variance.

Décomposition de l'inertie

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

L'inertie totale du nuage de points, T, peut se décomposer en

- ✔ La part d'inertie expliquée par les classes, inter-classes ou between B
- ✔ La part d'inertie résiduelle, à l'intérieur des classes, intra-classes ou withins W

$$T = B + W$$

Si on met un individu par classe, on peut avoir l'impression que toute l'inertie est expliquée, en pratique absurde.

Si on met tous les individus dans la même classe, il n'y a plus que de la résiduelle.

À chaque étape, il semblerait donc logique de choisir la fusion de classe qui dégrade le moins...

Décomposition de l'inertie

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

L'inertie totale du nuage de points, T, peut se décomposer en

- ✔ La part d'inertie expliquée par les classes, inter-classes ou between B
- ✔ La part d'inertie résiduelle, à l'intérieur des classes, intra-classes ou withins W

$$T = B + W$$

Si on met un individu par classe, on peut avoir l'impression que toute l'inertie est expliquée, en pratique absurde.

Si on met tous les individus dans la même classe, il n'y a plus que de la résiduelle.

À chaque étape, il semblerait donc logique de choisir la fusion de classe qui dégrade le moins... la partie "expliquée" de notre inertie, donc...

Décomposition de l'inertie

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

L'inertie totale du nuage de points, T, peut se décomposer en

- ✔ La part d'inertie expliquée par les classes, inter-classes ou between B
- ✔ La part d'inertie résiduelle, à l'intérieur des classes, intra-classes ou withins W

$$T = B + W$$

Si on met un individu par classe, on peut avoir l'impression que toute l'inertie est expliquée, en pratique absurde.

Si on met tous les individus dans la même classe, il n'y a plus que de la résiduelle.

À chaque étape, il semblerait donc logique de choisir la fusion de classe qui dégrade le moins... la partie "expliquée" de notre inertie, donc... l'inertie inter.

distance de Ward

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La distance qui permet de minimiser la perte d'inertie inter lors de la fusion de deux classes est appelée distance de Ward.
Il est possible de montrer qu'elle s'écrit :

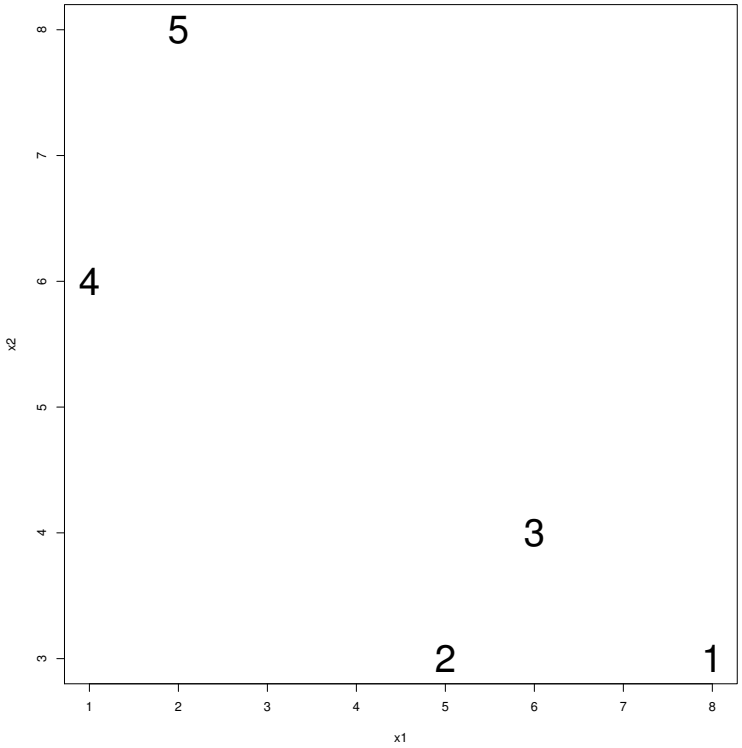
$$d^2(C_1, C_2) = \frac{\sharp C_1 \sharp C_2}{\sharp C_1 + \sharp C_2} \times d^2(\mu_1, \mu_2)$$

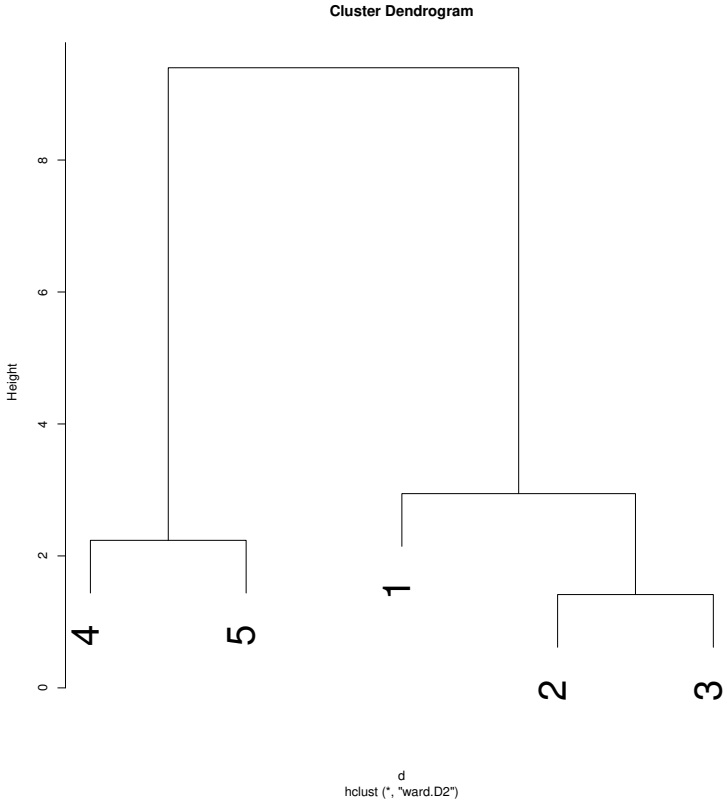
où $\sharp C_k$ désigne le cardinal de la classe k, et μ_k est le centre de gravité de la classe k.

Cette perte d'inertie inter est utilisée pour interpréter le **dendrogramme** obtenu à l'issu de la CAH.

Grimpons aux arbres

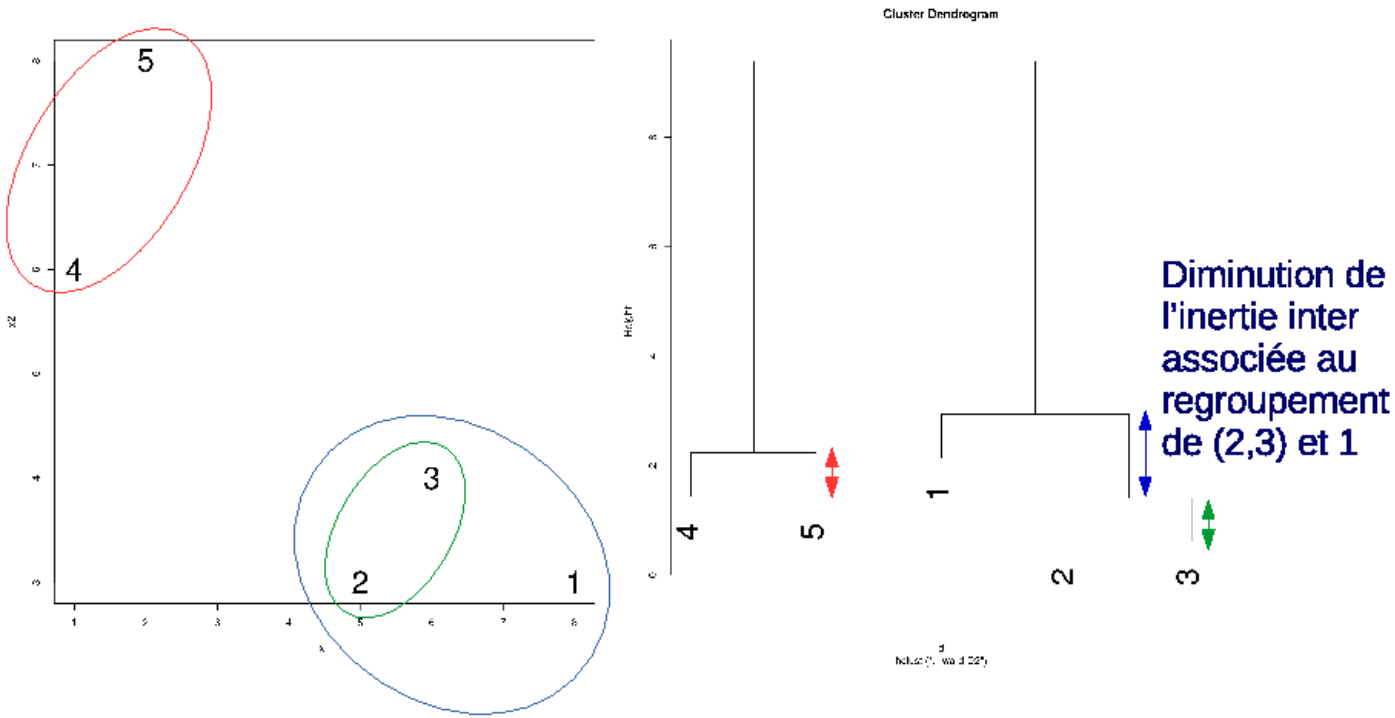
Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage





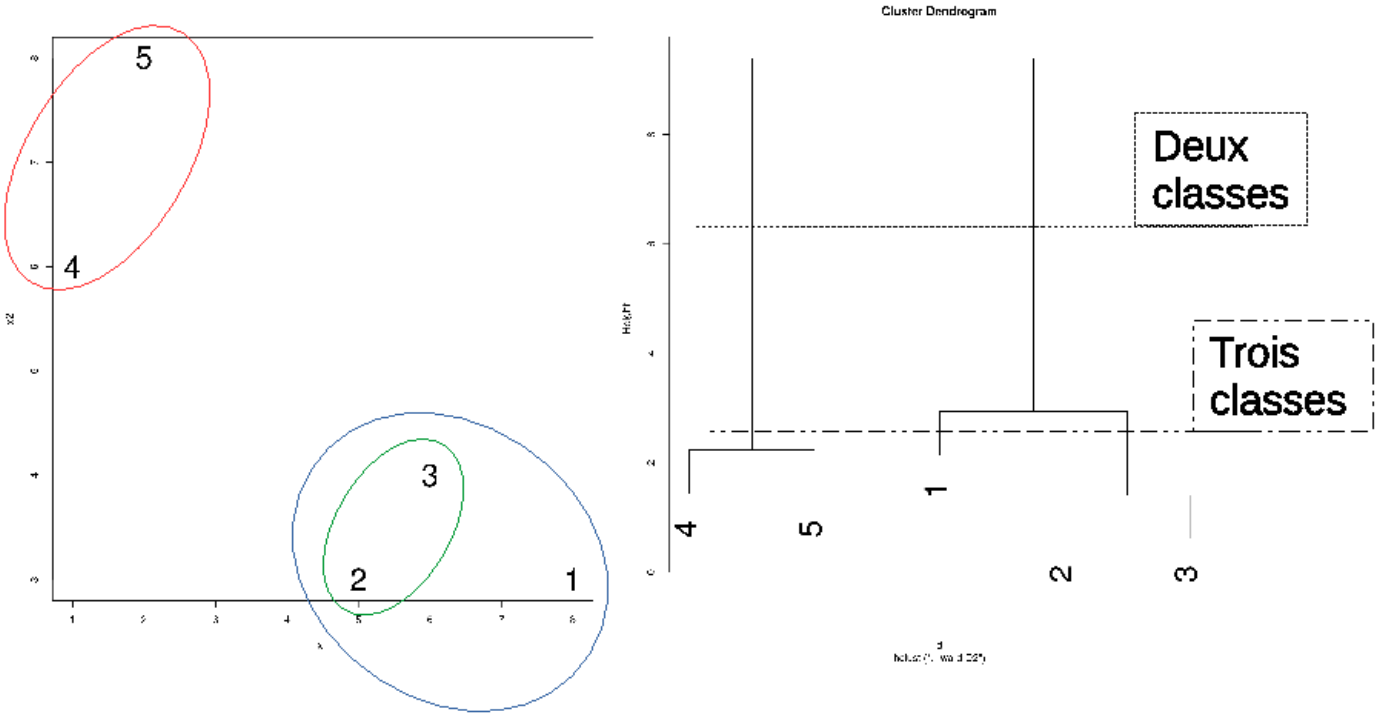
Grimpons aux arbres

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage



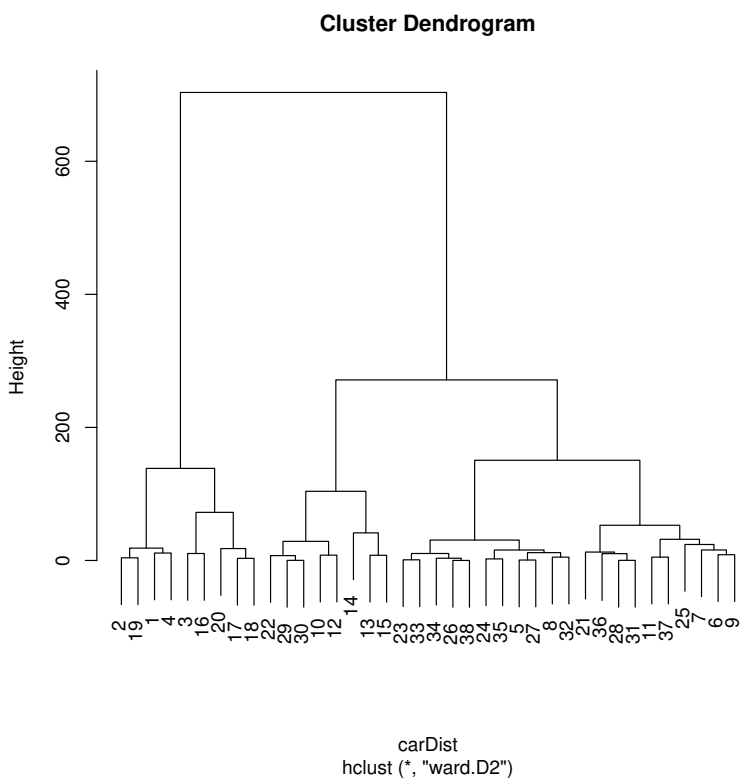
Pour mieux les couper

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage



Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

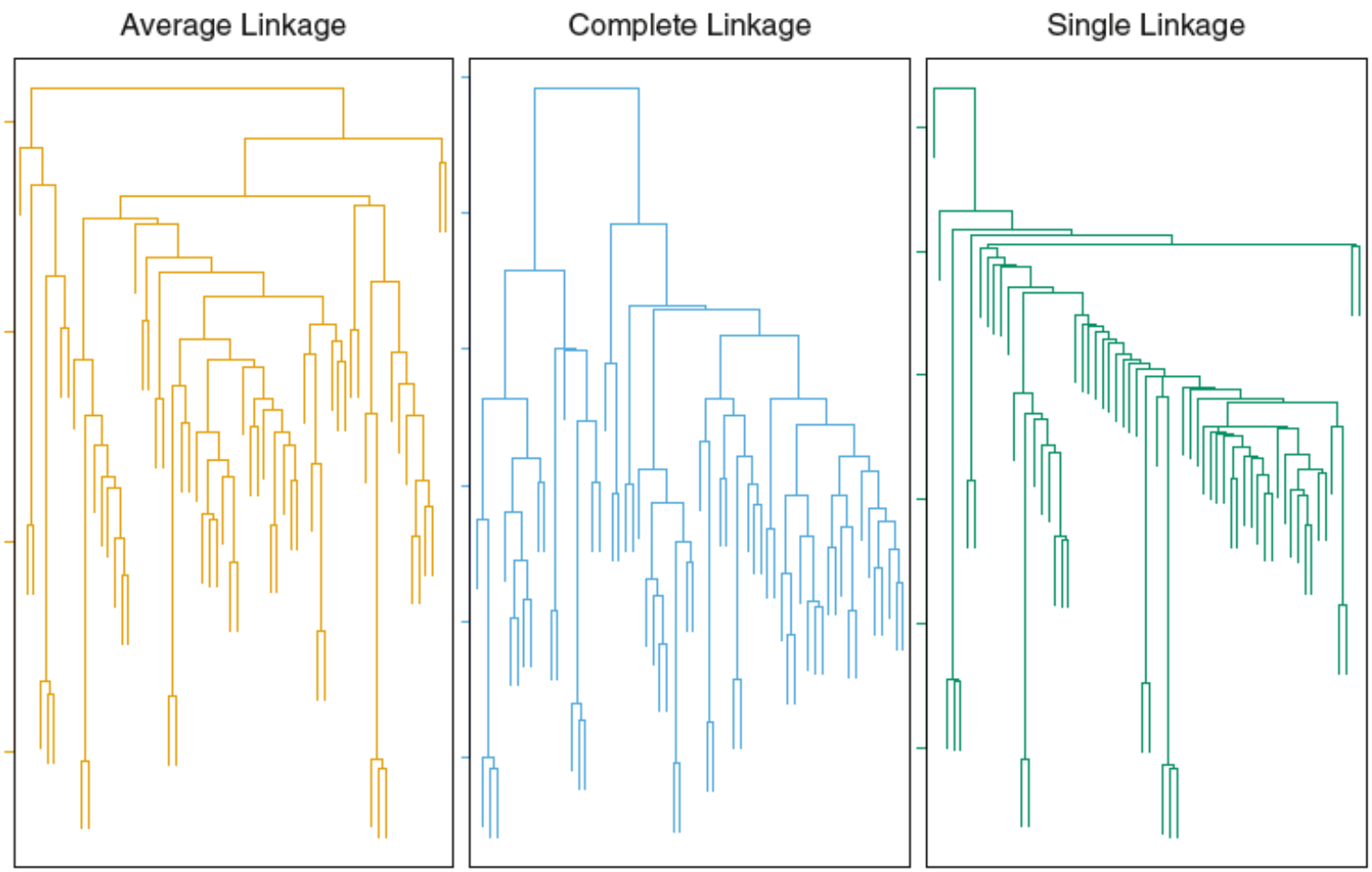
Précautions dans la lecture d'un arbre



- ✔ Au sein d'une classe on peut permuter les branches : la 'distance' entre deux classes se lit donc verticalement et non horizontalement
- ✔ Toujours regarder au-delà de la solution à deux classes

Exemples d'impact du critère d'agrégation

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage



Un petit pas vers la data science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La classification hiérarchique est elle résistante à la taille du jeu de données ?

Un petit pas vers la data science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La classification hiérarchique est elle résistante à la taille du jeu de données ?
Quelle quantité de calculs à effectuer ?

Un petit pas vers la data science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La classification hiérarchique est elle résistante à la taille du jeu de données ?
Quelle quantité de calculs à effectuer ? Combien de distances à calculer ?

Un petit pas vers la data science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La classification hiérarchique est elle résistante à la taille du jeu de données ?
Quelle quantité de calculs à effectuer ? Combien de distances à calculer ?

$n \times n$ distances à calculer dès le premier pas !

Un petit pas vers la data science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La classification hiérarchique est elle résistante à la taille du jeu de données ?
Quelle quantité de calculs à effectuer ? Combien de distances à calculer ?

$n \times n$ distances à calculer dès le premier pas ! Pour passer de $k+1$ classes à k classes : $\frac{(k+1)\times k}{2}$ critères d'agrégation : complexité de l'algorithme de l'ordre de n^3

Un petit pas vers la data science et le big data ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

La classification hiérarchique est elle résistante à la taille du jeu de données ?
Quelle quantité de calculs à effectuer ? Combien de distances à calculer ?

$n \times n$ distances à calculer dès le premier pas ! Pour passer de $k+1$ classes à k classes : $\frac{(k+1)\times k}{2}$ critères d'agrégation : complexité de l'algorithme de l'ordre de n^3

A comparer avec les seulement $n \times K$ distances à calculer à chaque itération avec les K-Means.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Avantages

Récapitulatif CAH

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Avantages

- ✔ Pas d'arbitraire dans l'initialisation (contrairement à K-means)
- ✔ Le dendrogramme aide à choisir le nombre de classes
- ✔ Le dendrogramme apporte une vision "génétique" de la classification
- ✔ Le choix du critère d'agrégation permet de construire des classes de forme différentes

Récapitulatif CAH

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
Algorithme général de la CAH
Un exemple très simple
Distances entre classes : critère d'agrégation
Single Linkage
Single Linkage sur notre exemple
Bon usage du single linkage
Complete Linkage
Complete Linkage
Au delà du complete linkage

Avantages

- ✔ Pas d'arbitraire dans l'initialisation (contrairement à K-means)
- ✔ Le dendrogramme aide à choisir le nombre de classes
- ✔ Le dendrogramme apporte une vision "génétique" de la classification
- ✔ Le choix du critère d'agrégation permet de construire des classes de forme différentes

Inconvénients

- ✔ Le nombre de calculs nécessaires (complexité) est de l'odrdre de n^3, pas scalable
- ✔ Pas de modèle sous jacent
- ✔ À chaque étape le regroupement dépend des regroupements précédents : pas d'optimisation globale.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

De la classe à la typologie

Quelle est la question ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ Rappel de l'hypothèse principale sous-jacente : il existe des sous populations dans notre population
- ✔ Quels sont les sous-groupes les plus pertinents au sein de ma population d'étude ?
- ✔ Comment mes groupes se distinguent les uns des autres ?

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement : par exemple B/T

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement : par exemple B/T
  - ✘ On peut caractériser la compacité d'une classe par rapport à une autre :

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement : par exemple B/T
  - ✘ On peut caractériser la compacité d'une classe par rapport à une autre : comparer les W

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement : par exemple B/T
  - ✘ On peut caractériser la compacité d'une classe par rapport à une autre : comparer les W
- ✔ D'un point de vue pratique :

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement : par exemple B/T
  - ✘ On peut caractériser la compacité d'une classe par rapport à une autre : comparer les W
- ✔ D'un point de vue pratique :
  - ✘ Un indicateur de centralité par classe : **moyennes conditionnelles aux classes (médiane, mode pour une quali) et pour chaque variable mesurée**

Caractériser les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ D'un point de vue statistique :
  - ✘ On peut caractériser le partitionnement : par exemple B/T
  - ✘ On peut caractériser la compacité d'une classe par rapport à une autre : comparer les W
- ✔ D'un point de vue pratique :
  - ✘ Un indicateur de centralité par classe : **moyennes conditionnelles aux classes (médiane, mode pour une quali) et pour chaque variable mesurée**
  - ✘ Autres indicateurs éventuels (dispersion, effectif)

	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol
G 1	331.11765	262.7941	189.40000	27.15294	60.09412	9.711765	21.37647	82.35294
G 2	389.75000	126.0000	280.67500	30.65000	64.27500	2.525000	30.22500	107.50000
G 3	276.50000	235.5000	127.20000	22.82500	115.00000	34.050000	17.95000	65.00000
G 4	101.75000	44.7500	133.75000	6.27500	55.15000	16.475000	7.20000	18.25000
% epl.	87.97373	56.6772	41.27705	86.85973	64.89488	63.494807	82.70802	82.46284

De la classe au segment, au type etc.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

Le travail du data-miner ne s'arrête pas à l'application d'un algorithme ou à l'estimation d'un modèle.

De la classe au segment, au type etc.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

Le travail du data-miner ne s'arrête pas à l'application d'un algorithme ou à l'estimation d'un modèle.

Il faut pouvoir apporter des éléments de réponse ou de décision aux spécialistes métiers.

De la classe au segment, au type etc.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

Le travail du data-miner ne s'arrête pas à l'application d'un algorithme ou à l'estimation d'un modèle.

Il faut pouvoir apporter des éléments de réponse ou de décision aux spécialistes métiers.

Pour cela, il faut s'assurer que la classification puisse être comprise et appropriée.

Exemples

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

- ✔ Population d'étude : les clients d'une entreprise
  - ✘ Passer d'une classification à une segmentation
  - ✘ Segment opérationnel : sur lequel on va pouvoir faire une opération (marketing)
- ✔ Population d'étude : patients
  - ✘ Passer d'une classification à une typologie
  - ✘ Type de patient opérationnel : pour lequel on va pouvoir définir une stratégie thérapeutique

Nommer les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

"Mal nommer les choses c'est ajouter au malheur du monde."

A. Camus

Nommer les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

"Mal nommer les choses c'est ajouter au malheur du monde."

A. Camus

Donner un nom à chacune des classes trouvées permet de vérifier si on a réellement caractérisé des sous-populations.

Nommer les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

"Mal nommer les choses c'est ajouter au malheur du monde."

A. Camus

Donner un nom à chacune des classes trouvées permet de vérifier si on a réellement caractérisé des sous-populations.

L'impossibilité de nommer les sous-groupes doit questionner le praticien sur la pertinence opérationnelle de sa classification.

Nommer les classes

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

"Mal nommer les choses c'est ajouter au malheur du monde."

A. Camus

Donner un nom à chacune des classes trouvées permet de vérifier si on a réellement caractérisé des sous-populations.

L'impossibilité de nommer les sous-groupes doit questionner le praticien sur la pertinence opérationnelle de sa classification.

La classification obtenue va-t-elle faciliter la prise de décision des experts métiers ?

Caractériser les classes à l'aide d'un modèle

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les

Construire un arbre de décision : variable cible = le numéro de classe.

Si taux d'erreur faible, et que le nombre de règles n'est pas trop important, on peut même substituer à la classification initiale la classification définie par les règles de l'arbre de décision.

Permet de comprendre la formation des classes, donc de les expliquer, donc qu'elles soient utilisées par la suite.

Plus de détail dans tufféry, [7], pp235-236.

Trouver les variables qui sont significatives dans la description d'une classe

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Quelle est la question ?
Caractériser les classes
De la classe au segment, au type etc.
Exemples
Nommer les classes
Caractériser les classes à l'aide d'un modèle
Trouver les variables qui sont significatives dans la description d'une classe

$$v.test = \frac{\bar{x_k} - \bar{x}}{\sqrt{\frac{s^2}{n_k} \times \left(\frac{n-n_k}{n-1}\right)}} \sim \mathcal{N}(0,1)$$

avec $\bar{x_k}$ la moyenne de x dans la classe k, n_k l'effectif de la classe k, et $\bar{x}$ et s la moyenne et l'écart type de x dans le jeu de données (estimés par leurs estimateurs classiques), et sous l'hypothèse :

$$H_0 : \bar{x_k} = \bar{x}$$

On obtient alors une p-valeur pour chaque variable qui permet d'appréhender leur importance dans la classification obtenue.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Affecter un nouvel individu à une classe
Classifications Mixte
Classification sur axes factoriels
Miscellanées

Individus supplémentaires

Affecter un nouvel individu à une classe

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Affecter un nouvel individu à une classe
Classifications Mixte
Classification sur axes factoriels
Miscellanées

Conserver la cohérence avec les distances utilisées pour faire le clustering.

- ✔ K-means : affecter l'individu au centre le plus proche, au sens de la distance choisie pour mettre en place l'algorithme K-means
- ✔ CAH : selon la distance choisie :
  - ✘ single linkage : au groupe du plus proche voisin
  - ✘ complete linkage : au groupe dont le membre le plus éloigné est le plus proche
  - ✘ ward ou average $\equiv$ au groupe dont le centre est le plus proche

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Réduire la dimension "côté individu"
Optimiser une CAH à l'aide d'un k-means !
Classification sur axes factoriels

Classifications Mixte

Réduire la dimension "côté individus"

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Réduire la dimension "côté individu"
Optimiser une CAH à l'aide d'un k-means !
Classification sur axes factoriels

CAH : complexité d'ordre n^3. Rapidement infaisable avec un grand jeu de données.
Démarche : réduire la dimension côté "individus"

- ✔ Etape 1 : Effectuer un k-means avec k grand : "résumer" l'information générale
- ✔ Etape 2 : Effectuer une CAH des k centroïdes obtenus
  - ✘ Se servir du dendrogramme pour choisir le nombre de classes
  - ✘ Affecter à chaque individu la classe du centroïde auquel il était rattaché dans l'étape 1

Optimiser une CAH à l'aide d'un k-means !

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Réduire la dimension "côté individu"
Optimiser une CAH à l'aide d'un k-means !
Classification sur axes factoriels

- ✔ La CAH utilisant le critère de Ward minimise la perte d'inertie au passage de k à k-1 classes : pas forcément un optimum global.
- ✔ Effectuer un k-means en initialisant avec les centres des classes retenues à la CAH peut permettre de gagner un peu en inertie inter.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Réduire la dimension "côté variable"
Miscellanées

Classification sur axes factoriels

Réduire la dimension "côté variable"

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Réduire la dimension "côté variable"
Miscellanées

- ✔ Beaucoup de variables
- ✔ Éventuellement liées entre elles

$\Rightarrow$ Rechercher des combinaisons de variables, moins nombreuses, orthogonales entre elles, et qui expliquent au mieux l'information du nuage de points = axes factoriels.
Effectuer un clustering sur les m premiers axes factoriels.

Réduire la dimension "côté variable"

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Réduire la dimension "côté variable"
Miscellanées

- ✔ Beaucoup de variables
- ✔ Éventuellement liées entre elles

$\Rightarrow$ Rechercher des combinaisons de variables, moins nombreuses, orthogonales entre elles, et qui expliquent au mieux l'information du nuage de points = axes factoriels.
Effectuer un clustering sur les m premiers axes factoriels.

Permet la représentation mais attention à ne pas confondre les fonctions de réductions de l'information et de représentation ! Notamment, en pas prendre un nombre d'axes trop faible, au risque de perdre de l'information.

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

Miscellanées

Sur quelles variables travailler ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

Faut il garder toutes les variables ?

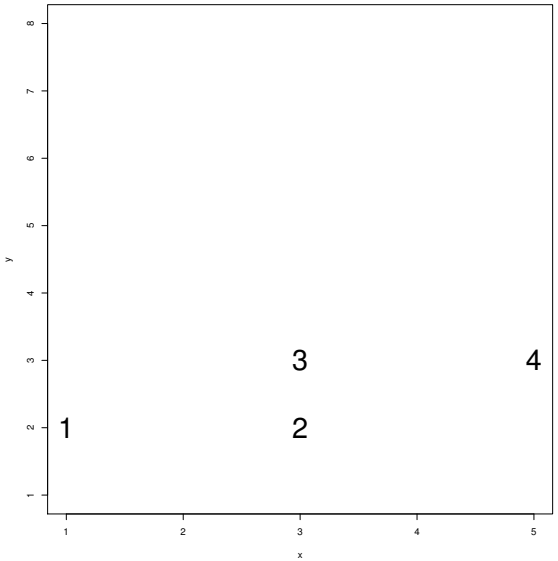
- ✔ Toutes les variables ne sont pas forcément pertinentes
- ✔ Variables corrélées $\rightarrow$ poids important au trait sous-jacent qu'elles mesurent :
  - ✘ Exemple : taille, longueur jambe, longueur bras, longueur torse
  - ✘ Une analyse préliminaire peut aider : Analyses factorielles (cf cours E. Miqueu.)
- ✔ Utiliser certaines variables comme variables illustratives (non utilisées pour partitionner). Variables géographiques, ou quali si toutes les autres sont quanti, ou classification prééexistante, par exemple à dire d'expert.

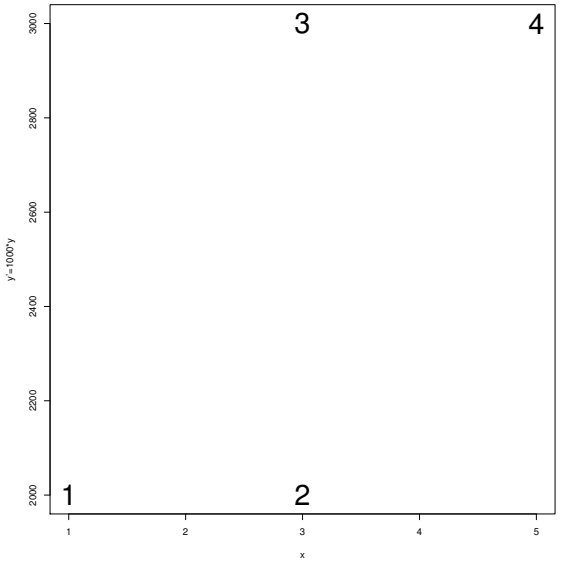
Réduire ou non ?

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

Beaucoup de distances calculées sur des variables quantitatives.

Faut-il réduire les variables ?
L'influence d'une variable dans un algorithme de clustering peut dépendre de l'unité de mesure choisie.





Réduire ou non : une heuristique

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non : une heuristique

- ✔ Si les variables sont exprimées dans des unités différentes
- ✔ Si on n'a pas de raison explicite de conserver les valeurs initiales

On peut au moins essayer de normaliser les variables et de voir quel effet cela peut avoir sur le clustering.
Si au contraire même unité et les différences ont du sens, ne pas réduire.

Réduire ou non : une heuristique

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non : une heuristique

- ✔ Si les variables sont exprimées dans des unités différentes
- ✔ Si on n'a pas de raison explicite de conserver les valeurs initiales

On peut au moins essayer de normaliser les variables et de voir quel effet cela peut avoir sur le clustering.
Si au contraire même unité et les différences ont du sens, ne pas réduire.
Exemple typique : variables de longitude et latitudes sur des mesures au Chili.

Comparer des classifications entre elles : la question

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

On a vu qu'il existait plusieurs manières de partitionner un ensemble de données.

Pour un nombre de classes donné, on a vu qu'il était possible de comparer les classes selon le critère de l'inertie expliquée B/T.

Comment savoir si deux classifications sont proches l'une de l'autre ou très différentes ?

Comparer des classifications entre elles : Indice de Rand

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

L'indice de Rand permet de mesurer la proximité entre deux classes.

$$R(P_1, P_2) = \frac{1}{n^2}\left[\sum_i \sum_{i' \neq i} c^1_{ii'} c^2_{ii'} + \sum_i \sum_{i' \neq i} \bar{c}^1_{ii'} \bar{c}^2_{ii'}\right]$$

avec P_1 et P_2 deux partitions des individus,

$$c^k_{ii'} = \begin{cases} = 1 \text{ si l'individu i et l'individu i' appartiennent à la même cla} \\ = 0 \text{ sinon} \end{cases}$$

$$\bar{c}^k_{ii'} = \begin{cases} = 1 \text{ si l'individu i et l'individu i' appartiennent à deux classes} \\ = 0 \text{ sinon} \end{cases}$$

Si les partitions sont identiques, l'indice de Rand vaut 1. Si les deux partitions diffèrent systématiquement, il vaut 0.

Références sur le clustering

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

Une première approche synthétique : wiki stats [8] : https://www.math.univ-toulouse.fr/ besse/Wikistat/

Le poly de fouille de données de J. Jacques. [4]. Disponible à cette url : http://eric.univ-lyon2.fr/j̃jacques/enseignement.html

Sur le clustering, une approche par l'exemple de Ricco Rakotomalala : [5].
Disponible à cette url :
http://eric.univ-lyon2.fr/řicco/cours/supports_data_mining.html
qui regorge de ressources sur le data-mining.

Sur la CAH, voir Husson [3], à cette url : http://factominer.free.fr/docs/tutorial_2010_clustering.pdf Sur le clustering en Général : Gasso et Leray [2], Durrieu [1]

References

Rappel sur le partitionnement
Distances
Centres mobiles
Décomposition de l'inertie du nuage de points
Classification ascendante hiérarchique
De la classe à la typologie
Individus supplémentaires
Classifications Mixte
Classification sur axes factoriels
Miscellanées
Sur quelles variables travailler ?
Réduire ou non ?
Réduire ou non :

[1] G. Durrieu. Analyse de données et data mining, inconnue. Poly de cours Université Bretagne Sud.

[2] G. Gasso and P. Leray. clustering, inconnue. Poly de cours pour l'INSA de Rouen.

[3] Husson. Classification ascendante hiérarchique, inconnue.

[4] J. Jacques. Fouilles de données / data mining, 2014. Poly de cours pour l'université de Lyon 2.

[5] R. Rakotomalala. Classification automatique sous r, Inconnue.

[6] P.N. and Tan. *Introduction to Data Mining*. Pearson, 2006.

[7] S. Tuffery. *Data mining et statistique décisionnelle, l'intelligence des données*. Technip, 2010.

[8] . Wikistat. Classification non-supervisée — wikistat, 2016. [En ligne; Page disponible le 21-janvier-2016].