01-03
Regroupement hiérarchique

Sommaire

Sommaire

Rappel des principaux types de partitionnement

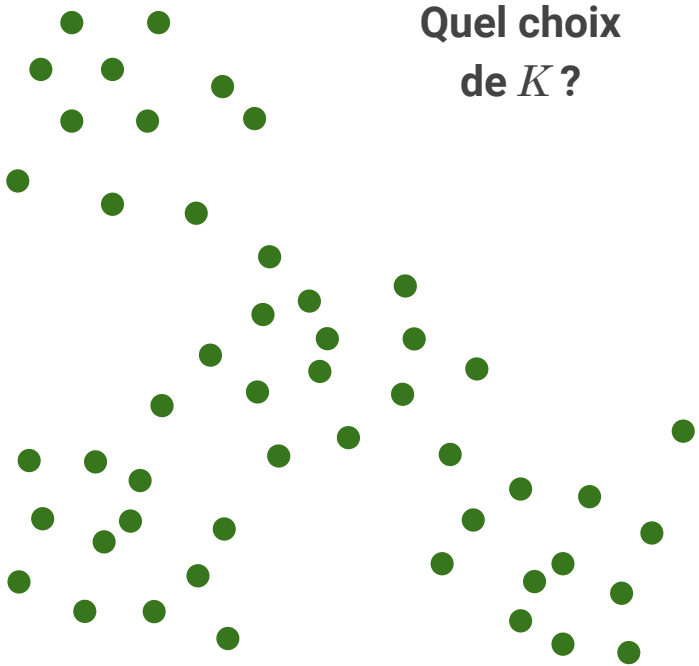
- Partitionnement basé sur
  - les centroïdes (K-moyennes, CURE, …)
  - la connectivité (hiérarchique, ...)
  - la distribution (BFR, …)
  - la densité (DBSCAN, OPTICS, …)
  - les grilles
- Et d’autres ….

Rappel des principaux types de partitionnement

- Partitionnement basé sur
  - les centroïdes (K-moyennes, CURE, …)
  - la connectivité (**hiérarchique**, ...) 👈 On parle alors de **regroupement hiérarchique**
  - la distribution (BFR, …)
  - la densité (DBSCAN, OPTICS, …)
  - les grilles
- Et d’autres ….

K-moyennes, choix de K

- Le partitionnement basé sur les *K*-moyennes nécessite de **spécifier le nombre de clusters K**
- Ceci est un **inconvénient** car il n'existe pas de méthode "universelle" et robuste pour le **choix** de K
- Le **regroupement hiérarchique** est une technique de partitionnement alternative à l'algorithme des K-moyennes ne nécessitant pas le choix préalable du nombre de clusters

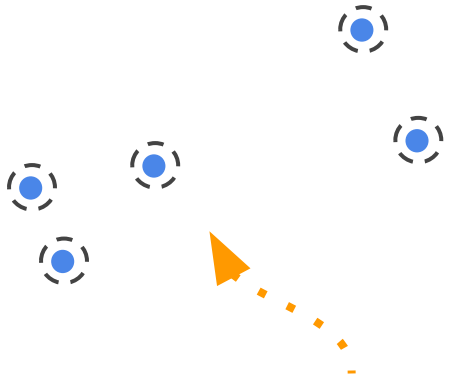


Regroupement hiérarchique

- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)

Regroupement hiérarchique

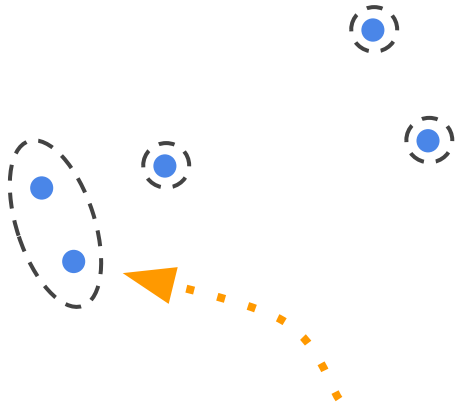
- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)



À l'état initial, chaque observation (feuille) est son propre cluster

Regroupement hiérarchique

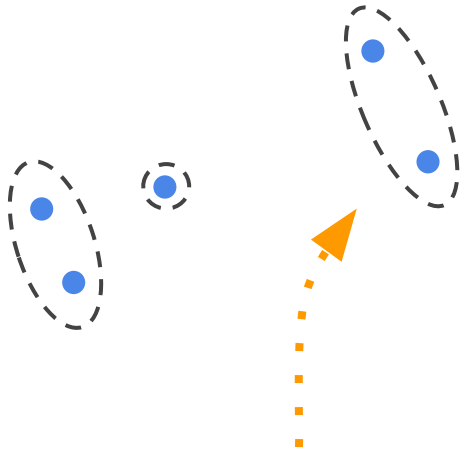
- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)



Les deux observations / clusters les plus proches sont regroupées en un cluster

Regroupement hiérarchique

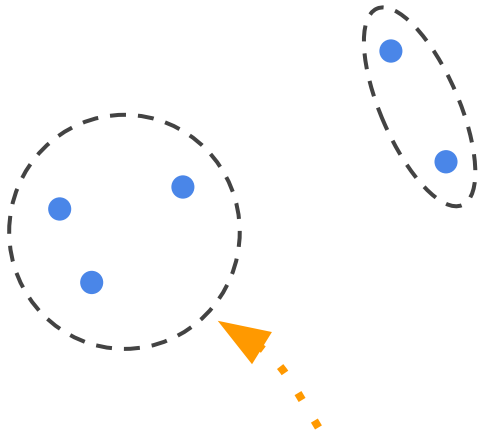
- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)



Les deux clusters les plus proches sont regroupées en un cluster

Regroupement hiérarchique

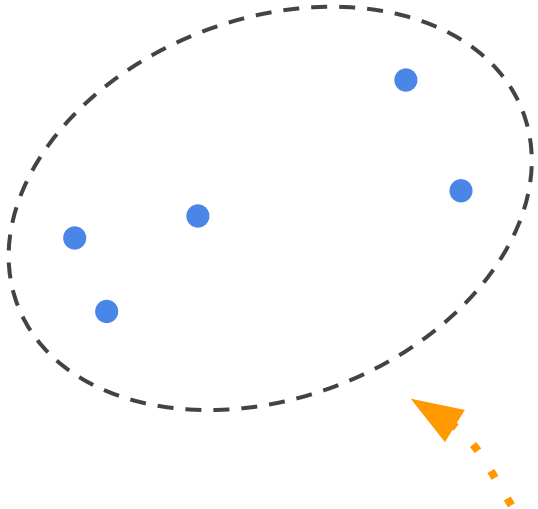
- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)



Les deux clusters les plus proches sont regroupées en un cluster

Regroupement hiérarchique

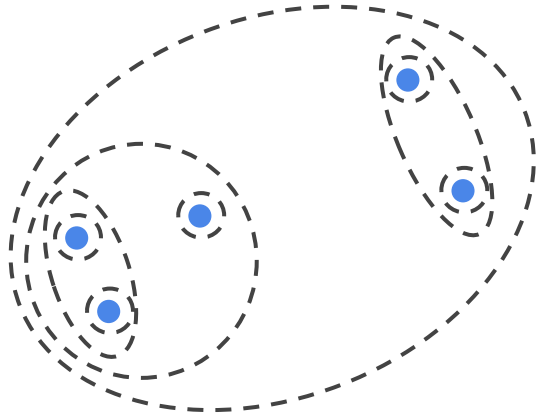
- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)



Au final, un seul cluster (racine) contenant toutes les observations est obtenu

Regroupement hiérarchique

- Nous allons décrire ici le regroupement hiérarchique de type **agglomératif** (ou ascendant / bottom-up)
- Il s'agit du type le plus **courant** de regroupement hiérarchique
- Réfère au fait qu'un **dendrogramme** est construit à partir des observations (**feuilles**) en combinant successivement les clusters obtenus jusqu'à la **racine** (arbre / hiérarchie)



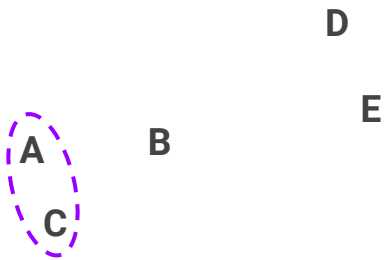
Les regroupements successifs permettent d'obtenir le **dendrogramme**

Sommaire

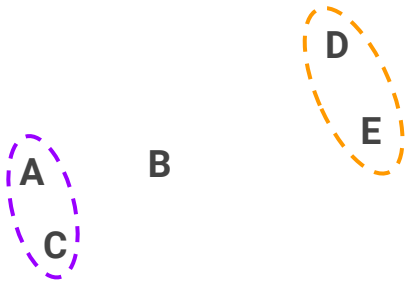
Dendrogramme

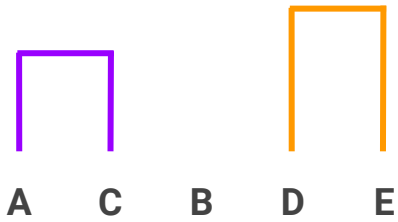
Clusters initiaux: **A** **C** **B** **D** **E**

Dendrogramme

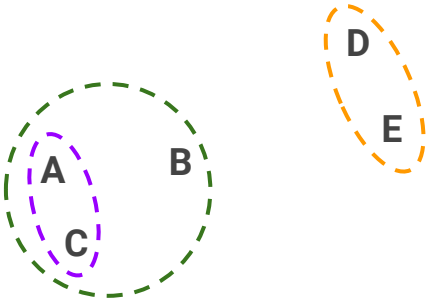


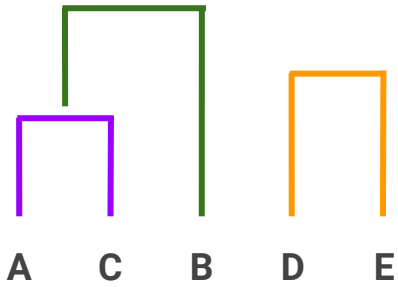
Dendrogramme



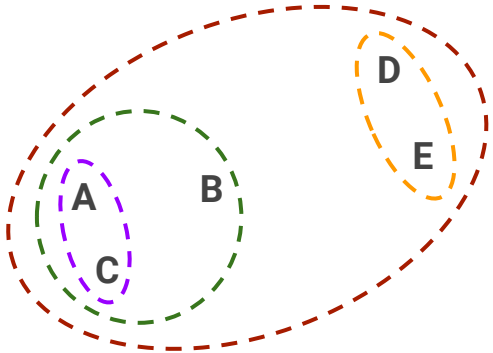


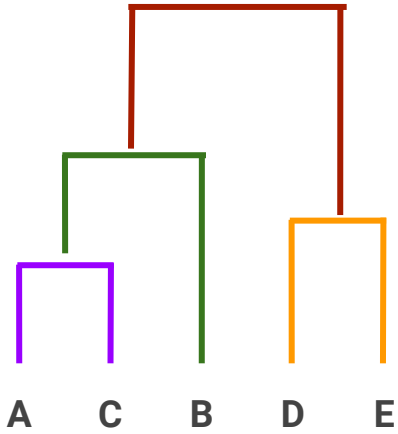
Dendrogramme





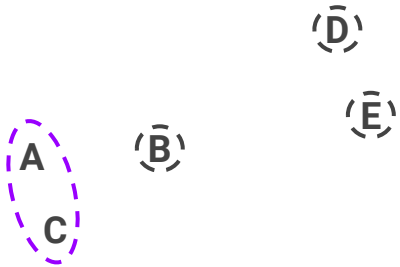
Dendrogramme

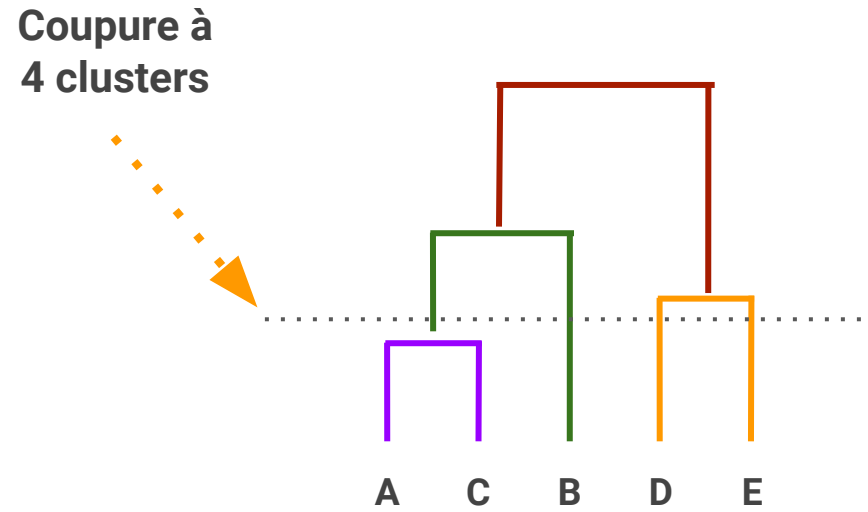




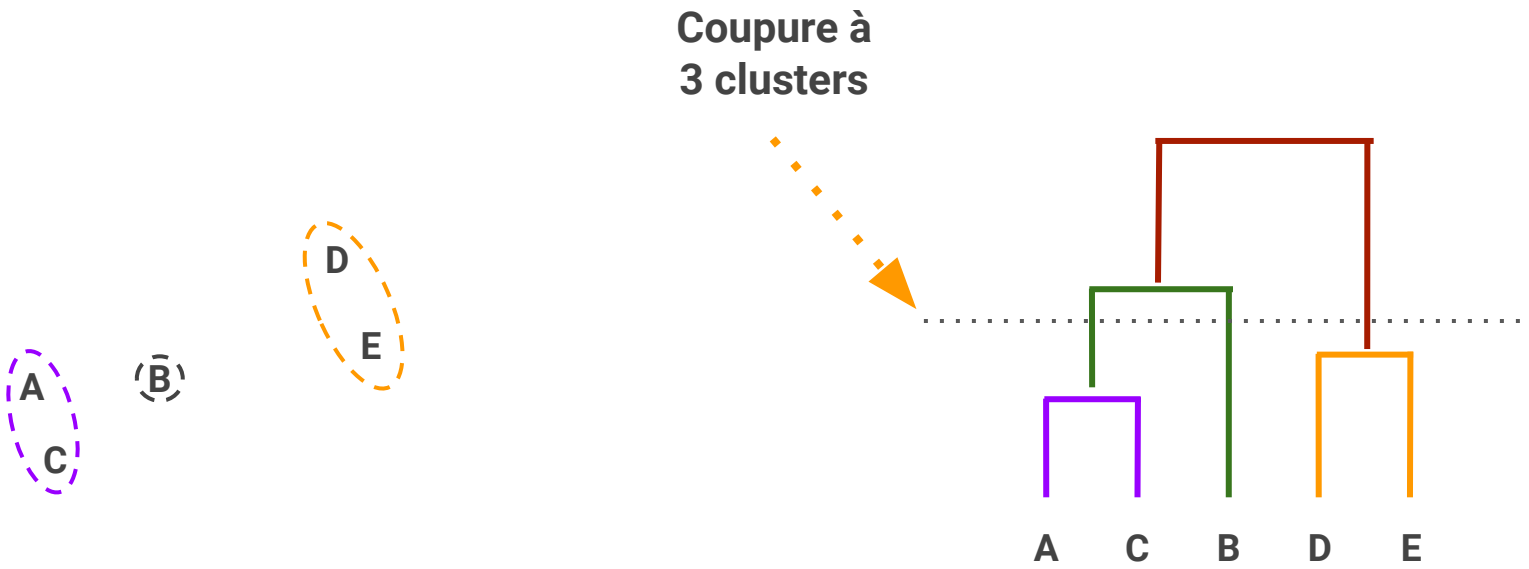
Le dendrogramme complet est obtenu lorsque toutes les observations appartiennent au même cluster

Coupure du dendrogramme

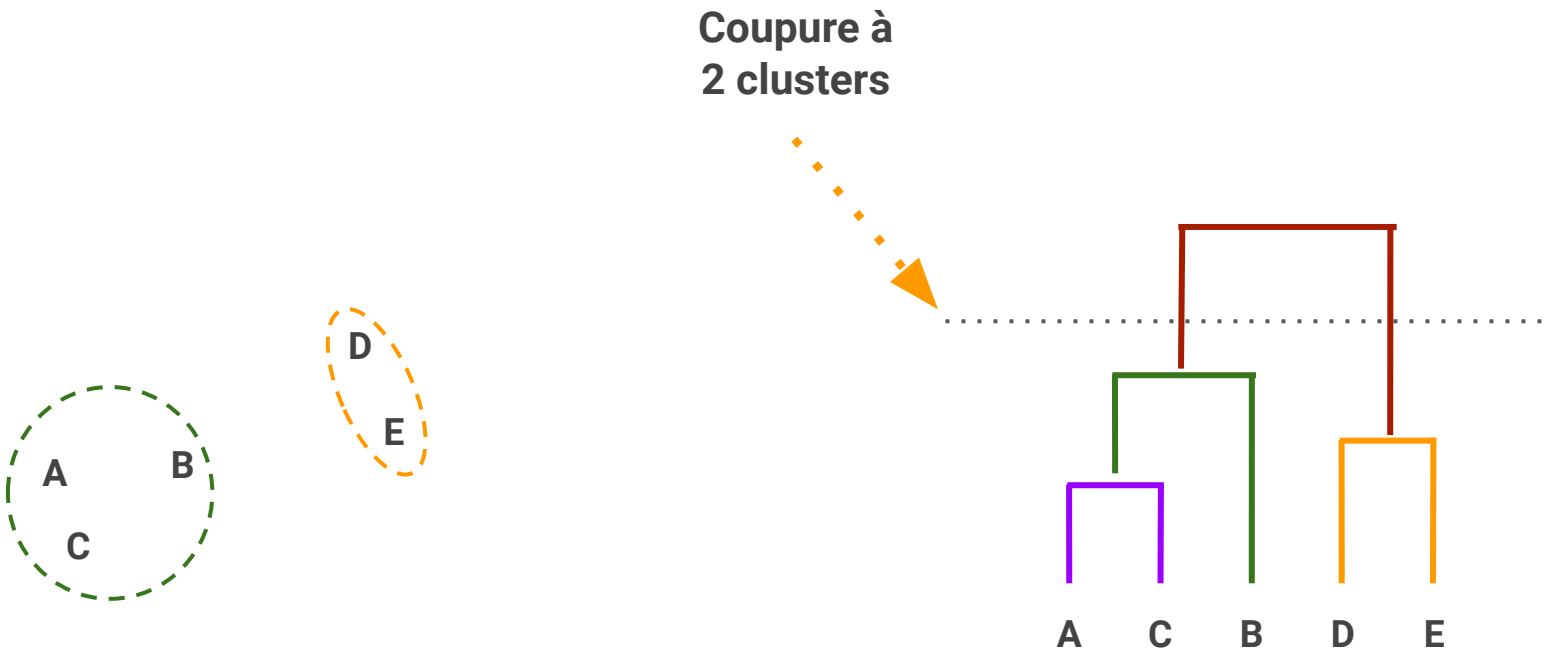




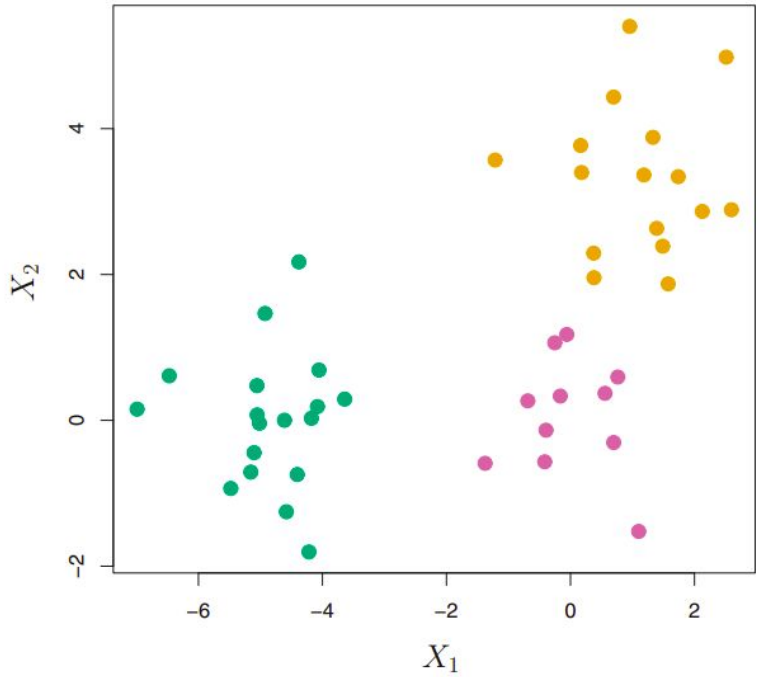
Coupure du dendrogramme



Coupure du dendrogramme

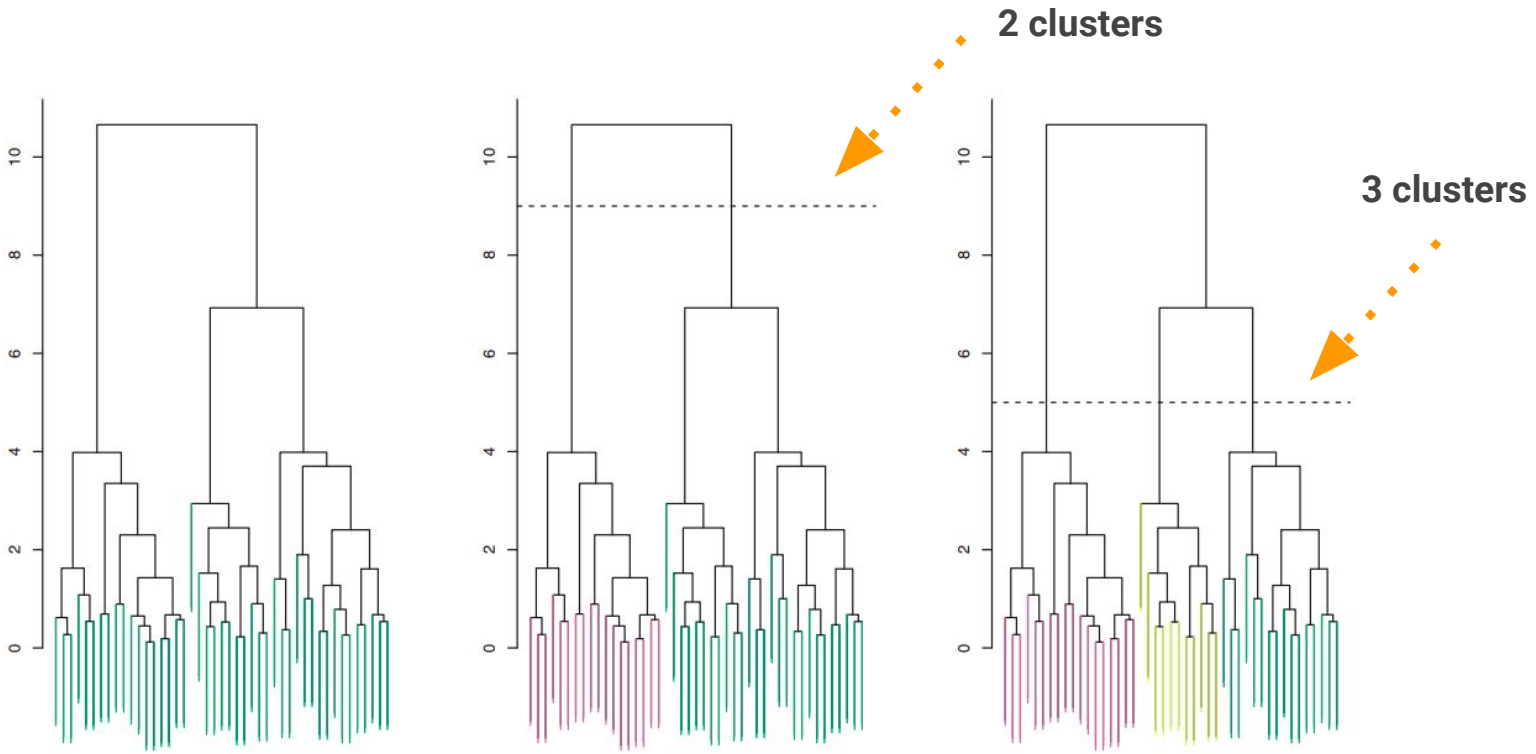


Exemple avec des données simulées



Extrait de [2] - 45 observations simulées dans une espace 2D

Exemple avec des données simulées



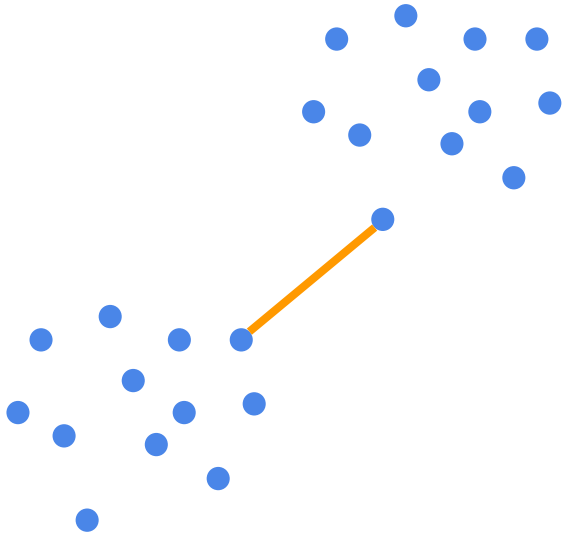
Extrait de [2] - Dendrogramme obtenu par regroupement hiérarchique (saut maximum) et seuil de coupure

Sommaire

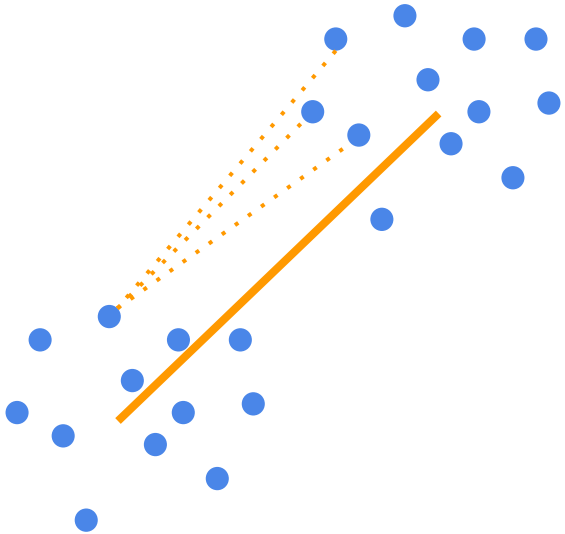
Mesure de dissimilarité

- La mesure de dissimilarité la plus souvent rencontrée est la **distance euclidienne**
- Il y a une analogie forte entre **distance** et **dissimilarité**. En effet
  - Des observations proches dans l'espace sont considérés plus similaires que des observations éloignées
  - Plus la distance augmente, plus les observations diffèrent, et donc, plus la dissimilarité augmente
- Il existe de nombreuses mesures de dissimilarité, qui seront étudiées au fur et à mesure du cours
- Plusieurs critères peuvent être utilisés pour calculer la dissimilarité: **saut minimum**, **saut maximum**, **lien moyen** et **distance de Ward**

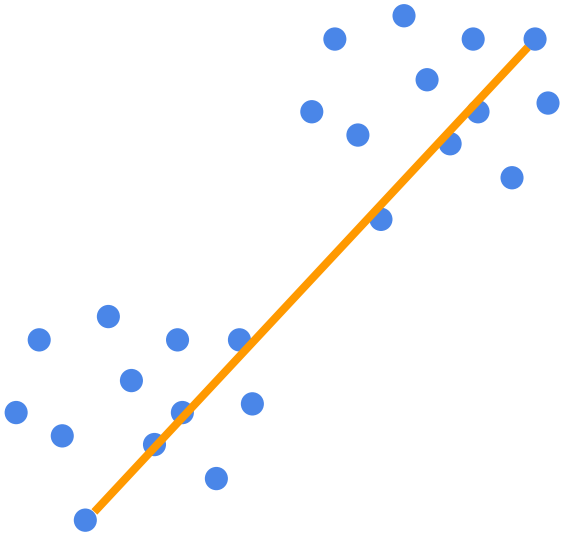
Saut minimum, lien moyen et saut maximum



Saut minimum
(simple link)



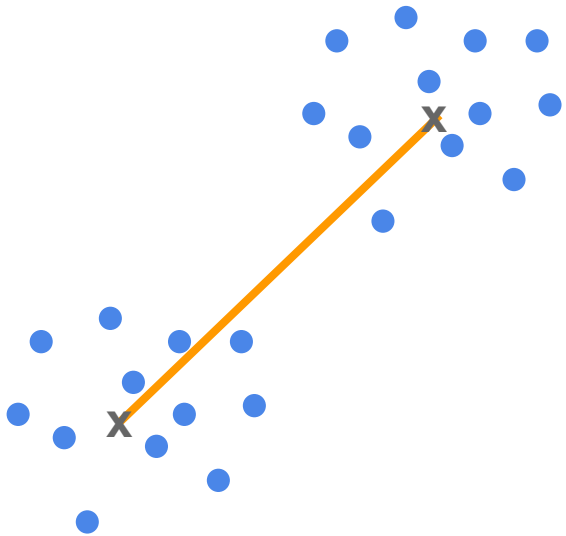
Lien moyen
(average link)



Saut maximum
(complete link)

Distance de Ward (ou centroïde)

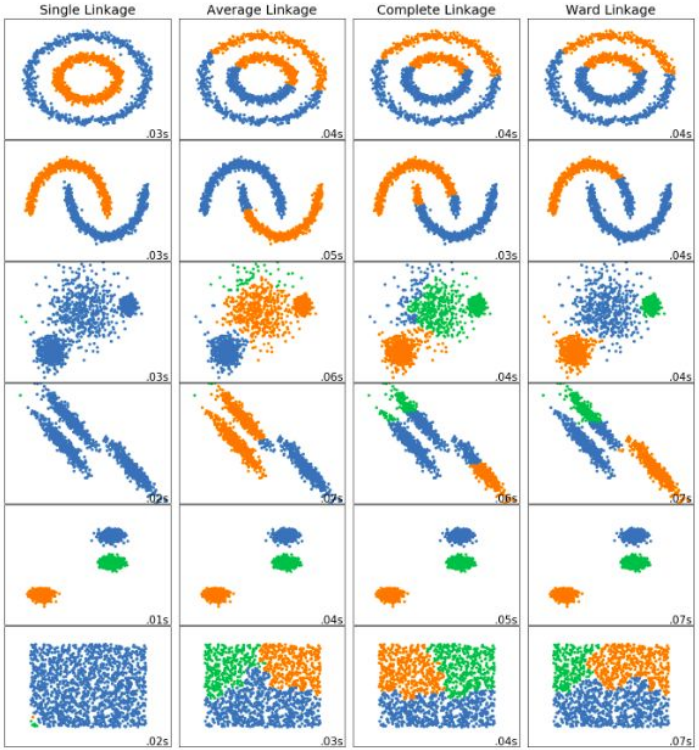
- La **distance de Ward** minimise la somme des carrés des différences à l'intérieur d'un cluster lors de la fusion
- Approche similaire à la fonction de coût de l'algorithme des K-moyennes
- **Applicable aux espaces euclidiens seulement**



Dissimilarité selon les critères (1/2)

Lien	Dissimilarité
Saut maximum	Maximale inter-clusters
Saut minimum	Minimale inter-clusters
Lien moyen	Moyenne inter-clusters
Ward / Centroïde	Entre les centroïdes de deux clusters

Dissimilarité selon les critères (2/2)



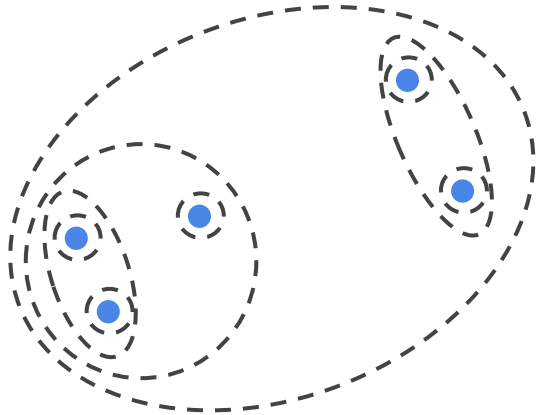
https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

Sommaire

Algorithme de regroupement hiérarchique

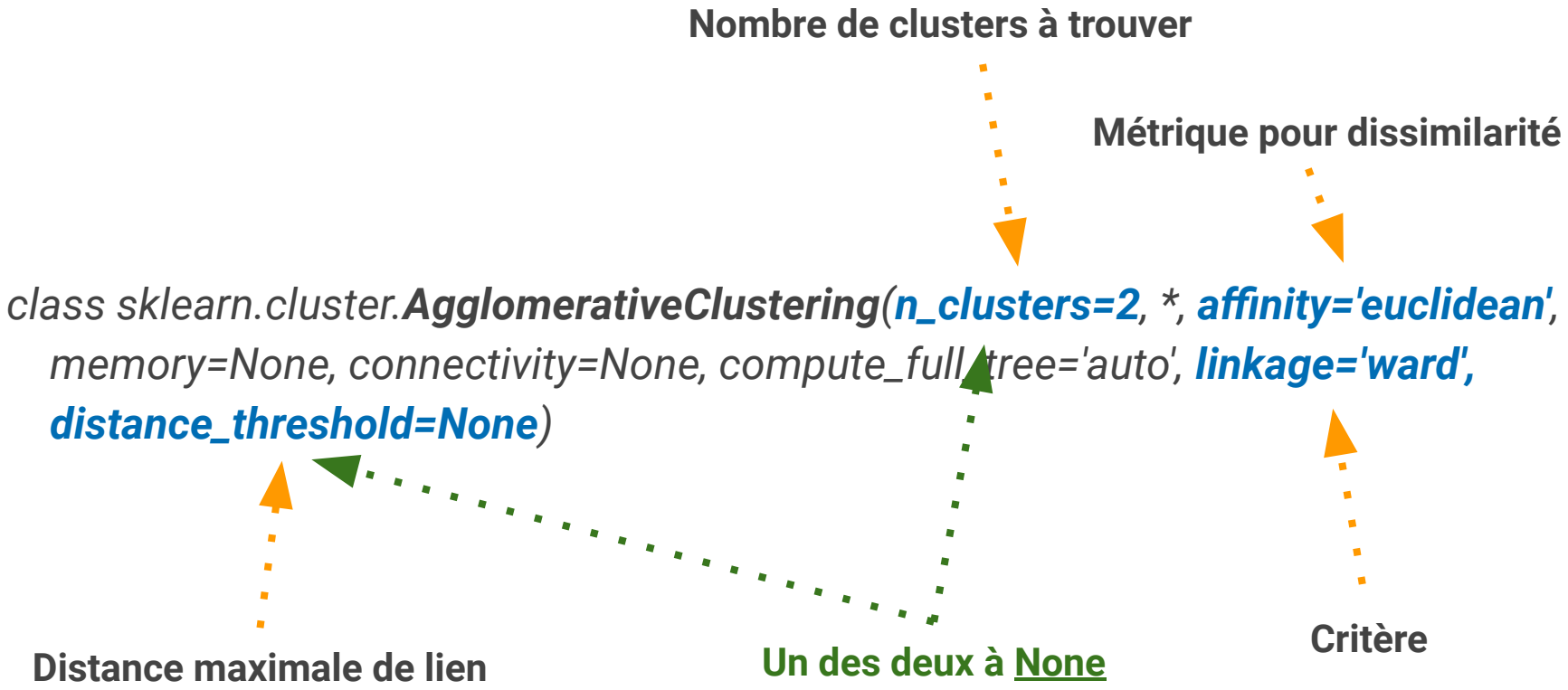
- Comparativement aux K-moyennes, l'algorithme de regroupement hiérarchique est relativement simple. Voici le pseudo-code

```
Démarrer avec un cluster par
observation

Répéter {

    Fusionner les deux clusters
    les plus similaires

} jusqu'à ce que toutes les
observations appartiennent à un
seul et même cluster
```



Sommaire

AgglomerativeClustering (scikit-learn 0.24.2)



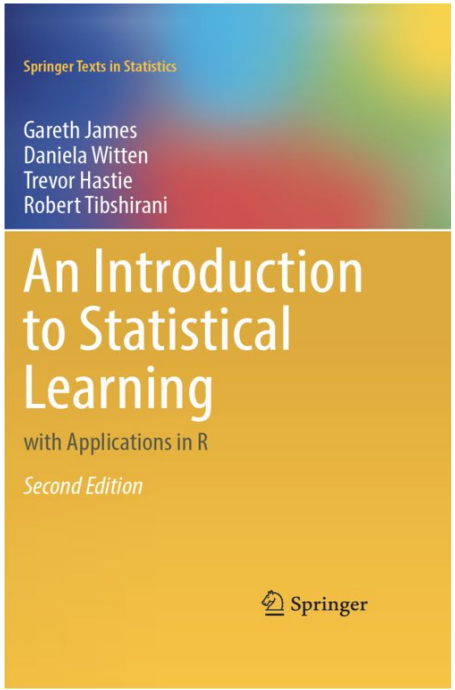
Sommaire

https://github.com/mswawola-cegep/420-a58-sf.git

01-03

Sommaire

Lectures



- Introduction to Statistical Learning with Applications in R Second edition (2021)

  → 12.4 Clustering Methods

Références

[1] CS229: Machine Learning - Stanford University

[2] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, “Introduction to Statistical Learning with Applications in R - Second edition”