Machine Learning
Apprentissage non supervisé

Stéphanie Bricq
stephanie.bricq@u-bourgogne.fr

Université de Bourgogne

2022

Plan

Apprentissage supervisé

- données d'entraînement fournies à l'algo comportent les solutions désirées, appelées *étiquettes* (*labels*)

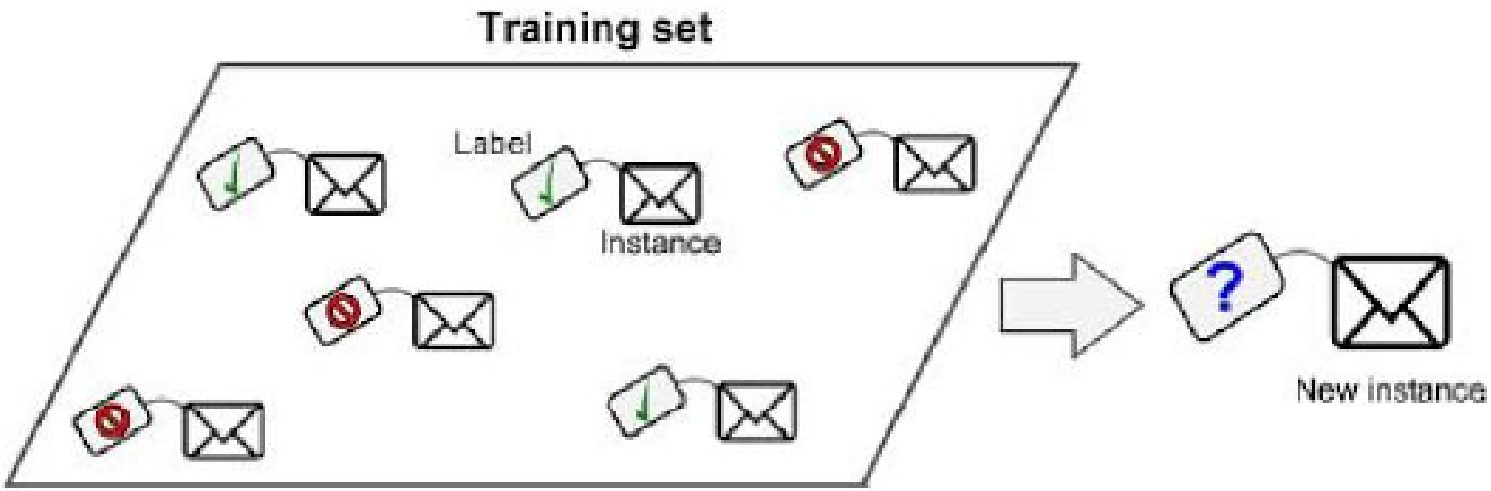


Fig. Jeu d'entraînement étiqueté pour apprentissage supervisé [Géron17]

Apprentissage supervisé

Exemples

- K plus proches voisins
- Régression linéaire
- Régression logistique
- Machines à vecteurs de support
- Arbres de décisions et forêts aléatoires
- Réseaux neuronaux

Apprentissage non supervisé

- les données d'apprentissage ne sont pas étiquetées

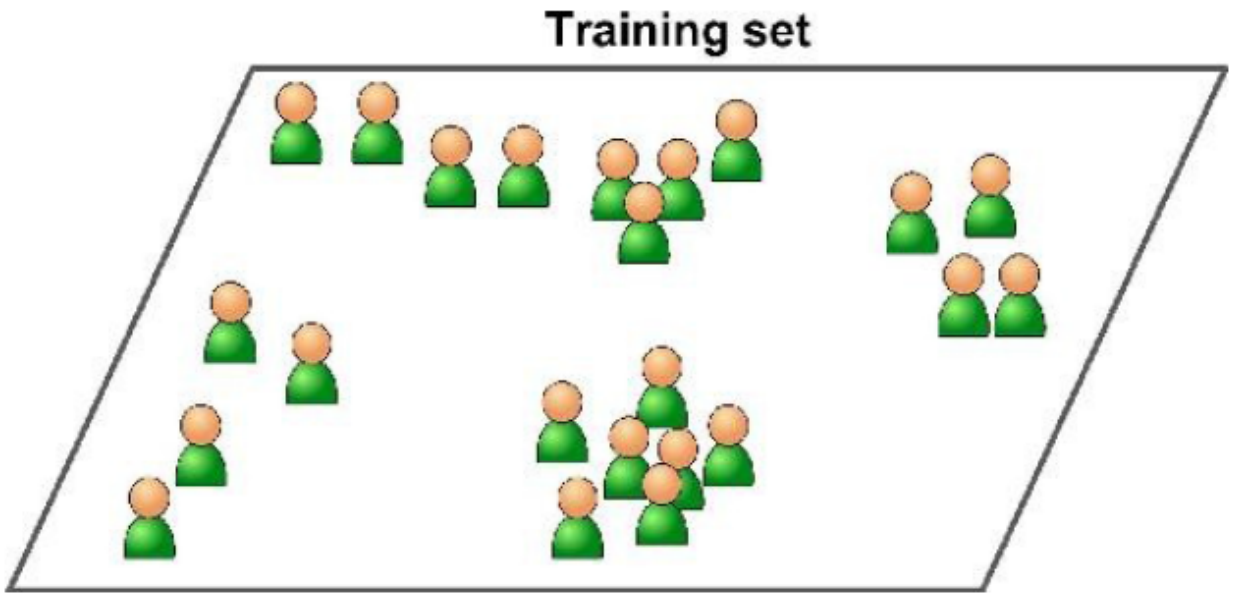


Fig. Jeu d'entraînement non étiqueté pour apprentissage non supervisé [Géron17]

Apprentissage non supervisé

Exemples

- Partitionnement
  - K-moyennes (*clustering*)
  - Partitionnement hiérarchique
- Visualisation et réduction de dimension
  - Analyse en composants principales
- Détection d'anomalies ou de nouveauté
  - One-class SVM
  - Isolation Forest
- Apprentissage par association de règles

Plan

Déterminer des groupes dans des données

Observer les individus et déterminer des motifs de comportement entre les individus (pas nécessairement humains), des groupes est une tâche essentielle pour produire de nouvelles connaissances.
On travaille avec les hypothèses suivantes :

- les groupes vont révéler des similarités
- l'identification des groupes (*clustering*) doit ressembler à un processus d'observation

Déterminer des groupes dans des données

Observer les individus et déterminer des motifs de comportement entre les individus (pas nécessairement humains), des groupes est une tâche essentielle pour produire de nouvelles connaissances.
On travaille avec les hypothèses suivantes :

- les groupes vont révéler des similarités
- l'identification des groupes (*clustering*) doit ressembler à un processus d'observation

Dans la suite de ce cours vous allez :

- comprendre en quoi le clustering diffère de la classification
- étudier un algorithme classique dont le principe est simple : k-means
- appliquer l'algorithme sur les données, les nettoyer et évaluer le résultat (TD)

Le clustering

Définition :

le clustering est une méthode non supervisée de machine learning qui consite à diviser un ensemble de données en groupes d'éléments similaires ou cluster.

Le clustering

Définition :

le clustering est une méthode non supervisée de machine learning qui consite à diviser un ensemble de données en groupes d'éléments similaires ou cluster.

- fonctionne sans avoir de connaissance a priori sur les caractéristiques des groupes.
- utilisé en analyse exploratoire pour découvrir des connaissances.
- ne produit pas directement de modèle prédictif comme le ferait une régression.
- peut s'intégrer dans une approche d'apprentissage semi-supervisée (par ex. k-means + decision tree) pour établir un modèle prédictif.

Le clustering

Définition :

le clustering est une méthode non supervisée de machine learning qui consite à diviser un ensemble de données en groupes d'éléments similaires ou cluster.

- fonctionne sans avoir de connaissance a priori sur les caractéristiques des groupes.
- utilisé en analyse exploratoire pour découvrir des connaissances.
- ne produit pas directement de modèle prédictif comme le ferait une régression.
- peut s'intégrer dans une approche d'apprentissage semi-supervisée (par ex. k-means + decision tree) pour établir un modèle prédictif.
- Le terme de classification est souvent utilisé comme synonyme, catégorisation serait plus précis.

Utilisations

Les clusters extraits des données peuvent être utilisées pour :

- segmenter des groupes de consommateurs, avec des caractéristiques démographies similaires, des comportements semblables, etc.

Utilisations

Les clusters extraits des données peuvent être utilisées pour :

- segmenter des groupes de consommateurs, avec des caractéristiques démographies similaires, des comportements semblables, etc.
- détecter des anomalies de comportement comme des intrusions sur des systèmes informatiques par identification des éléments qui ne correspondent pas à des clusters connus.

Utilisations

Les clusters extraits des données peuvent être utilisées pour :

- segmenter des groupes de consommateurs, avec des caractéristiques démographies similaires, des comportements semblables, etc.
- détecter des anomalies de comportement comme des intrusions sur des systèmes informatiques par identification des éléments qui ne correspondent pas à des clusters connus.
- simplifier, synthétiser des grands jeux de données en regroupant les éléments aux caractéristiques similaires, c-à-d en groupes homogènes partageant une ou plusieurs caractéristiques.

Utilisations

Les clusters extraits des données peuvent être utilisées pour :

- segmenter des groupes de consommateurs, avec des caractéristiques démographies similaires, des comportements semblables, etc.
- détecter des anomalies de comportement comme des intrusions sur des systèmes informatiques par identification des éléments qui ne correspondent pas à des clusters connus.
- simplifier, synthétiser des grands jeux de données en regroupant les éléments aux caractéristiques similaires, c-à-d en groupes homogènes partageant une ou plusieurs caractéristiques.

idée directrice : réduire la complexité pour pouvoir guider des actions.

Clustering et classification

- Les groupes n'ont **pas de label** : par exemple les groupes 1,2,3,4 sont extraits des données
- Aucun recours à un expert pour annoter les données

Clustering et classification

- Les groupes n'ont **pas de label** : par exemple les groupes 1,2,3,4 sont extraits des données
- Aucun recours à un expert pour annoter les données
- Le clustering produit de nouvelles données : les groupes
- C'est aux experts d'interpréter les résultats et de leur donner un signification : que signifient les groupes 1,2,3,4 ?

Le clustering est souvent qualifié de classification non supervisée.

Clustering et classification

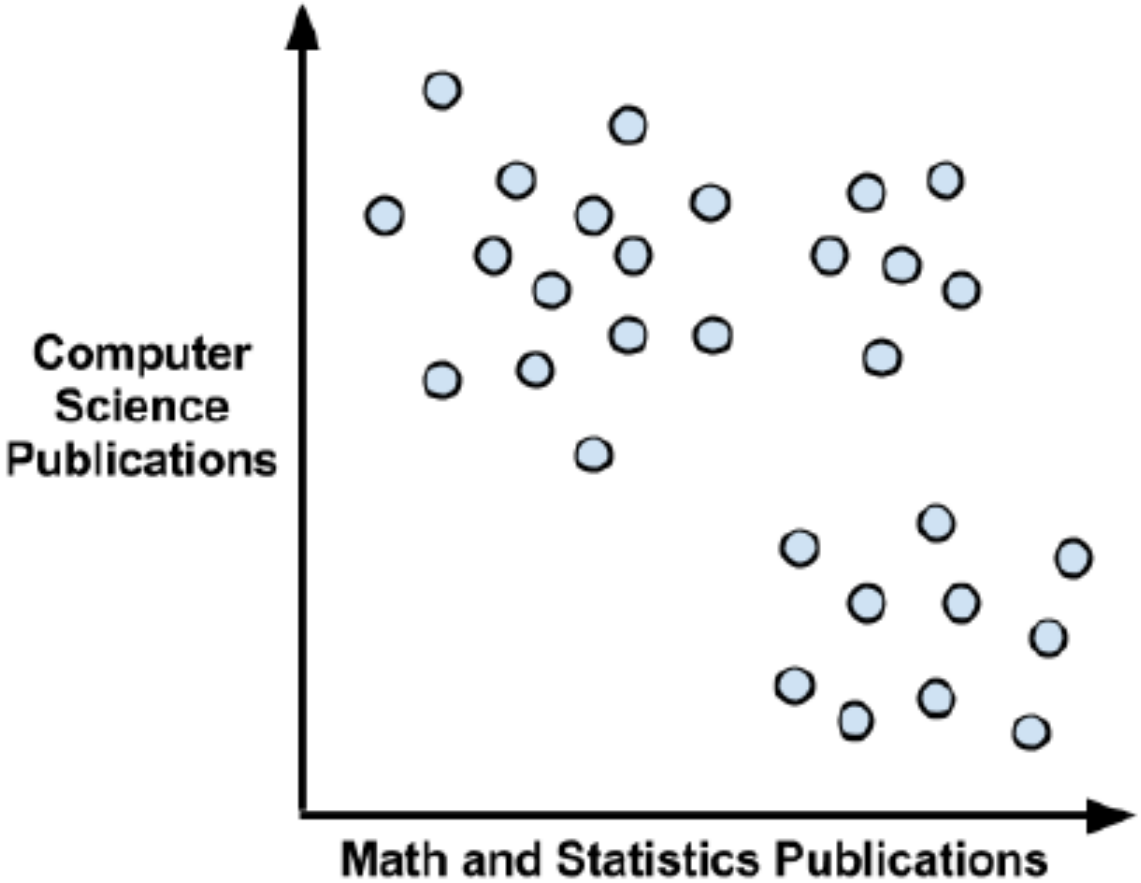
- Les groupes n'ont **pas de label** : par exemple les groupes 1,2,3,4 sont extraits des données
- Aucun recours à un expert pour annoter les données
- Le clustering produit de nouvelles données : les groupes
- C'est aux experts d'interpréter les résultats et de leur donner un signification : que signifient les groupes 1,2,3,4 ?

Le clustering est souvent qualifié de classification non supervisée.

Exemple :

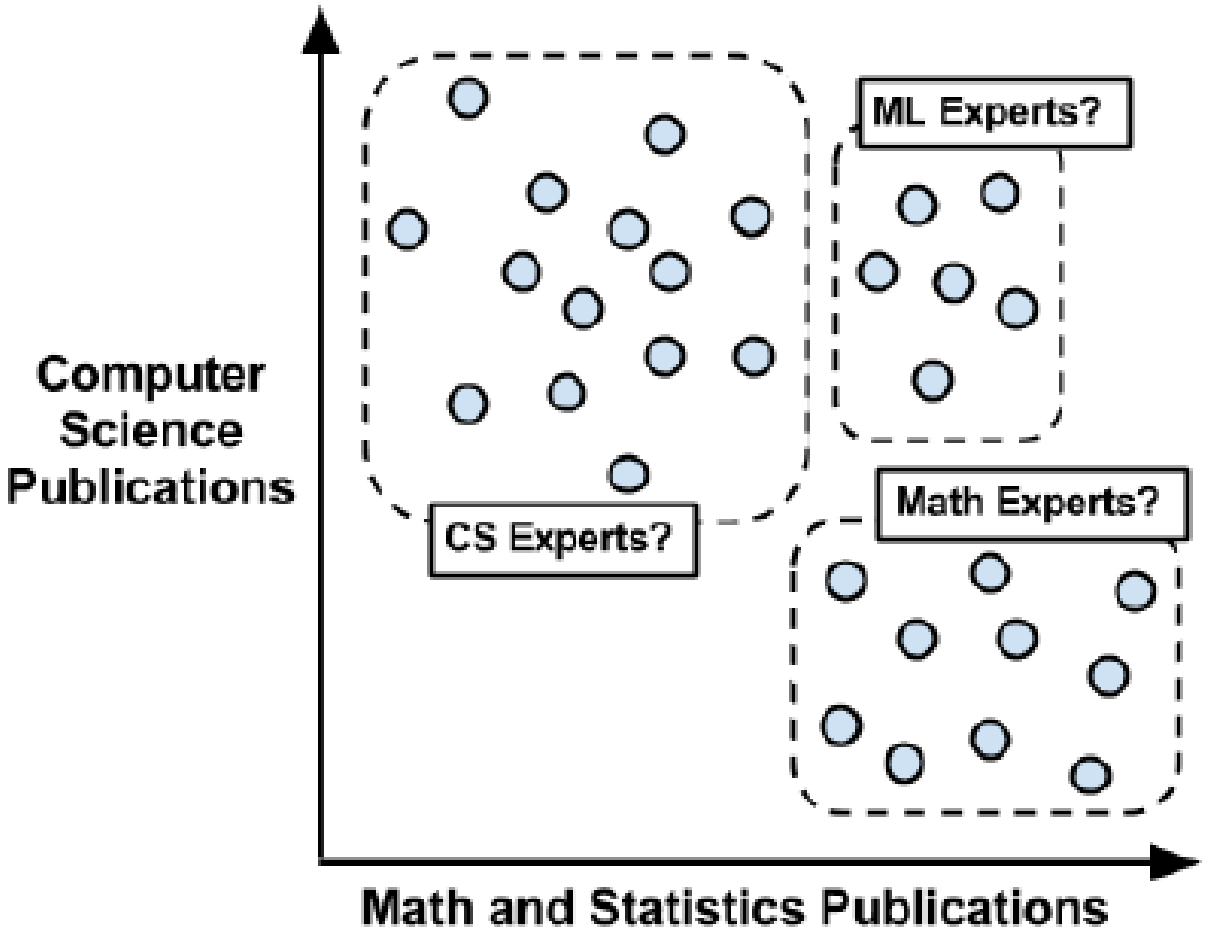
Dans un congrès de Science des Données, on souhaite regrouper les participants en spécialités mais on ne possède pas d'information sur leur discipline. À partir des bases de données de publications dans 2 domaines (informatique et maths-appliquées), on va essayer de déterminer des groupes en considérant que les participants ont deux coordonnées dans l'espace des publications (nb publis info, nb publis maths).

Clustering et classification



Machine Learning With R - Brett Lantz

Clustering et classification



Machine Learning With R - Brett Lantz

K-means

Avantages	Inconvénients
- basé sur des principes simples	- pas aussi sophisitqué que certains algos de clustering récents
- très flexible	- basé sur un élément aléatoire, pas de garantie de trouver les clusters optimales
- fonctionne bien dans la plupart des cas réels	- nécessite d'avoir une idée du nb de clusters

L'algorithme k-means

Il s'agit plutôt d'une famille d'algorithme relatifs à un même principe.

Principe :

Affecter n individus à k groupes avec la contrainte qu'un individu appartient à un seul groupe, en minimisant les différences intra-groupe et en maximisant les différences inter-groupes.

L'algorithme k-means

Il s'agit plutôt d'une famille d'algorithme relatifs à un même principe.

Principe :

Affecter n individus à k groupes avec la contrainte qu'un individu appartient à un seul groupe, en minimisant les différences intra-groupe et en maximisant les différences inter-groupes.

- Si k et n sont petits, il est possible de trouver la solution optimale, dans le cas contraire il faut un algorithme heuristique pour trouver une solution optimale locale.
- Cette heuristique utilise une mesure de qualité et un processus itératif

Principe : (Heuristique utilisée)

Démarrer avec une affectation (aléatoire) et modifier légèrement les affectations de manière à augmenter "l'homogénéité" des clusters.

L'algorithme k-means

L'algorithme est constitué de deux phases :

1. initialisation : configuration initiale des k clusters et affectation des n individus (aléatoire)

L'algorithme k-means

L'algorithme est constitué de deux phases :

1. initialisation : configuration initiale des k clusters et affectation des n individus (aléatoire)
2. mise-à-jour : réévaluation des clusters (mesure d'homogénéité) tant qu'il y a une amélioration ou un changement de la configuration

De part sa nature heuristique l'algorithme peut ne pas donner le même résultat d'une exécution à l'autre (modification légère des conditions initiales).

L'algorithme k-means : distance et mise-à-jour des clusters

On travaille avec plusieurs caractéristiques (*features*) qui forment un espace multi-dimensionnel de dimension d ($d = 2$ dans l'exemple)

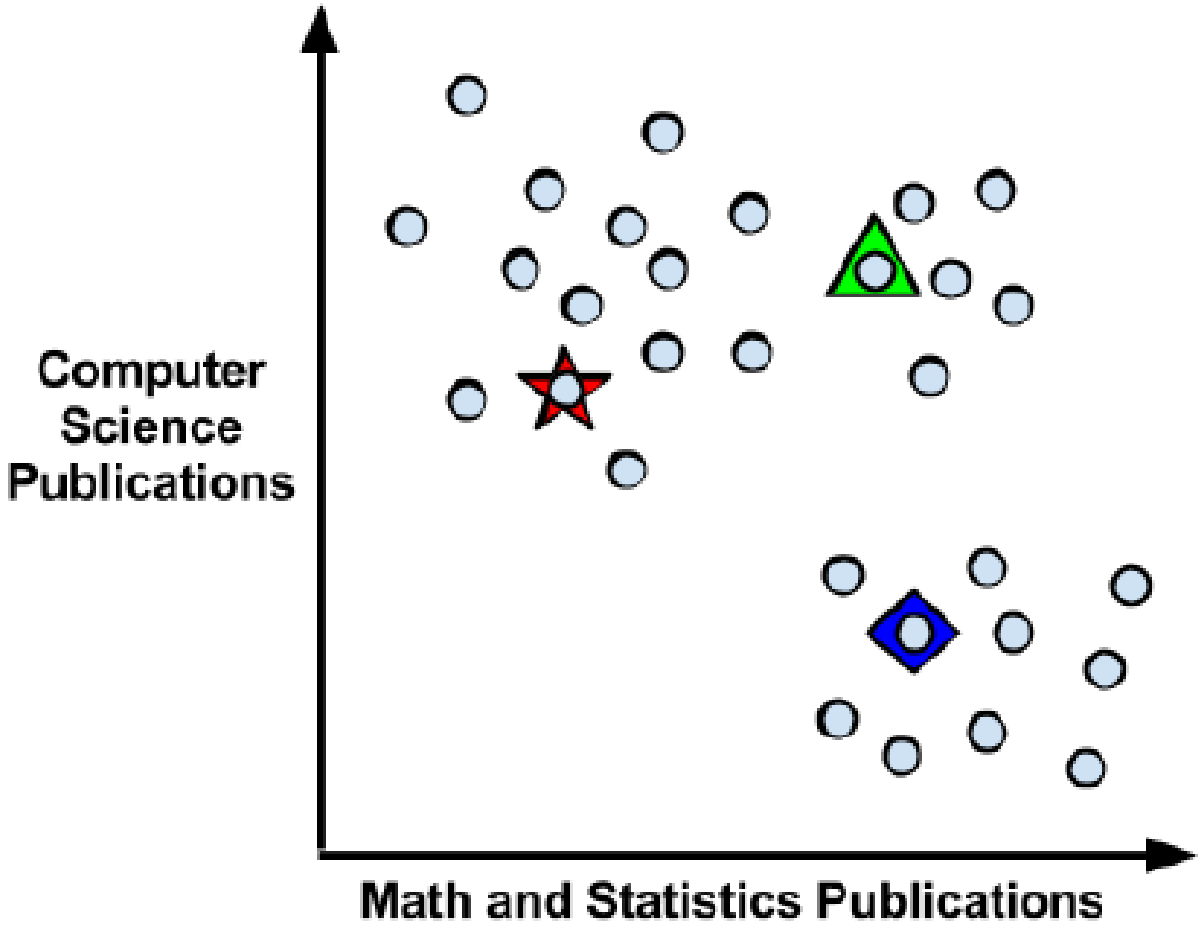
- configuration initiale : choisir k points dans l'espace des caractéristiques pour servir de centre aux clusters
  - au hasard mais pas trop proches
  - k représentants parmi les n individus
  - autre configuration initiale : affecter au hasard les n individus aux k clusters

L'algorithme k-means : distance et mise-à-jour des clusters

On travaille avec plusieurs caractéristiques (*features*) qui forment un espace multi-dimensionnel de dimension d ($d = 2$ dans l'exemple)

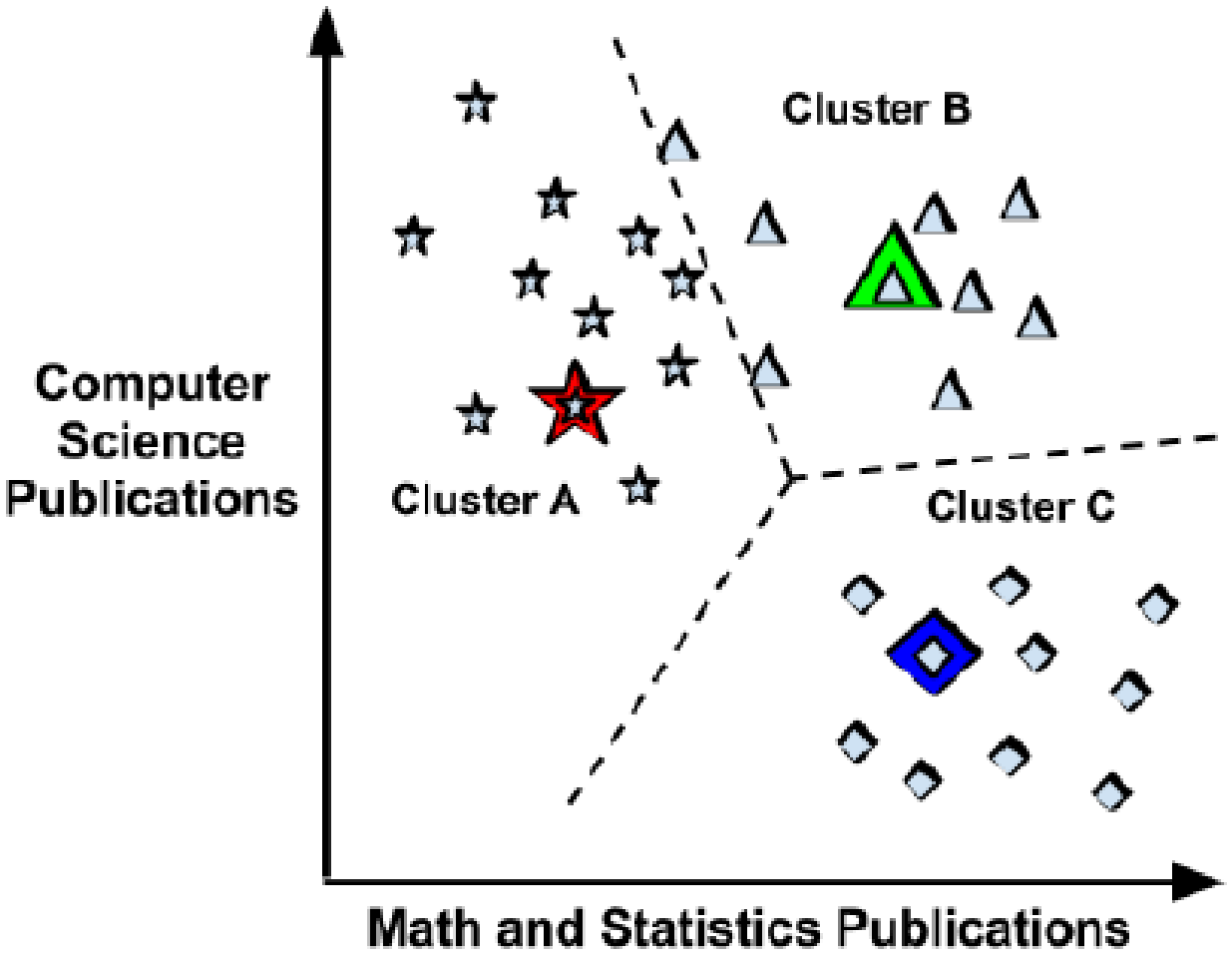
- configuration initiale : choisir k points dans l'espace des caractéristiques pour servir de centre aux clusters
  - au hasard mais pas trop proches
  - k représentants parmi les n individus
  - autre configuration initiale : affecter au hasard les n individus aux k clusters
- affectation : les (autres) individus sont affectés au cluster le plus proche (ou similaire)
  - $dist(u_1, u_2) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$
  - la distance est évaluée par rapport au centre du cluster (centroïde), à un représentant ?
  - d'autres distances sont utilisables

k-means : initialisation



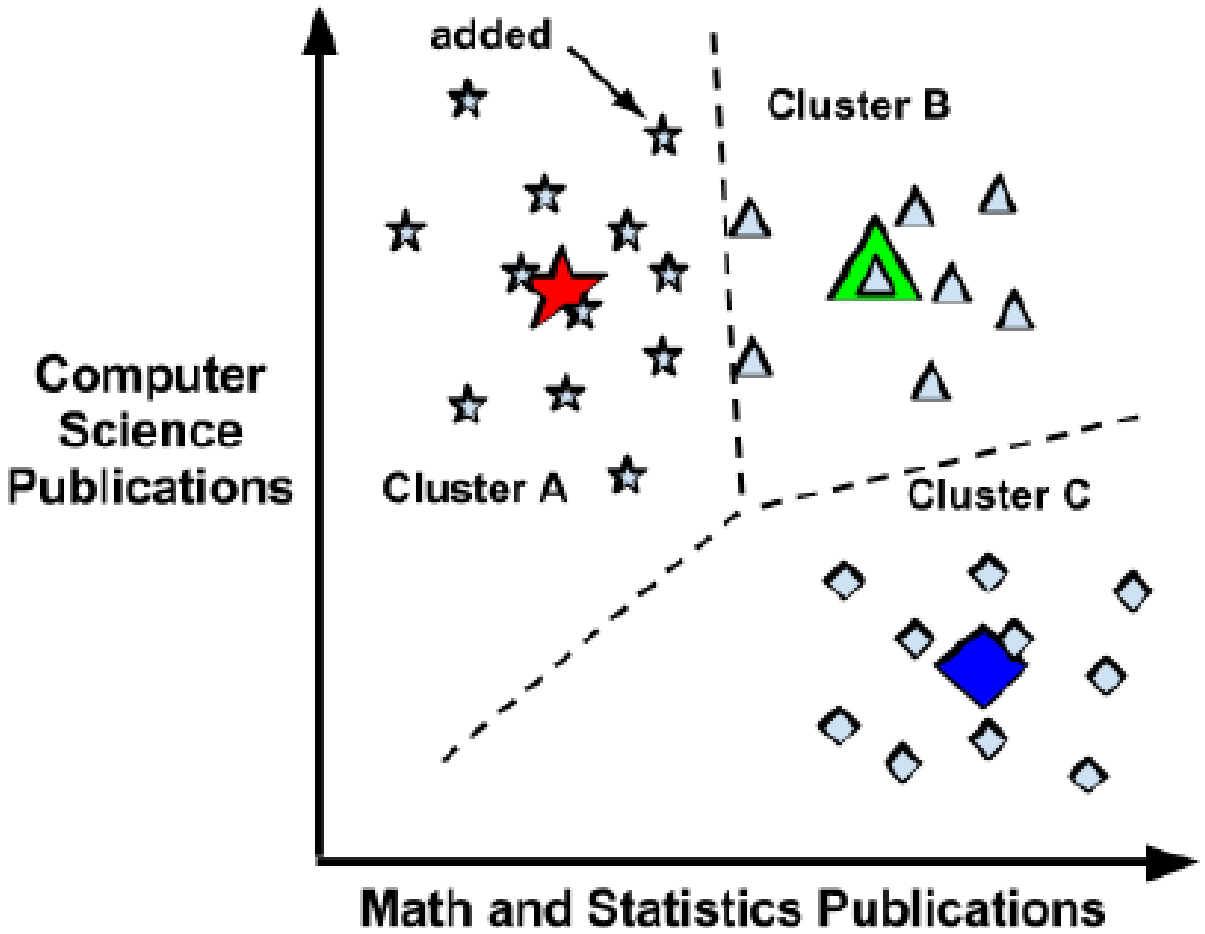
Machine Learning With R - Brett Lantz

k-means : première affectation



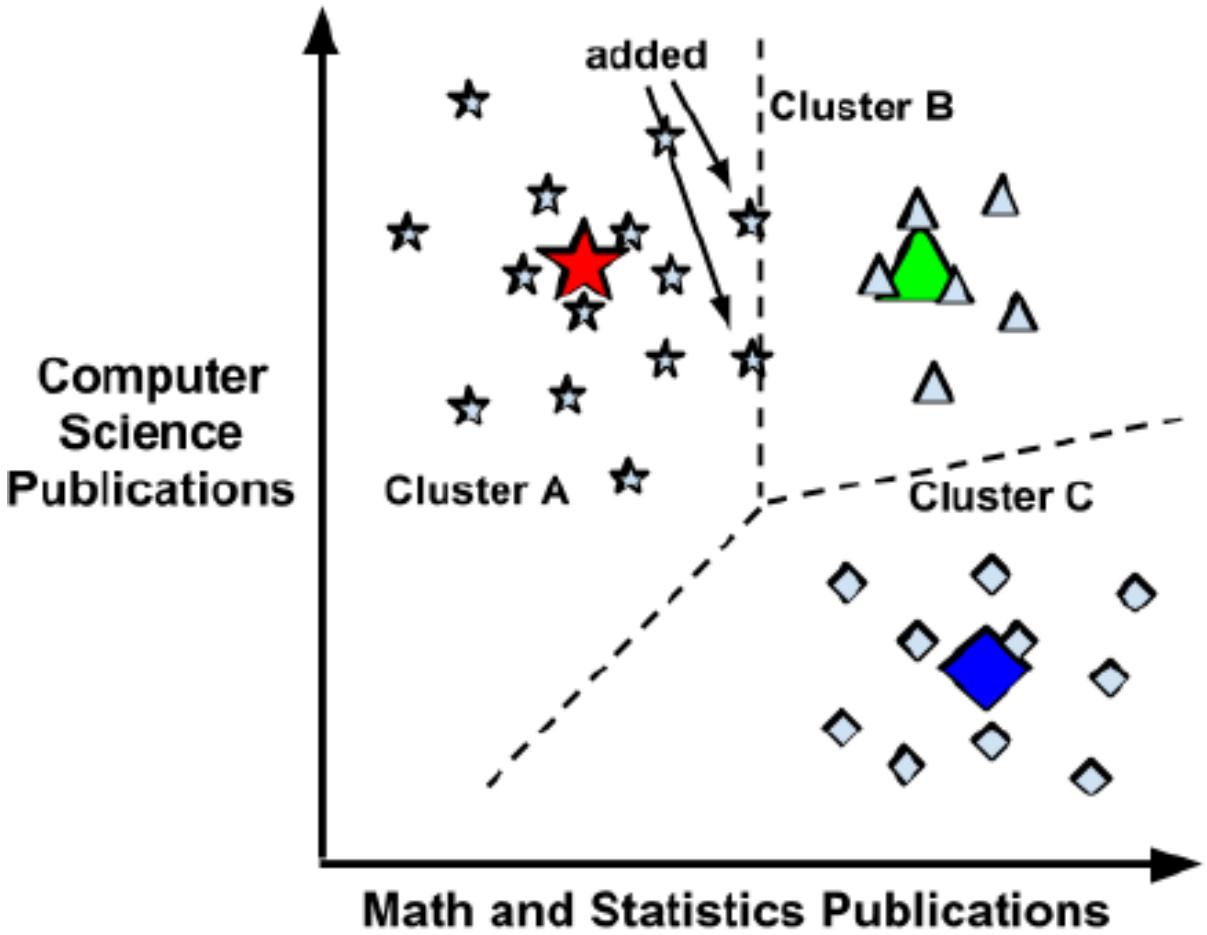
Machine Learning With R - Brett Lantz

k-means : mise-à-jour



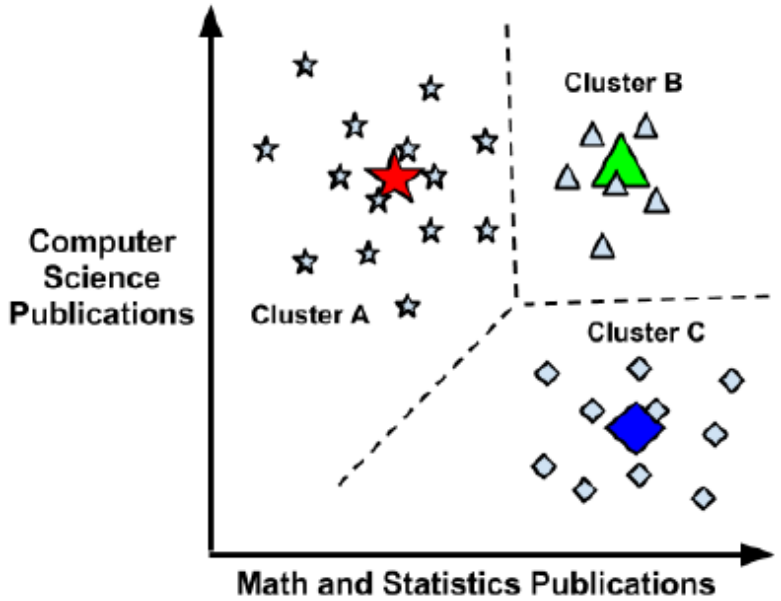
Machine Learning With R - Brett Lantz

k-means : mise-à-jour



Machine Learning With R - Brett Lantz

k-means : résultat



Machine Learning With R - Brett Lantz

Les éléments du résultat :

- liste des affectations
- centroïdes
- régions, diagrammes de Voronoï (`https://freakonometrics.hypotheses.org/19156`)

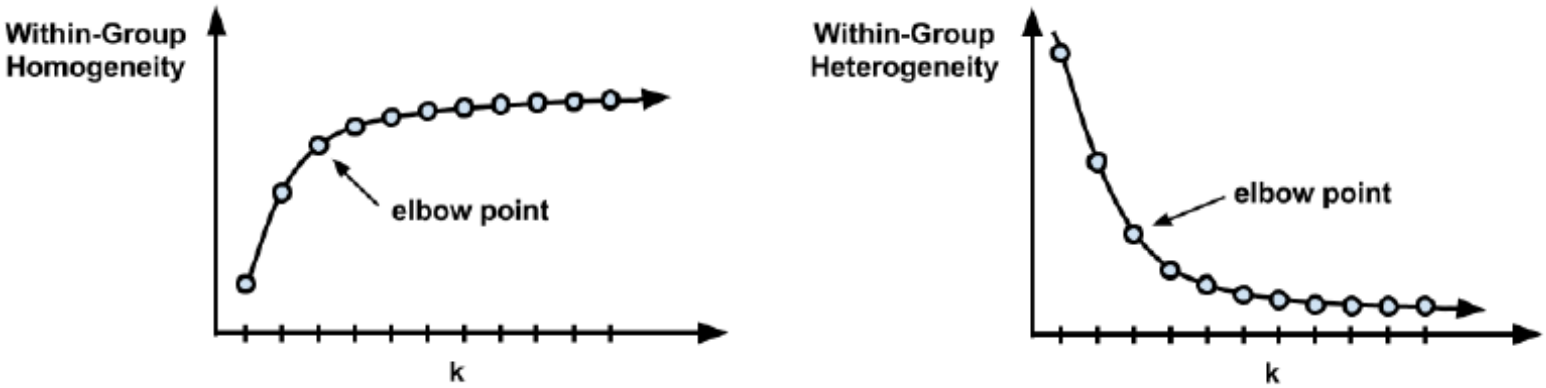
k-means : choisir k

Le choix de k peut être fait par :

- la connaissance du domaine
- une méthode objective : cohésion des clusters n'augmente plus avec k
  - évaluer la moyenne des distances au centre pour chaque k
  - nécessite plus d'exécution de k-means (peut être long)
  - mais peut donner d'autres informations
  - il existe de nombreuses mesures de la qualité des clusters
- empiriquement tester $k \approx \sqrt{n/2}$

k-means : choisir k

- *elbow method* : on regarde comment l'homogénéité, hétérogénéité des clusters change pour différentes valeurs de *k*



Machine Learning With R - Brett Lantz

Plan

Réduction de dimension

- dans certaines applications, le nombre de variables p utilisé pour représenter les données est très élevé.
- représentation des données contenant + de variables
  - $\Rightarrow$ + riche
    - mais + difficile d'apprendre un modèle performant

Réduction de dimension

Objectifs

But : transformer une représentation $X \in \mathbb{R}^{n \times p}$ des données en une représentation $X^* \in \mathbb{R}^{n \times m}$ où $m << p$

Réduction de dimension

Objectifs

But : transformer une représentation $X \in \mathbb{R}^{n \times p}$ des données en une représentation $X^* \in \mathbb{R}^{n \times m}$ où $m << p$

- visualiser les données
- réduire les coûts algorithmiques
- améliorer la qualité des modèles

Visualiser les données

- pour mieux définir les variables à utiliser
- pour éliminer si besoin les données aberrantes
- pour guider le choix des algos

Visualiser les données

- pour mieux définir les variables à utiliser
- pour éliminer si besoin les données aberrantes
- pour guider le choix des algos

$\Rightarrow$ utile de visualiser les données mais compliqué avec p variables

$\Rightarrow$ limiter les variables à un faible nb de dimensions pour visualiser + facilement les données (au risque de perdre un peu d'information)

Principe de l'ACP

- Matrice $X = [x_i^j], i = 1, ..., n; j = 1, ..., p$
  - ligne i -> observation
  - colonne j -> variable

Principe de l'ACP

- Matrice $X = [x_i^j], i = 1, ..., n; j = 1, ..., p$
  - ligne i -> observation
  - colonne j -> variable
- Questions ?
  - observations proches ou éloignées ?

  $\Rightarrow$ distance entre observations. On cherche une représentation qui les déforme le moins possible

Principe de l'ACP

- Matrice $X = [x_i^j], i = 1, ..., n; j = 1, ..., p$
  - ligne i -> observation
  - colonne j -> variable
- Questions ?
  - observations proches ou éloignées ?

  $\Rightarrow$ distance entre observations. On cherche une représentation qui les déforme le moins possible
  - variables généralement corrélées peuvent-elles être transformées pour donner d'autres variables aux propriétés plus intéressantes ?

Objectif de l'ACP

- résumer et visualiser un tableau de données individus * variables
- permet d'étudier les ressemblances entre individus du point de vue de l'ensemble des variables et dégage des profils d'individus

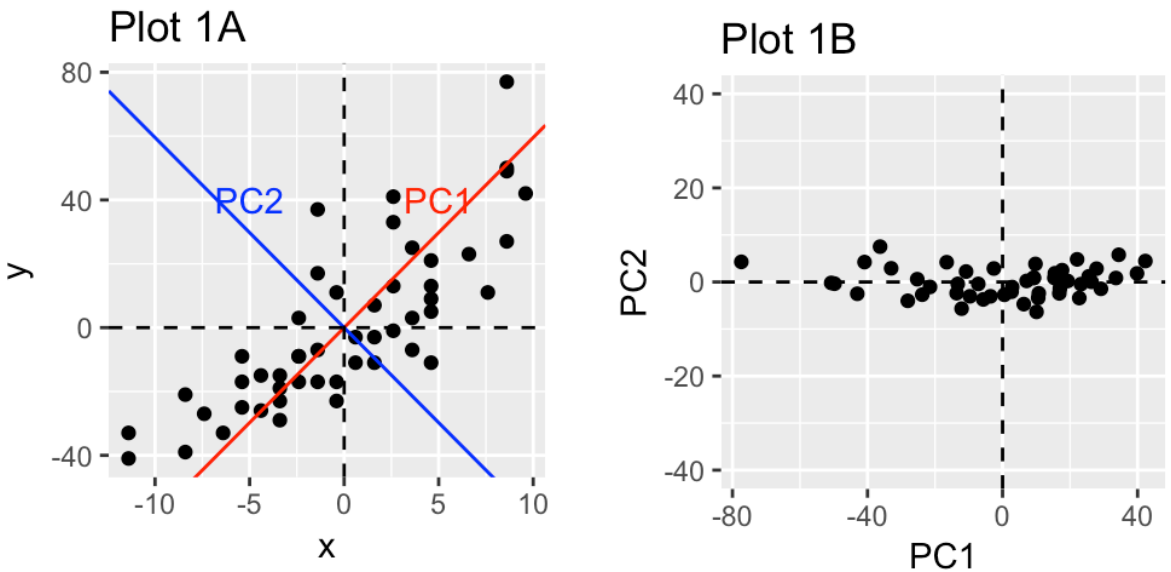
Objectif de l'ACP

- résumer et visualiser un tableau de données individus * variables
- permet d'étudier les ressemblances entre individus du point de vue de l'ensemble des variables et dégage des profils d'individus
- permet de réaliser un bilan des liaisons linéaires entre variables à partir des coefficients de corrélation
- on peut caractériser les individus ou groupes d'individus par les variables et illustrer les liaisons entre variables à partir d'individus caractéristiques

ACP

Une ACP de la matrice $X \in \mathbb{R}^{n \times p}$ est une transformation linéaire orthogonale qui permet d'exprimer X dans une nouvelle base orthonormée, de sorte que

- la plus grande variance de X par projection s'aligne sur le 1er axe de cette nouvelle base,
- la 2nde plus grande variance sur le 2ème axe ...



ACP

L'ACP vise à fournir une image simplifiée la + fidèle possible $\Leftrightarrow$
Trouver le sous-espace qui résume au mieux les données

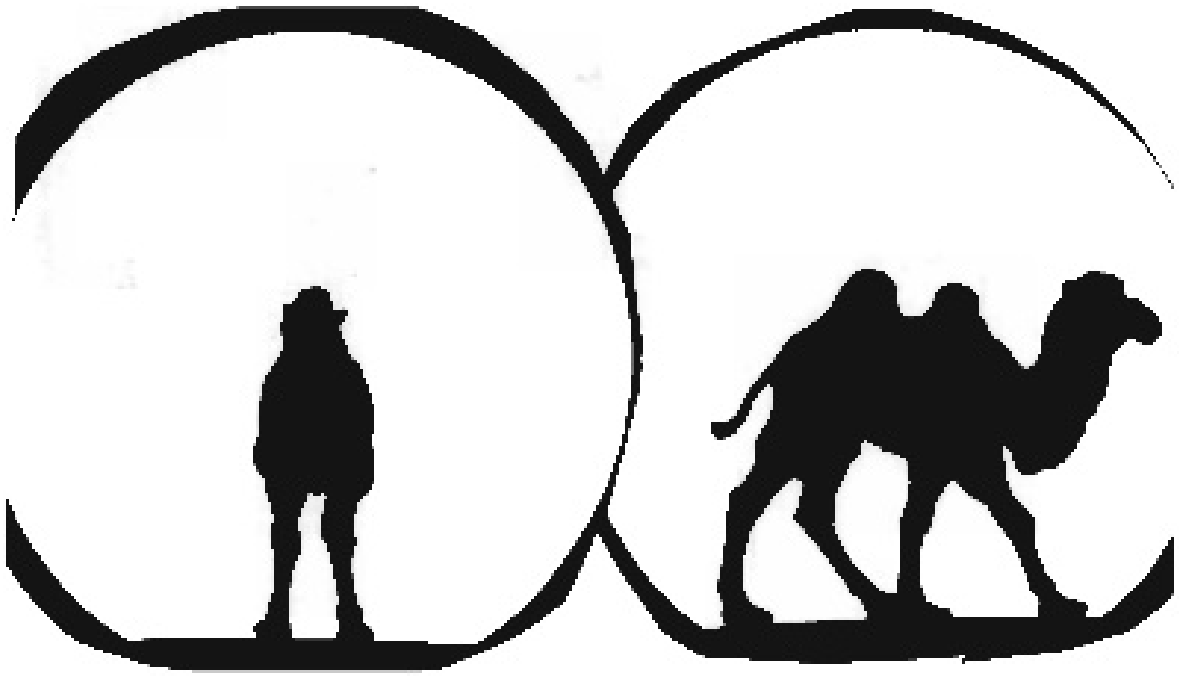


FIGURE: Chameau ou dromadaire ? source J.P. Fenelon

ACP : Cas d'une matrice centrée

- normalisation des variables pour avoir une moyenne de 0 et une variance de 1

Théorème

soit $X \in \mathbb{R}^{n \times p}$ une matrice centrée de covariance $\Sigma = \frac{1}{n} X^T X$.
Les composantes principales de X sont les vecteurs propres de Σ, ordonnés par valeur propre décroissante.

Décomposition en valeurs singulières(SVD)

Théorème

soit $X \in \mathbb{R}^{n \times p}$ une matrice centrée.
Les composantes principales de X sont ses vecteurs singuliers à droite ordonnés par valeur singulière décroissante

Décomposition en valeurs singulières(SVD)

Théorème

soit $X \in \mathbb{R}^{n \times p}$ une matrice centrée.
Les composantes principales de X sont ses vecteurs singuliers à droite ordonnés par valeur singulière décroissante

SVD

$A = U\Sigma V^\intercal$ avec A matrice $n \times m$

- U et V sont des matrices orthogonales ;
- Σ est une matrice diagonale (ou pseudo-diag) σ_i avec $\sigma_i^2 = \lambda_i$ et λ_i valeur propre de $AA^\intercal$
- **matrice orthogonale** : matrice carrée qui vérifie : $Q^\intercal Q = QQ^\intercal = I$

Décomposition en valeurs singulières(SVD)

Théorème

soit $X \in \mathbb{R}^{n \times p}$ une matrice centrée.
Les composantes principales de X sont ses vecteurs singuliers à droite ordonnés par valeur singulière décroissante

SVD

$A = U\Sigma V^\intercal$ avec A matrice $n \times m$

- U et V sont des matrices orthogonales ;
- Σ est une matrice diagonale (ou pseudo-diag) σ_i avec $\sigma_i^2 = \lambda_i$ et λ_i valeur propre de $AA^\intercal$
- **matrice orthogonale** : matrice carrée qui vérifie : $Q^\intercal Q = QQ^\intercal = I$
- On veut obtenir 2 ens de vecteurs singuliers **u** et **v** tels que :
  - les **u** sont les vecteurs propres de $AA^\intercal$
  - les **v** sont les vecteurs propres de $A^\intercal A$
  - $Av_i = \sigma_i u_i$

ACP

Choix du nb de composantes principales

- on utilise la proportion de variance expliquée par ses composantes
- la variance de X s'exprime comme la trace de Σ (= somme de ses valeurs propres)

ACP

Choix du nb de composantes principales

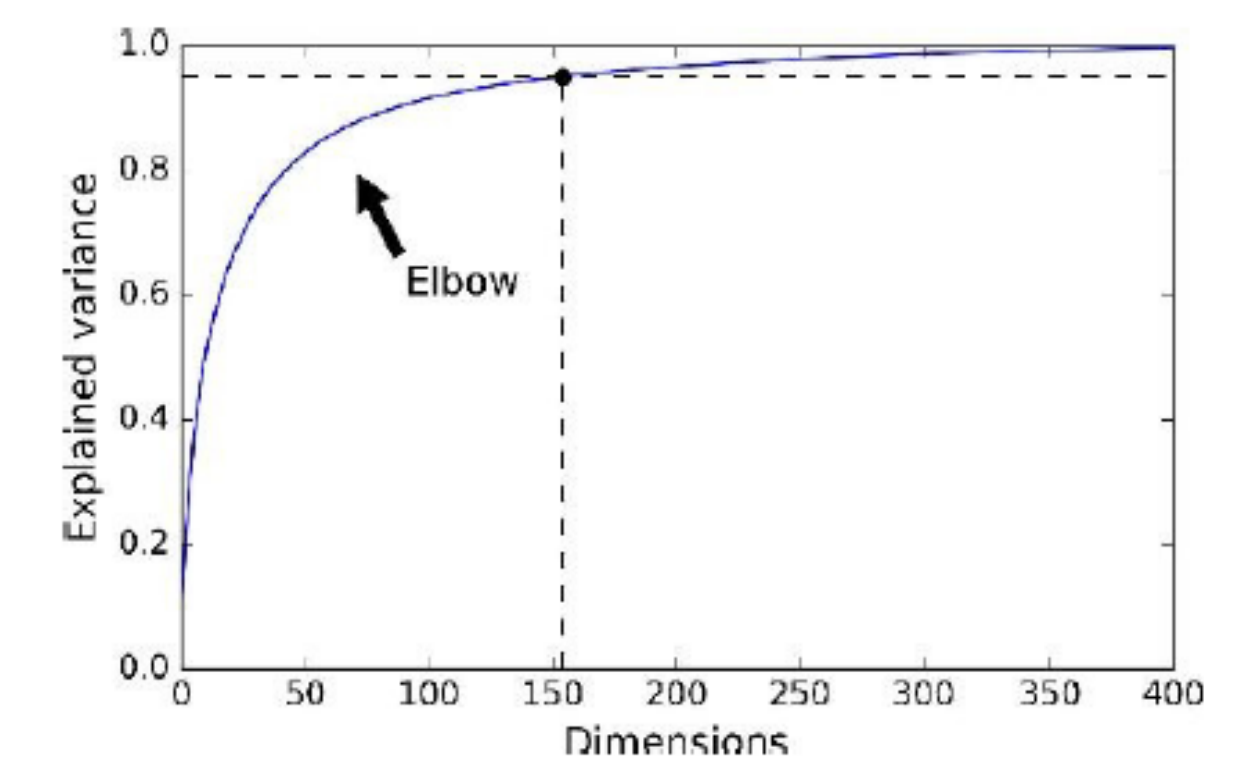
- on utilise la proportion de variance expliquée par ses composantes
- la variance de X s'exprime comme la trace de Σ (= somme de ses valeurs propres)
- ex : si on décide de garder les m premières composantes principales de X, la proportion de variance qu'elles expliquent est :

$$\frac{\alpha_1 + \alpha_2 + \cdots + \alpha_m}{Tr(\Sigma)}$$

où $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p$ sont les valeurs propres de Σ par ordre décroissant

ACP

Choix du nb de composantes principales



ACP

Factorisation de la matrice des données

- matrice $W \in \mathbb{R}^{p \times m}$ représentant les m composantes principales
- représentation réduite des n observations dans le nouvel espace de dim m s'obtient en projetant X sur les colonnes de W,

$$H = W^T X$$

ACP

Factorisation de la matrice des données

- matrice $W \in \mathbb{R}^{p \times m}$ représentant les m composantes principales
- représentation réduite des n observations dans le nouvel espace de dim m s'obtient en projetant X sur les colonnes de W,

$$H = W^T X$$

- $H \in \mathbb{R}^{m \times n}$ peut être interprétée comme une représentation latente des données
- les colonnes de W étant orthonormés (vecteurs propres de XX^T), on peut multiplier l'eq précédente à gauche par W pour obtenir une *factorisation* de X

$$X = WH$$

- lignes de H : *facteurs* de X

ACP : Exemple

- données : résultats d'épreuves de décathlon
- étapes
  1. importation du jeu de données
  2. choix des variables et des individus actifs
  3. standardiser ou non les variables
  4. choix du nb d'axes
  5. analyse des résultats

Exemple : données

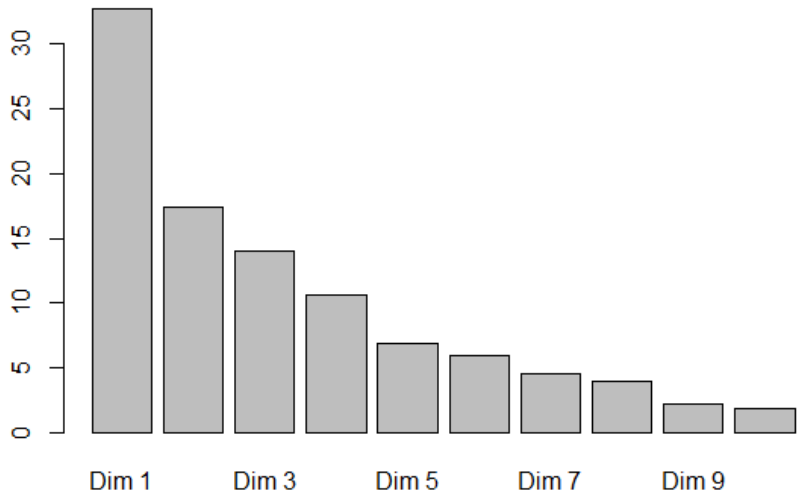
```
     100m          Longueur         Poids          Hauteur          400m          110m H
Min.   :10.44   Min.   :6.61   Min.   :12.68   Min.   :1.850   Min.   :46.81   Min.   :13.97
1st Qu.:10.85   1st Qu.:7.03   1st Qu.:13.88   1st Qu.:1.920   1st Qu.:48.93   1st Qu.:14.21
Median :10.98   Median :7.30   Median :14.57   Median :1.950   Median :49.40   Median :14.48
Mean   :11.00   Mean   :7.26   Mean   :14.48   Mean   :1.977   Mean   :49.62   Mean   :14.61
3rd Qu.:11.14   3rd Qu.:7.48   3rd Qu.:14.97   3rd Qu.:2.040   3rd Qu.:50.30   3rd Qu.:14.98
Max.   :11.64   Max.   :7.96   Max.   :16.36   Max.   :2.150   Max.   :53.20   Max.   :15.67
    Disque          Perche          Javelot          1500m        Classement        Points
Min.   :37.92   Min.   :4.200   Min.   :50.31   Min.   :262.1   Min.   : 1.00   Min.   :7313
1st Qu.:41.90   1st Qu.:4.500   1st Qu.:55.27   1st Qu.:271.0   1st Qu.: 6.00   1st Qu.:7802
Median :44.41   Median :4.800   Median :58.36   Median :278.1   Median :11.00   Median :8021
Mean   :44.33   Mean   :4.762   Mean   :58.32   Mean   :279.0   Mean   :12.12   Mean   :8005
3rd Qu.:46.07   3rd Qu.:4.920   3rd Qu.:60.89   3rd Qu.:285.1   3rd Qu.:18.00   3rd Qu.:8122
Max.   :51.65   Max.   :5.400   Max.   :70.52   Max.   :317.0   Max.   :28.00   Max.   :8893
  Competition
Decastar:13
JO      :28
```

Exemple

- choix des variables et des individus actifs
  - variables actives : les variables correspondant aux performances des athlètes : variables participant à la construction des axes
  - variables supplémentaires : nb de points et classement et competition. Utiles pour aider à l'interprétation
  - individus actifs
- standardiser ou non les variables : variables ont des unités différentes -> données centrées réduites

Exemple : Choix du nb d'axes

diagramme en barres des valeurs propres ou des inerties associées à chaque axe

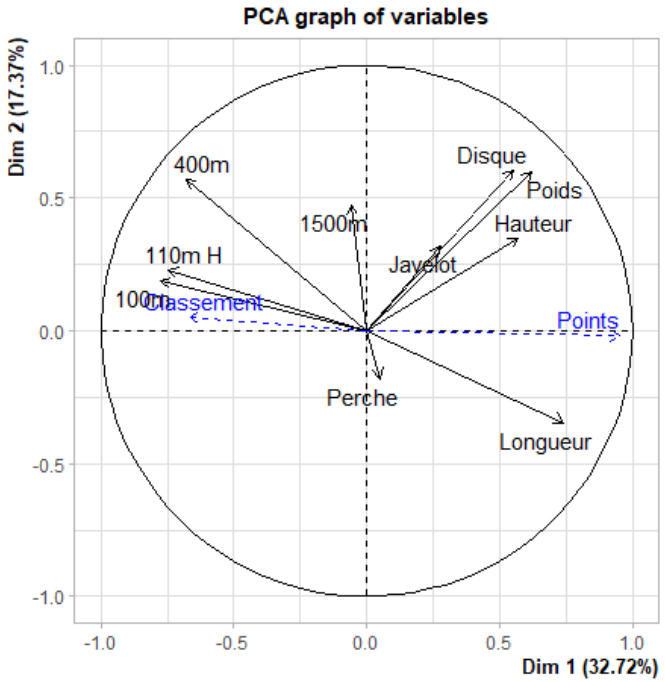


Pourcentage d'inertie associée à chaque dimension de l'ACP

- après 4 axes décroissance régulière
- les 2 premiers axes expriment 50% de l'inertie totale
- les 4 premiers axes expriment 75% de l'inertie totale

Exemple : analyse des résultats

Pour interpréter les résultat d'une ACP, l'usage est d'étudier simultanément les résultats sur les individus et sur les variables

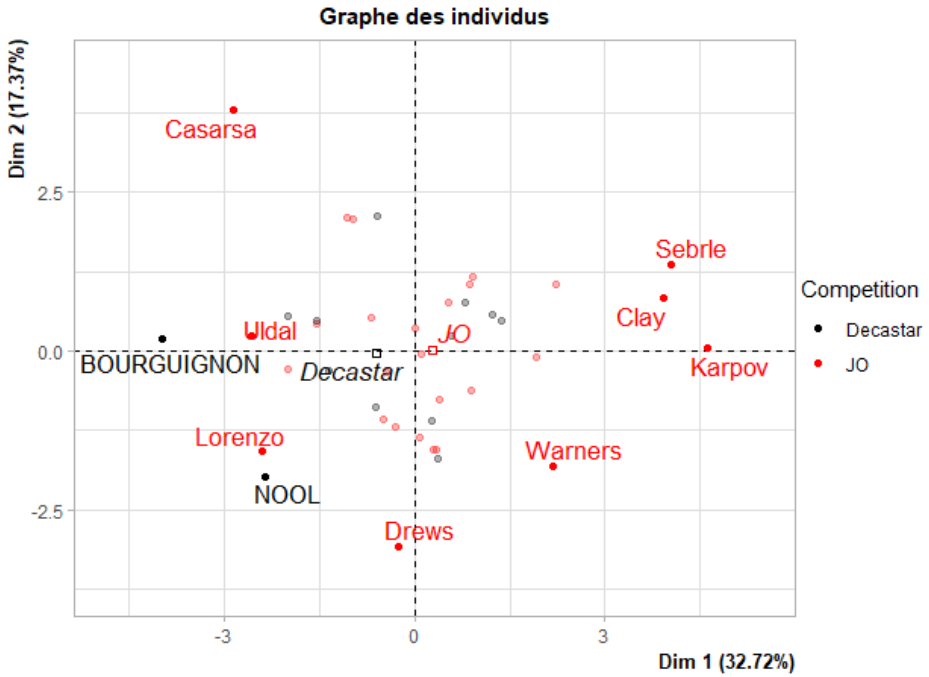


Graphe des variables pour les 2 premières dimensions

- première composante principale : combinaison linéaire des variables qui synthétise le mieux l'ensemble des variables

Exemple : analyse des résultats

Pour interpréter les résultat d'une ACP, l'usage est d'étudier simultanément les résultats sur les individus et sur les variables



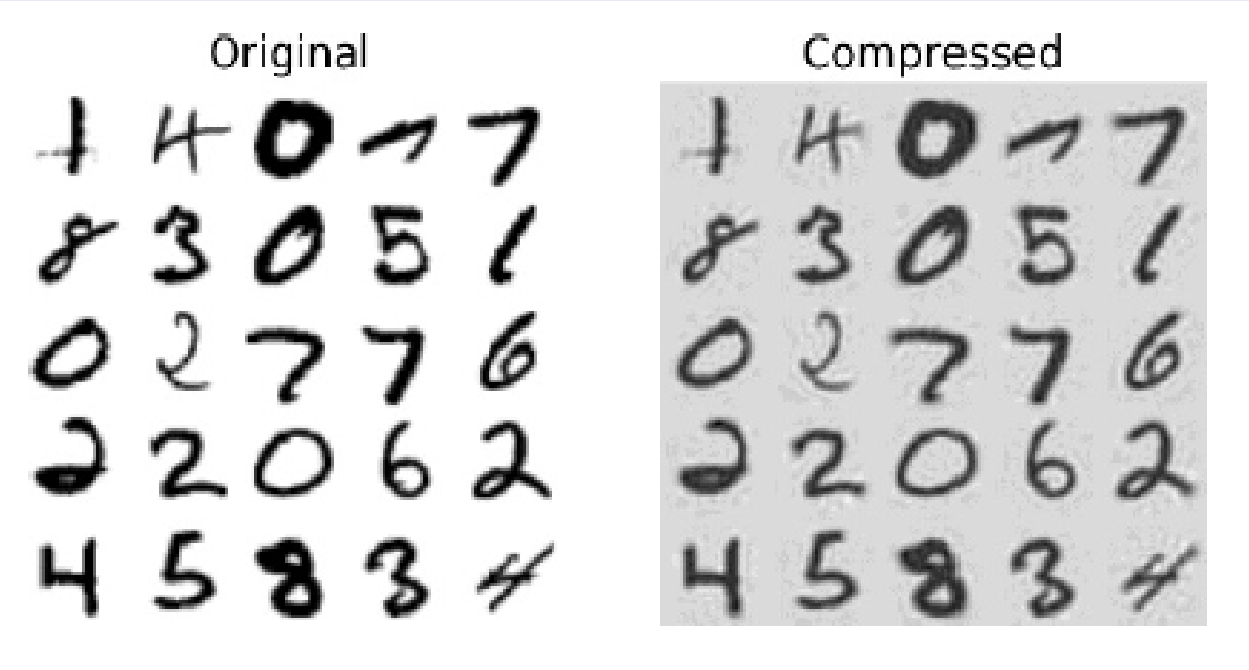
Graphe des individus pour les 2 premières dimensions

- 1er axe : oppose des profils de performance élevée (Karpov) aux profils (relativement) faibles partout (Bourguignon)

ACP

Application pour la compression

On garde un peu plus de 150 variables au lieu des 784.



Plan

Classification Ascendante Hiérarchique (CAH)

- outil très utilisé en analyse de données
- construire un arbre binaire des données qui successivement fusionne des groupes de points similaires
- visualisation de cet arbre fournit un résumé utile des données

Clustering hiérarchique vs K-Means

K-Means nécessite

- un nombre de classes K
- une affectation initiale des données aux clusters
- une mesure de distance entre les données $d(x_n, x_m)$

Clustering hiérarchique vs K-Means

K-Means nécessite

- un nombre de classes K
- une affectation initiale des données aux clusters
- une mesure de distance entre les données $d(x_n, x_m)$

CAH nécessite

- une mesure de similarité entre groupes de points de données.

Principes de la CAH

Quelles données ?

tableaux de données individus $\times$ variables quantitatives

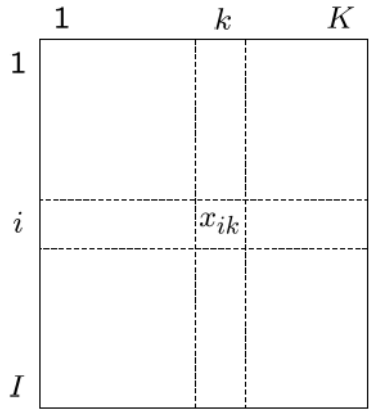
Principes de la CAH

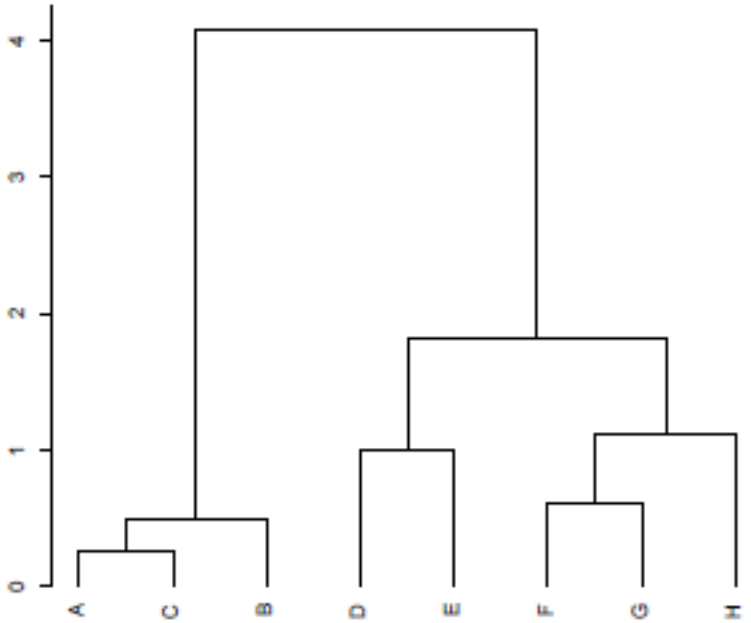
Quelles données ?

tableaux de données individus $\times$ variables quantitatives

Objectifs : arbre hiérarchique (*dendogramme*)

- mise en évidence de liens hiérarchiques entre individus ou groupes d'individus
- détection d'un nb de classes «naturel» au sein de la population





Ressemblance entre individus

- Distance euclidienne

$$d^2(i,l) = \sum_{k=1}^{K} (x_{ik} - x_{lk})^2$$

avec x_{ik} valeur de l'individu i pour la variable k (on suppose que les données x_{ik} ont été au préalable centrées et réduites)

Ressemblance entre individus

- Distance euclidienne

$$d^2(i,l) = \sum_{k=1}^{K} (x_{ik} - x_{lk})^2$$

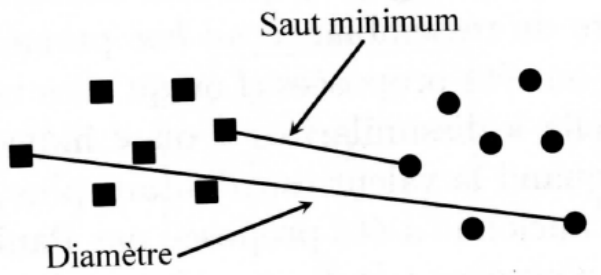
avec x_{ik} valeur de l'individu i pour la variable k (on suppose que les données x_{ik} ont été au préalable centrées et réduites)

- Ex de distance non euclidienne : distance de Manhattan

$$d(i,l) = \sum_{k=1}^{K} |x_{ik} - x_{lk}|$$

Ressemblance entre groupes d'individus

Plusieurs possibilités existent pour définir une distance ou dissimilarité entre groupes d'individus. Soit 2 groupes d'individus A et B

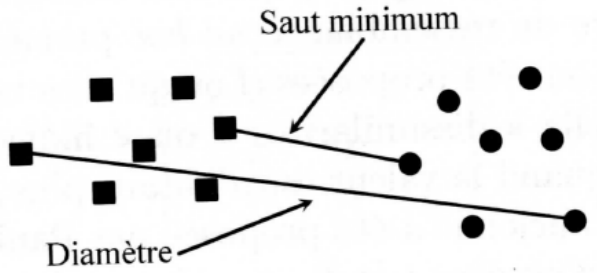


- saut minimum entre A et B = plus petite des distances entre un élément de A et un élément de B
- diamètre entre A et B = plus grande des distances entre un élément de A et un élément de B

$\Rightarrow$ définitions applicables à toutes les distances

Ressemblance entre groupes d'individus

Plusieurs possibilités existent pour définir une distance ou dissimilarité entre groupes d'individus. Soit 2 groupes d'individus A et B



- saut minimum entre A et B = plus petite des distances entre un élément de A et un élément de B
- diamètre entre A et B = plus grande des distances entre un élément de A et un élément de B

$\Rightarrow$ définitions applicables à toutes les distances

- Cas de distances euclidiennes : d'autres possibilités existent. Soit G_A et G_B les centres de gravité des ensembles d'individus A et B
  - distance entre les centres de gravité
  - inertie inter : inertie de $\{G_A, G_B\}$ par rapport à G (centre de gravité de $A \cup B$)

Algorithme classique de construction ascendante

- Point de départ : matrice de dissimilarités D entre individus dont le terme général $d(i, l)$ est la dissimilarité entre les individus i et l

$\Rightarrow$ matrice symétrique avec des 0 sur la diag

Algorithme classique de construction ascendante

- Point de départ : matrice de dissimilarités D entre individus dont le terme général $d(i, l)$ est la dissimilarité entre les individus i et l

$\Rightarrow$ matrice symétrique avec des 0 sur la diag

- on agrège les individus i et l les + similaires (si égalité -> choix arbitraire)
  - on constitue un nouvel élément (i, l)
  - valeur $d(i, l)$: indice de l'agrégation entre i et l -> hauteur à laquelle les branches de l'arbre qui correspondent à i et à l se rejoignent

Algorithme classique de construction ascendante

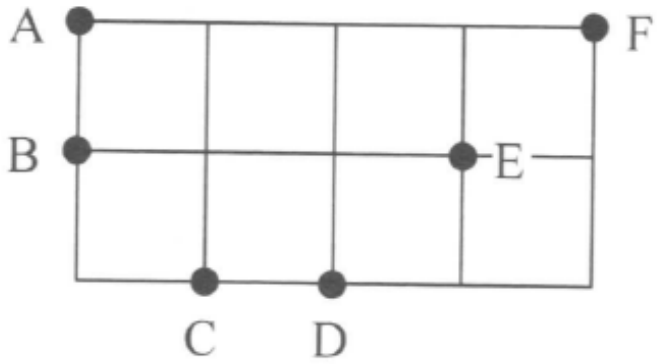
- Point de départ : matrice de dissimilarités D entre individus dont le terme général $d(i, l)$ est la dissimilarité entre les individus i et l

$\Rightarrow$ matrice symétrique avec des 0 sur la diag

- on agrège les individus i et l les + similaires (si égalité -> choix arbitraire)
  - on constitue un nouvel élément (i, l)
  - valeur $d(i, l)$: indice de l'agrégation entre i et l -> hauteur à laquelle les branches de l'arbre qui correspondent à i et à l se rejoignent
- mise à jour de la matrice D en supprimant les lignes et les colonnes correspondant aux individus i et l et en créant une nouvelle ligne et une nouvelle colonne pour le groupe (i, l) que l'on remplit avec les dissimilarités entre ce groupe et chacun des individus restants

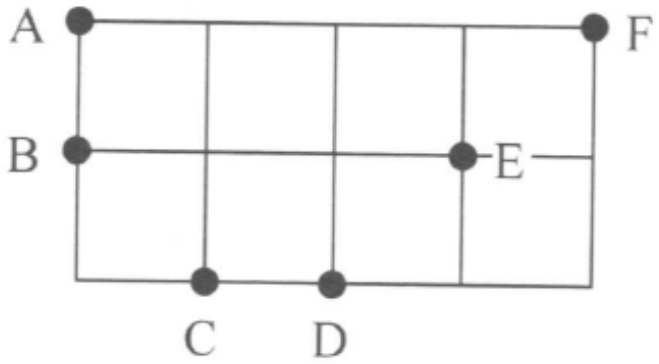
Exemple

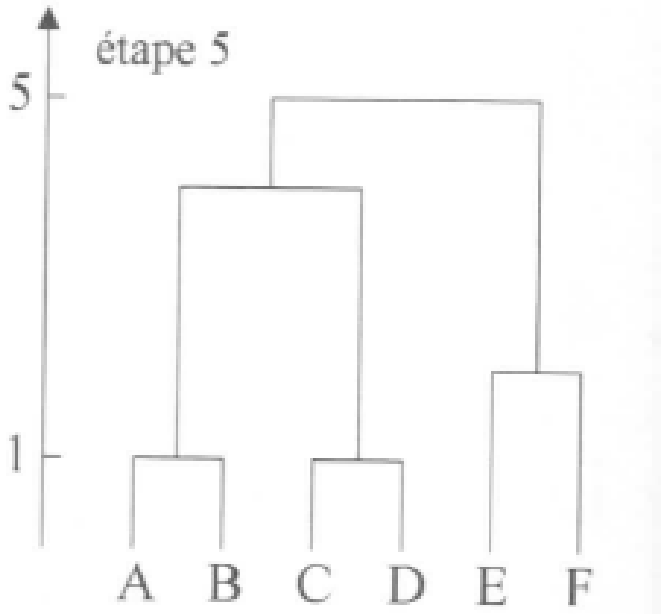
distance initiale : distance de Manhattan
recalcul des distances selon le diamètre



Exemple

distance initiale : distance de Manhattan
recalcul des distances selon le diamètre



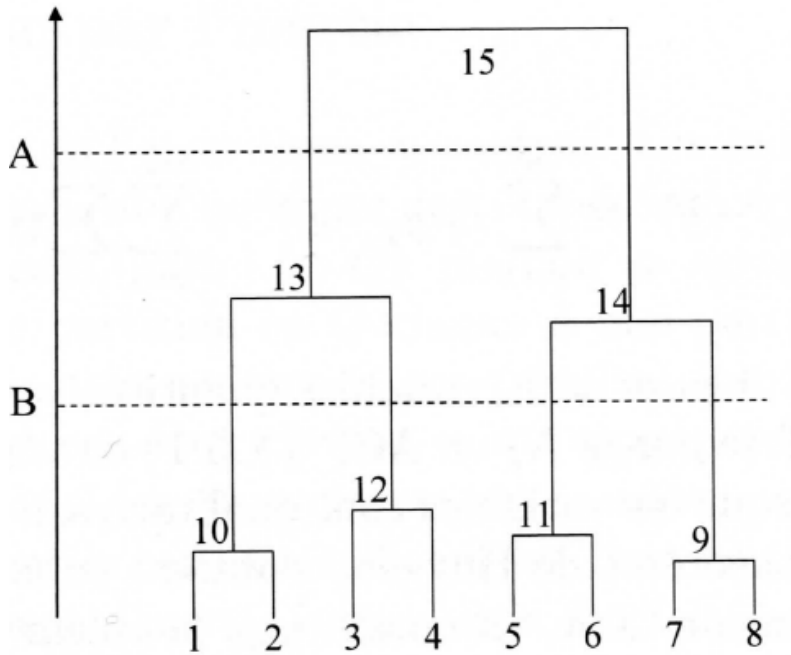


Hiérarchie et partition

- si I individus : $(I - 1)$ noeuds
- si on trace une ligne horizontale à un indice donné, on définit une partition (on coupe l'arbre)

Hiérarchie et partition

- si I individus : $(I - 1)$ noeuds
- si on trace une ligne horizontale à un indice donné, on définit une partition (on coupe l'arbre)
- arbre hiérarchique : suite de partition emboitées
  - allant de la plus fine (chaque individu constitue une classe)
  - à la plus grossière (une seule classe)



Qualité d'une partition

Une bonne partition est telle que

- individus homogènes à l'intérieur d'une classe

$\Rightarrow$ variabilité intra-classe faible

Qualité d'une partition

Une bonne partition est telle que

- individus homogènes à l'intérieur d'une classe

$\Rightarrow$ variabilité intra-classe faible

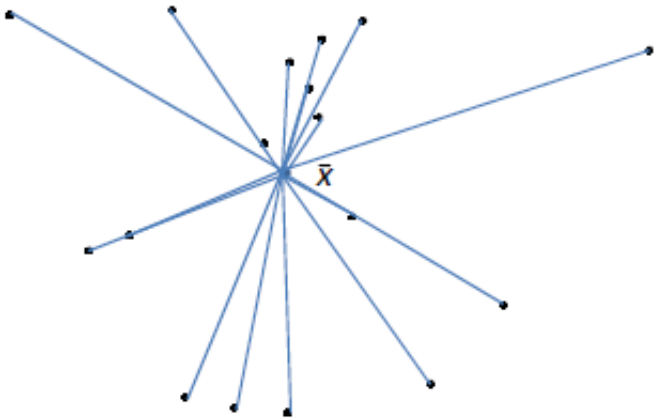
- individus différents d'une classe à une autre

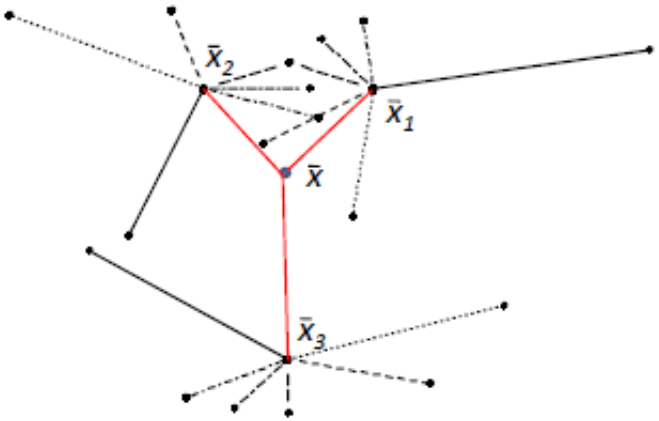
$\Rightarrow$ variabilité inter-classes élevée

$\Longrightarrow$ 2 critères, lequel choisir ?

Qualité d'une partition

$$\underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_k)^2}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_{qk})^2}_{\text{Inertie intra}} + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(\bar{x}_{qk}-\bar{x}_k)^2}_{\text{Inertie inter}}$$





- x_{iqk} valeur pour la $k$ème variable du $i$ème individu de la classe q
- $\bar{x}_{qk}$ moyenne de x_k dans la classe q
- $\bar{x}_k$ moyenne de x_k

$\Longrightarrow$ 1 seul critère

Qualité d'une partition

La qualité d'une partition peut être mesurée par :

$$\frac{\text{Inertie inter-classes}}{\text{Inertie totale}}$$

Ce critère dépend du nb d'individus et du nb de classes

- partition dans laquelle chaque individu constitue une classe $\Rightarrow$ critère=1 mais aucun intérêt pratique

Méthode de Ward

- Initialisation : 1 classe =1 individu

$\Rightarrow$ Inertie inter= Inertie totale

Méthode de Ward

- Initialisation : 1 classe =1 individu

$\Rightarrow$ Inertie inter= Inertie totale

- A chaque étape, les individus sont répartis en Q classes obtenues par les étapes précédents
  - comment choisir les 2 classes (parmi les Q) à agréger ?
  - en agrégeant 2 classes, on passe d'une partition en Q classes à une partition en Q-1 classes

  $\Rightarrow$ inertie intra-classe va augmenter

Méthode de Ward

- Initialisation : 1 classe =1 individu

$\Rightarrow$ Inertie inter= Inertie totale

- A chaque étape, les individus sont répartis en Q classes obtenues par les étapes précédents
  - comment choisir les 2 classes (parmi les Q) à agréger ?
  - en agrégeant 2 classes, on passe d'une partition en Q classes à une partition en Q-1 classes

  $\Rightarrow$ inertie intra-classe va augmenter
  - agrégation par inertie : on choisit les 2 classes p et q (de centre de gravité g et d'effectif l) à agréger de façon à minimiser l'accroissement d'inertie intra-classe $\Delta(p,q) = \frac{l_p l_q}{l_p + l_q} d^2(g_p, g_q)$

Méthode de Ward

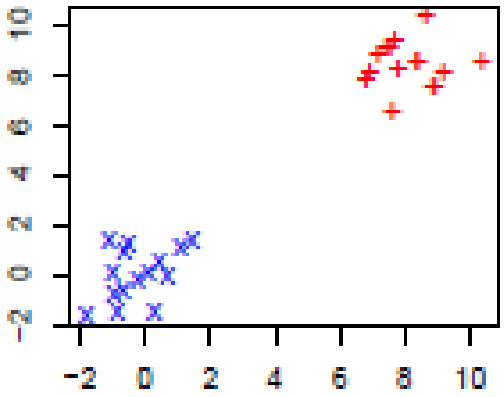
- Initialisation : 1 classe =1 individu

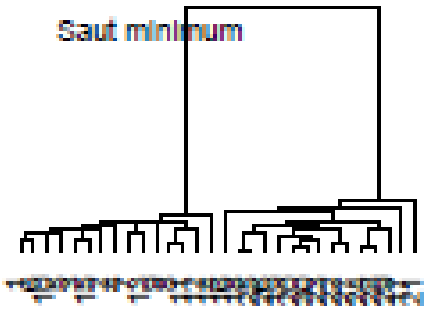
$\Rightarrow$ Inertie inter= Inertie totale

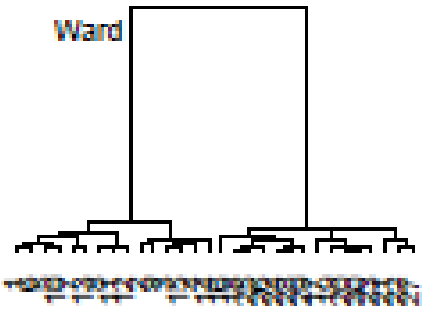
- A chaque étape, les individus sont répartis en Q classes obtenues par les étapes précédents
  - comment choisir les 2 classes (parmi les Q) à agréger ?
  - en agrégeant 2 classes, on passe d'une partition en Q classes à une partition en Q-1 classes

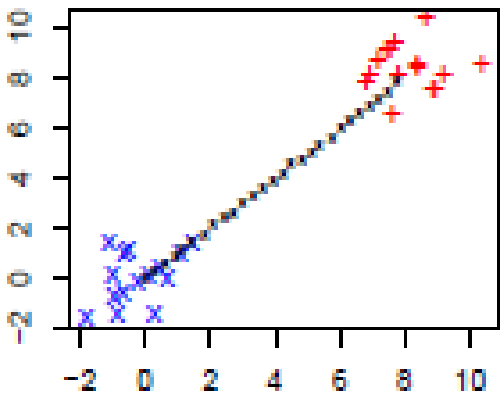
  $\Rightarrow$ inertie intra-classe va augmenter
  - agrégation par inertie : on choisit les 2 classes p et q (de centre de gravité g et d'effectif l) à agréger de façon à minimiser l'accroissement d'inertie intra-classe $\Delta(p,q) = \frac{l_p l_q}{l_p+l_q} d^2(g_p, g_q)$
  - choisir les classes p et q telles que $\Delta(p,q)$ soit minimum revient à choisir :
    - des classes dont les centres de gravité sont proches ($d^2(g_p, g_q)$ petit)
    - des classes d'effectifs faibles ($\frac{l_p l_q}{l_p+l_q}$ petit)

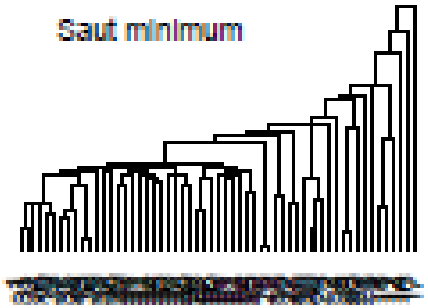
Méthode de Ward

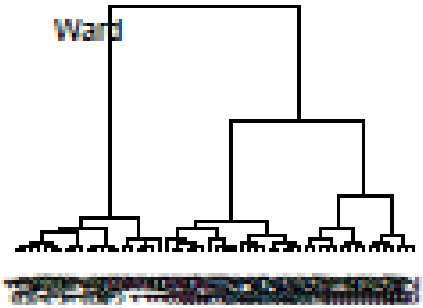












- regroupe les objets de faible poids et évite l'effet de chaîne
- regroupe des classes ayant des centres de gravité proches

$\Rightarrow$ intérêt pour la classification

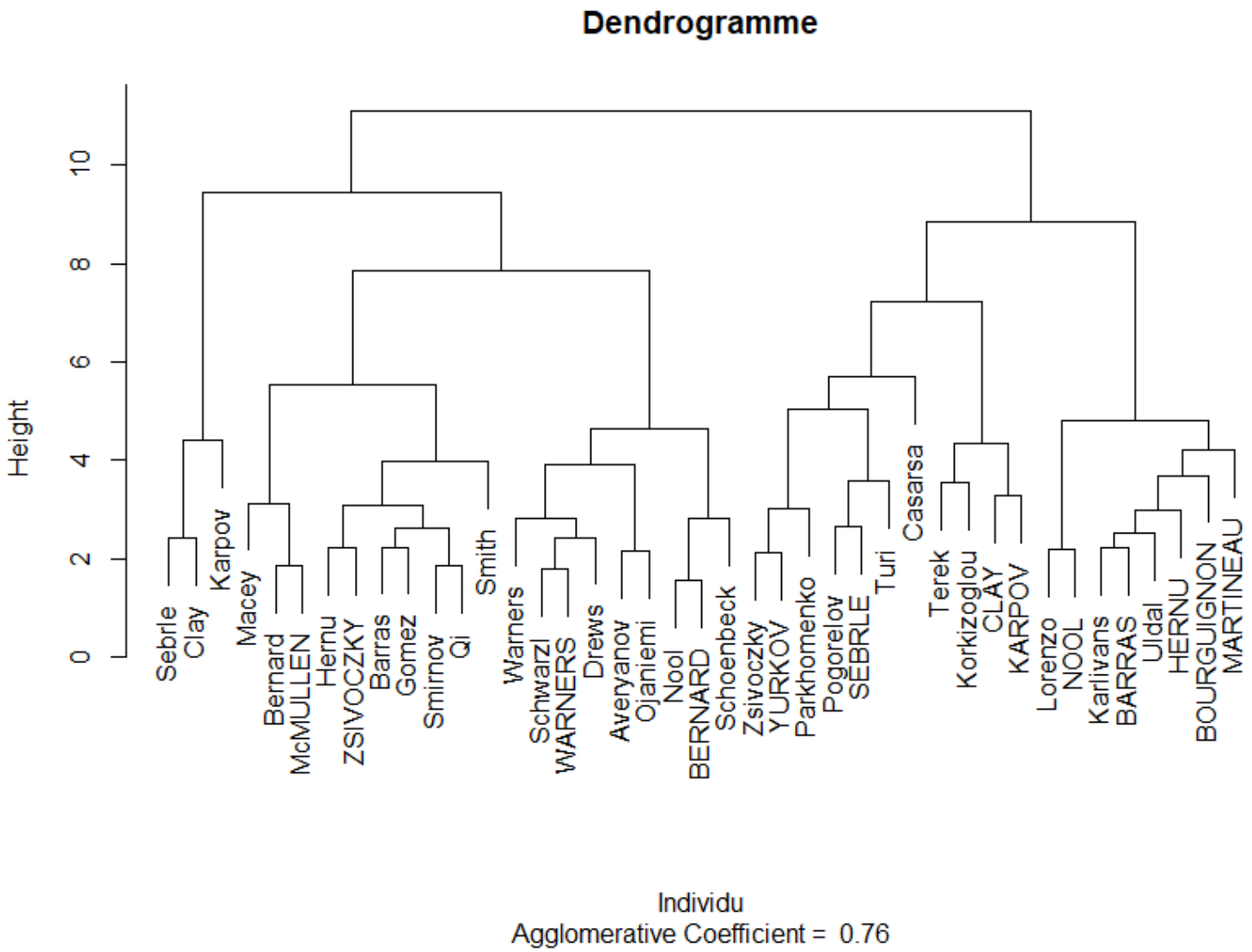
Exemple

- jeu de données : athlètes participant au décathlon (10 var de performance)
- distance euclidienne et indice d'agrégation de Ward

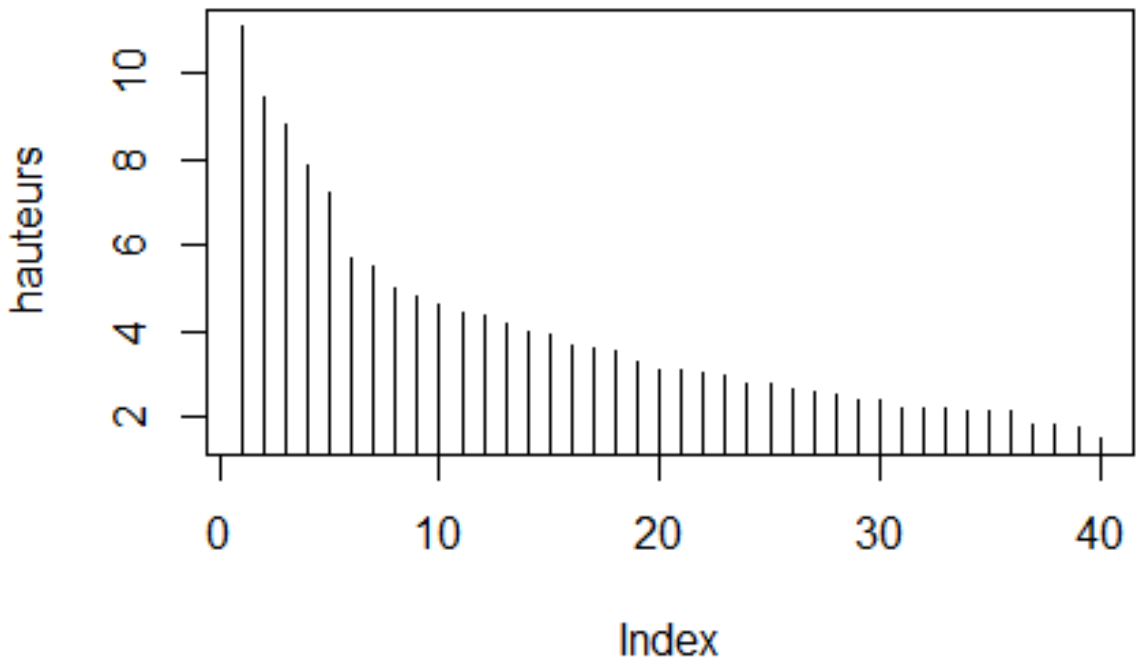
Exemple

- jeu de données : athlètes participant au décathlon (10 var de performance)
- distance euclidienne et indice d'agrégation de Ward
- Etapes
  1. importer les données
  2. standardiser ou non les données : indispensable ici car les variables ont des unités différentes
  3. construire la CAH
  4. couper l'arbre de classification
  5. caractériser les exemples

Exemple : Dendogramme



Exemple : Hauteur de coupe



- saut entre le 5ème et le 6ème baton -> 6 groupes
- saut entre le 3ème et le 4ème baton -> 4 classes

$\Rightarrow$ $k = 4$ pour avoir un effectif suffisant dans chacune des classes

CAH en grandes dimensions

- Si beaucoup de variables : faire une ACP et ne conserver que les premières dimensions $\Rightarrow$ on se ramène au cas classique

CAH en grandes dimensions

- Si beaucoup de variables : faire une ACP et ne conserver que les premières dimensions $\Rightarrow$ on se ramène au cas classique
- Si beaucoup d'individus : algorithme de CAH trop long

CAH en grandes dimensions

- Si beaucoup de variables : faire une ACP et ne conserver que les premières dimensions $\Rightarrow$ on se ramène au cas classique
- Si beaucoup d'individus : algorithme de CAH trop long
  - Faire une partition (par K-means) en une centaine de classes
  - Construire la CAH à partir des classes (utiliser l'effectif des classes dans le calcul)