IA et science des données

Cours 9 – mardi 29 mars 2022
Clustering

Christophe Marsala
Vincent Guigue

Sorbonne Université

LU3IN026 - 2021-2022

Plan du cours

Objectif du projet : analyser un jeu de données

- Résultat attendu
  - définition d'un certain nombre de problématiques et leur résolution
  - au moins un problème d'apprentissage supervisé et un autre non-supervisé
- Compte-rendu
  - un unique notebook complété par un package avec vos fonctions
  - un poster (électronique) expliquant de façon synthétique en une page les différentes expériences réalisées et leurs résultats
- Calendrier (groupe de TDTME)
  - jeu de données mis en ligne dans les jours qui viennent
  - séance 10 : prévue (en partie) pour travailler sur le projet
  - séance 11 : soutenance et rendu

Plan du cours

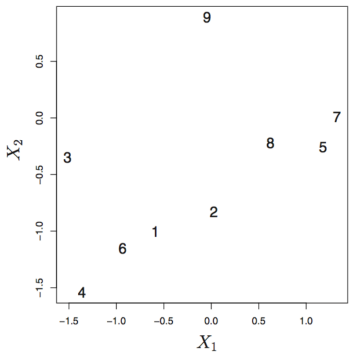
Rappels

- Classification : trouver des classes de descriptions
- Un ensemble de données sans classe connue
  - on recherche à faire des regroupements de descriptions similaires
  - on souhaite mettre en évidence des classes, des catégories
- But : former des groupes de données qui se ressemblent
  - clustering : faire des groupes parmi les données
  - cluster : ensemble de données regroupées ensemble
- Exemple :
  - le clustering hiérarchique
  - l'algorithme des K-moyennes

Mesurer la distance entre 2 clusters

- Utiliser une distance entre 2 exemples : $d(x_1, x_2)$
  - Euclidienne, Manhattan, Minkowski, "infinie",...
  - étape de normalisation nécessaire
- Distances entre 2 clusters : $dist(A, B)$
  - A) complete linkage
  - B) average linkage
  - C) simple linkage
  - D) centroid linkage
- Centre de gravité (centroid) d'un cluster

Exemple : méthode par agglomération



(source : "An introduction to statistical learning", G. James, D. Witten, T. Hastie, R. Tibshirani)

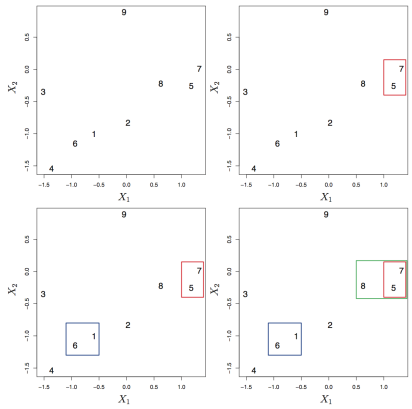
Algorithme : clustering hiérarchique (version ascendante)

- Soit $\mathbb{E}$ un ensemble d'éléments (exemple ou groupe d'exemples)

1. calculer les distances entre chaque élément de l'ensemble
2. fusionner en un seul groupe les 2 éléments les plus proches : ce groupe remplace les 2 éléments dans l'ensemble $\mathbb{E}$
3. recommencer en 1) jusqu'à ce qu'il ne reste qu'un seul groupe unique dans $\mathbb{E}$

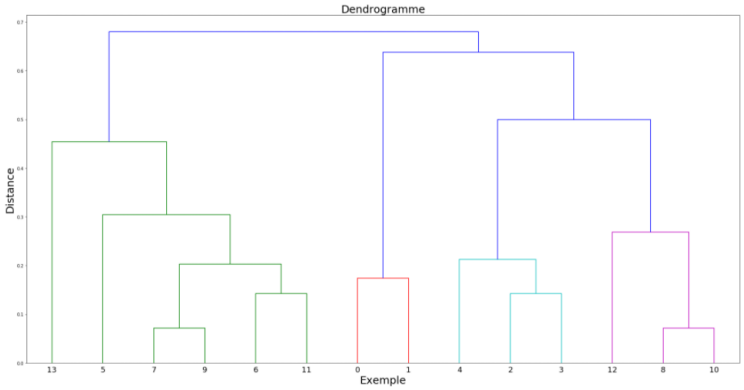
- Au départ : $\mathbb{E}$ est initialisé avec $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - chaque exemple forme un groupe à lui tout seul
- Au final : $\mathbb{E}$ contient un groupe avec tous les exemples de $\mathbf{X}$
- Cet algorithme permet de construire un dendrogramme

Exemple : méthode par agglomération



(source : "An introduction to statistical learning", G. James, D. Witten, T. Hastie, R. Tibshirani)

Exemple de dendrogramme final



Conclusion sur le clustering hiérarchique

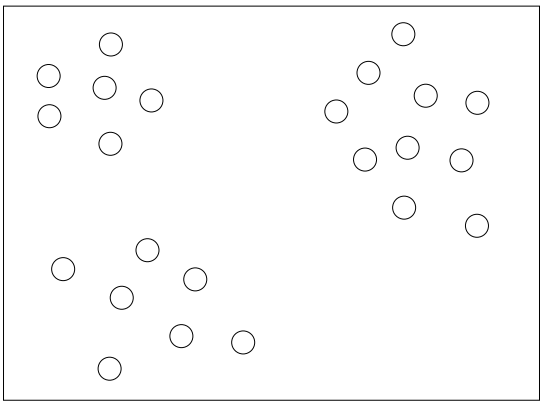
- Algorithme très efficace sur des jeux de données assez réduit, sinon ça devient vite peu lisible
- Le nombre de classes à trouver n'est pas défini : il est estimé par l'étude du dendrogramme
- Les calculs sont très coûteux ! ($\geq o(n^2)$)

Algorithme K moyennes (ou K-means)

- Un des algorithmes de clustering le plus courant
- Idée : *ceux qui se ressemblent, s'assemblent*
  - trouver des clusters qui séparent les données de façon équitable
    - les clusters seront repérés par leur centre
- Mise en œuvre
  - choix du nombre de clusters à trouver : $K > 0$, entier naturel
  - mesure de la proximité entre données : mesure de distance
    - par exemple, distance euclidienne entre leurs descriptions
  - choisir K centres de clusters et affecter les données au cluster qui leur est le plus proche
  - modifier les K centres en fonction des données qui sont dans leur cluster

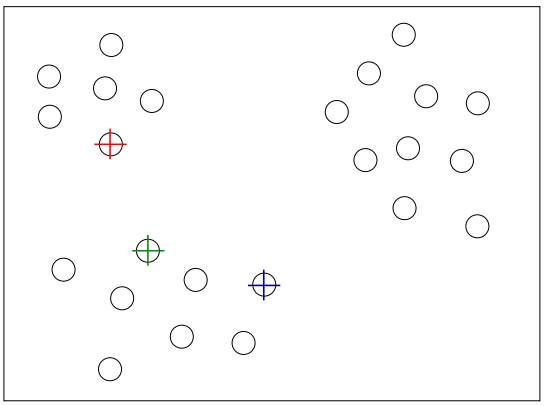
Un petit exemple

- Un ensemble de données quelconque



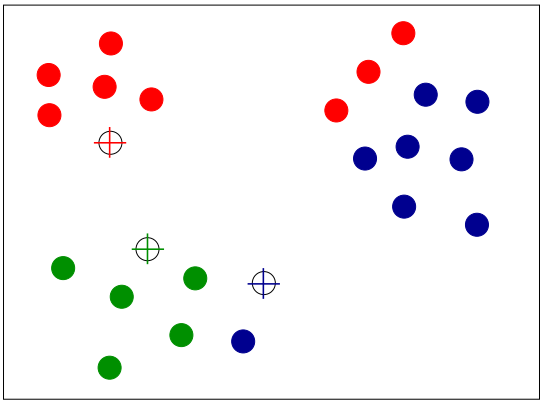
Un petit exemple

- Choix aléatoire de centres de clusters (ici $K = 3$)



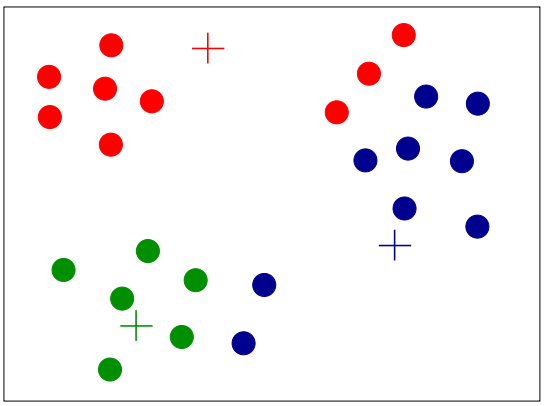
Un petit exemple

- Affectation des données : utilisation des médiatrices



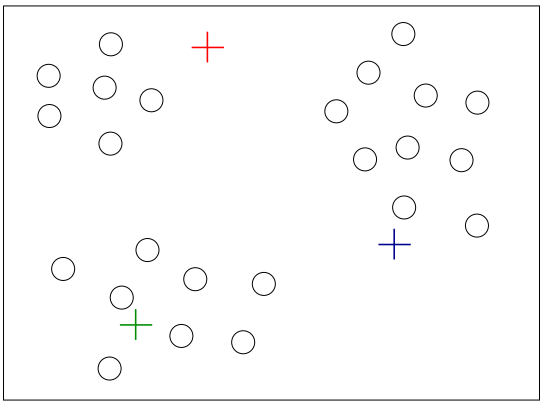
Un petit exemple

- Mise à jour des centres



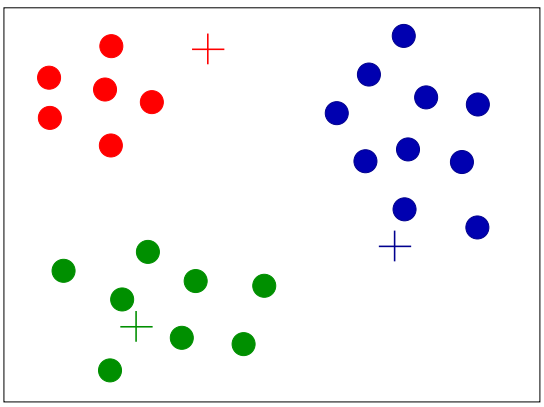
Un petit exemple

- Il faut refaire l'affectation des données



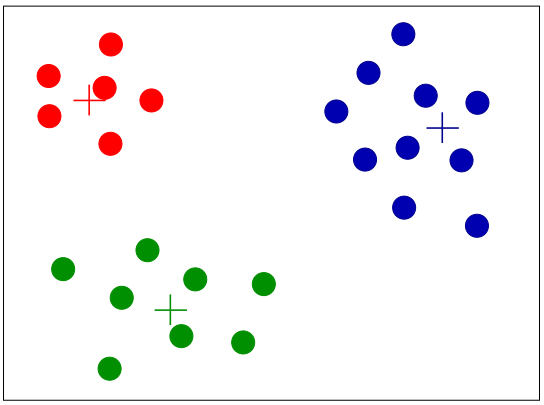
Un petit exemple

- Affectation des données aux centres les plus proches



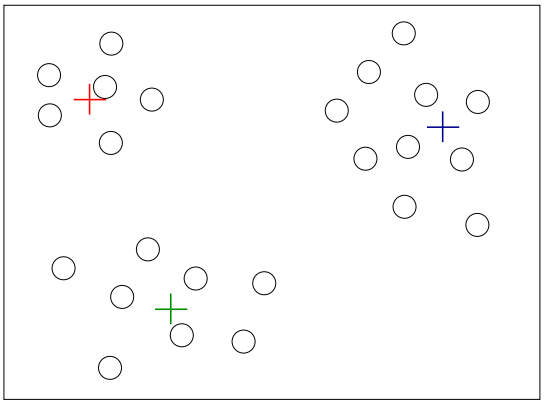
Un petit exemple

- Nouvelle mise à jour des centres



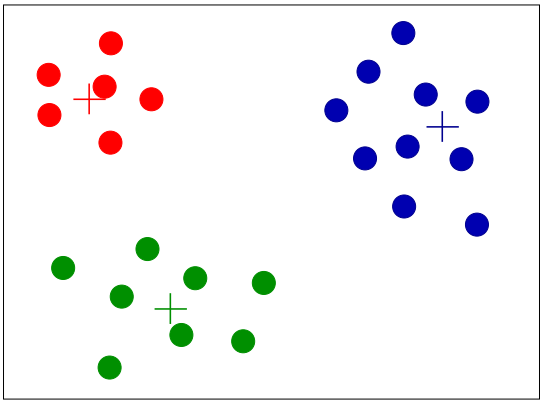
Un petit exemple

- Il faut refaire l'affectation des données



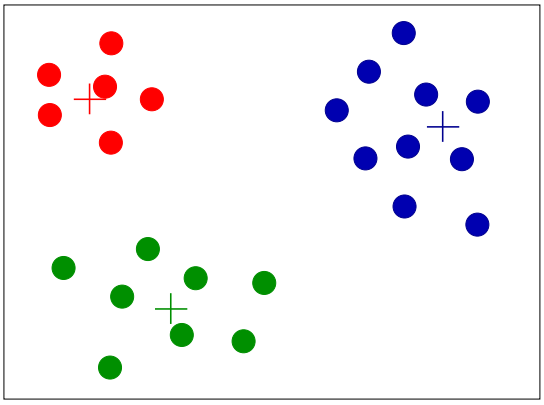
Un petit exemple

- Affectation des données aux centres les plus proches



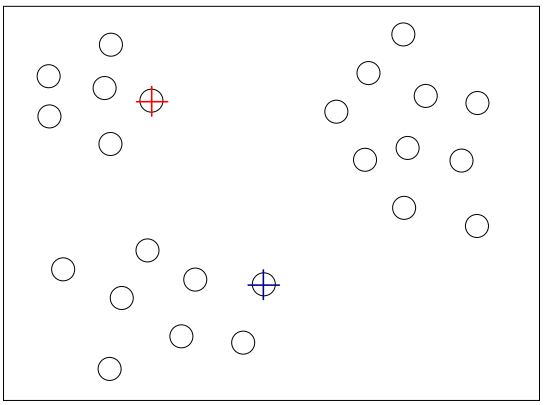
Un petit exemple

- Convergence de l'algorithme : les centres ne changent pas



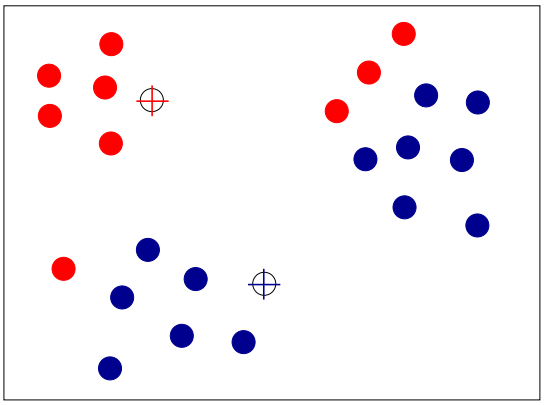
Un autre exemple

- Choix aléatoire de centres de clusters (ici $K = 2$)



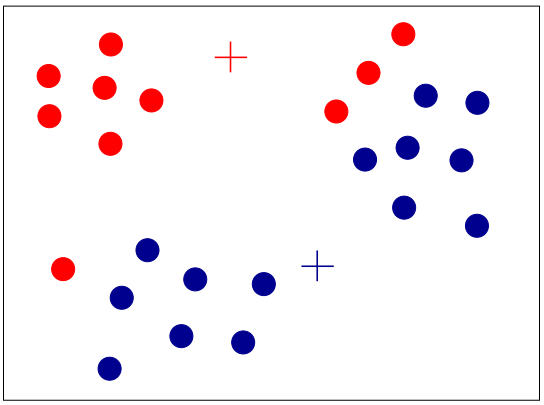
Un autre exemple

- Affectation des données aux centres les plus proches



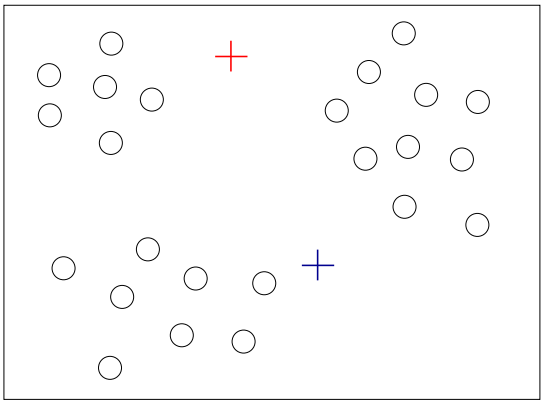
Un autre exemple

- Mise à jour des centres



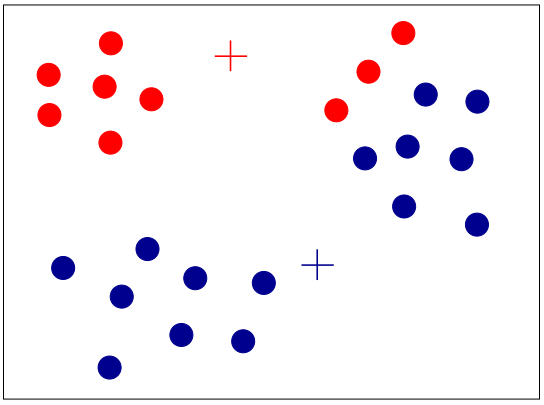
Un autre exemple

- Il faut refaire l'affectation des données



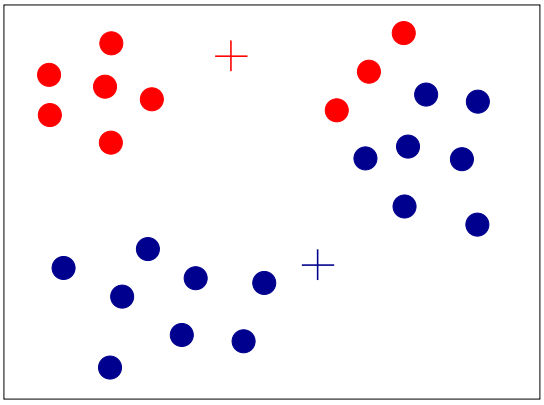
Un autre exemple

- Affectation des données aux centres les plus proches



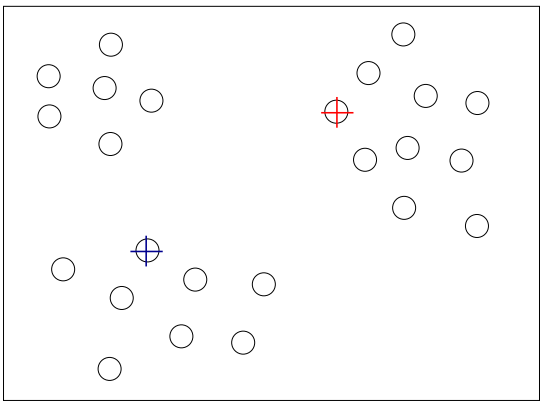
Un autre exemple

- Mise à jour des centres : ils ne changent pas



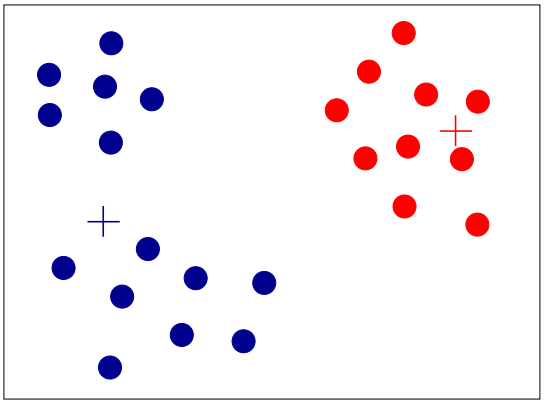
Un dernier exemple

- Autre choix des centres de clusters initiaux (toujours $K = 2$)



Un dernier exemple

- Clusters trouvés



Clusters différents au final

- Le choix initial des centres est important ! (ici avec $K = 2$)