Apprentissage automatique non supervisé

Amira Barhoumi

amira.barhoumi@univ-grenoble-alpes.fr

Année universitaire : 2025-2026

Apprentissage non supervisé

- Données décrites par un ensemble de descripteurs $\{var_1, var_2, ..., var_m\}$
- Données non-étiquetées sans classes connues à priori

$$D = \{obj_1, obj_2, \ldots, obj_N \text{ ; avec } \forall\ i \in \{1, \ldots, N\}, obj_i : \text{une observation}\}$$

- Chercher d'une division de ces données en catégories (groupes)
- Chercher une partition $\pi = \{D_1, D_2, \ldots, D_k\}$ du corpus D tels que

$$D = \bigcup D_j;\ \forall\ j \in \{1, \ldots, k\} \quad \text{et} \quad D_i \cap D_j = \varnothing$$

- Trouver la meilleure partition en fonction d'un critère de similarité sur D

	var_1	…	var_m
obj_1			
…			
obj_N			

Identification des catégories

	var_1	…	var_m
new			

Quelle catégorie**?**

- Approches de *clustering*
  - **Algorithmes de partitionnement** : construire plusieurs partitions puis les évaluer selon certains critères

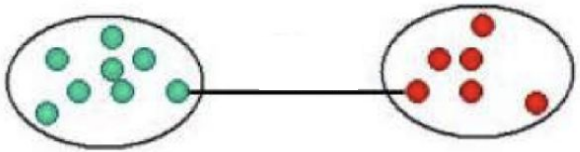
    exemple : k-means, k-medoids
  - **Algorithmes hiérarchiques** : créer une décomposition hiérarchique des objets selon certains critères

    exemple : Classification ascendante hiérarchique
  - **Algorithmes basées sur les grilles** : diviser l'espace des objets dans une grille et estimer la densité

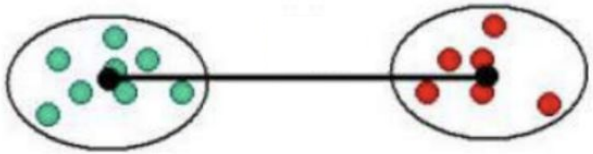
    exemple : CLIQUE, STING

- Représentation d'un cluster :
  - Le centroïde : le vecteur moyen des objets constituant le cluster. Il peut correspondre à un élément du cluster
  - Le médoïde (médiane) : un des objets du cluster qui est proche de la notion du centroïde

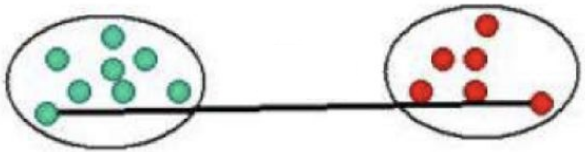
- Mesures de similarité entre 2 clusters
  - Lien simple (*single linkage*) : la plus petite distance entre les objets des 2 clusters
  - Lien moyen (*average linkage*) : la distance moyenne entre les objets des 2 clusters
  - Lien complet (*complete linkage*) : la plus grande distance entre tous les objets des 2 clusters



Lien simple



Lien moyen



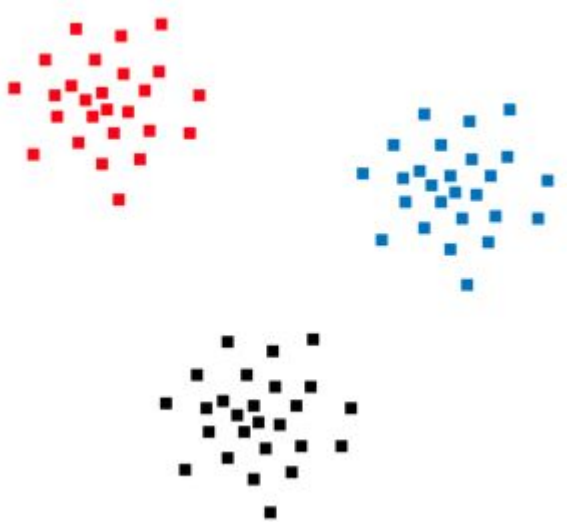
Lien complet

Apprentissage non supervisé

- Algorithmes de partitionnement
  - Partitionner un corpus de N objets en K clusters
  - Méthodes heuristiques :
    - **k-means** [MacQueen, 1967] : chaque cluster est représenté par son centre
    - **k-médoïdes** (*Partition around medoids*) [Kaufman & Rousseeuw, 1987] : chaque cluster est représenté par un de ses objets

K-means

- Méthode d'apprentissage non supervisé
- Données (objets) non étiquetées
- Très utilisé pour la classification de documents, la segmentation d'images, la segmentation marketing
- Modèle de classification (*clustering*)
- Partitionnement en groupes (*clusters*)
- Mise en clusters consiste à
  - regrouper les objets similaires
  - séparer les objets dissimilaires

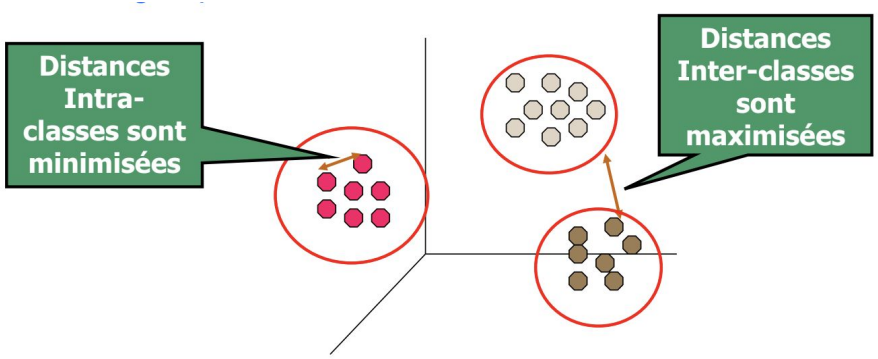


K-means

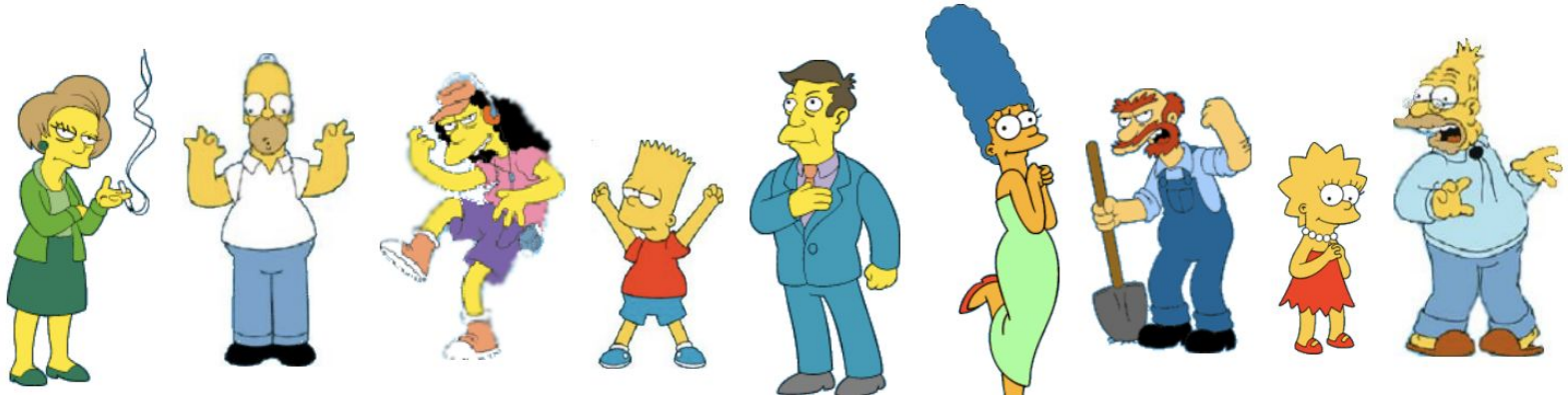
- Exploration des données
  - connaissance des données
  - connaissance de la nature des groupes recherchés (prototype par groupe)

- k-means = regrouper les données en **k** groupes
  - Maximiser la similarité entre objets à l'intérieur d'un groupe
  - Minimiser la similarité entre les objets des groupes

- Un individu appartient à un seul cluster



K-means



Objectif : partitionner cette population en 2 groupes (k=2)

Quels sont les éventuels groupes?

K-means

- Algorithme de k-means :

  1- Choisir k objets formant ainsi k clusters C_i; quelque soit i entier naturel $0 < i < k+1$

  2- Affecter chaque objet O au cluster C_i de centre M_i telle que la distance en O et M_i est minimale

  3- Calculer M_i de chaque cluster C_i

  4- Répéter les étapes 2- et 3-

  5- Arrêt si pas de changement

K-means

- Exemple 1 :

Soit le corpus D={1, 2, 3, 6, 7, 8, 13, 15, 17}

Appliquer l'algorithme k-means pour partitionner l'ensemble D en 3 clusters

Initialisation : prendre 3 objets au hasard 1, 2 et 3 M_1=1, M_2=2 et M_3=3

Similarité : d(a,b) = |a-b|

K-means

- Exemple 1 :

Soit le corpus D={1, 2, 3, 6, 7, 8, 13, 15, 17}

Appliquer l'algorithme k-means pour partitionner l'ensemble D en 3 clusters

Initialisation : prendre 3 objets au hasard 1, 2 et 3 M_1=1, M_2=2 et M_3=3

Similarité : d(a,b) = |a-b|

Résultat :

C1={1, 2, 3}

C2={6, 7, 8}

C3={13, 15, 17}

À quels groupes appartiennent les objets 18, 4, 14?

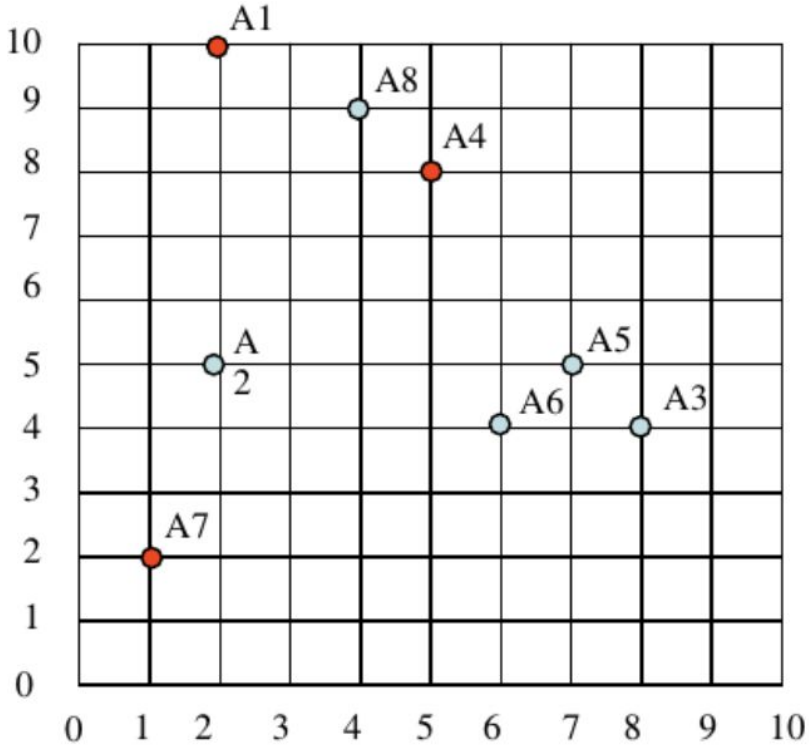
- Exemple 2 :

Soit le corpus D composé des points A1, A2, A3, A4, A5, A6, A7 et A8.

Les objets marqués en rouge représentent une initialisation des centroïdes.

Partitionner en groupes l'ensemble D en appliquant l'algorithme k-means

Montrer l'évolution des groupes en fonction des itérations d'applications de k-means



- Exemple 3 :

Soit le corpus D = {OB-1, OB-2, OB-3, OB-4, OB-5, OB-6, OB-7, OB-7}.

Partitionner en 2 groupes

Points	X	Y	Z
OB-1	1	4	1
OB-2	1	2	2
OB-3	1	4	2
OB-4	2	1	2
OB-5	1	1	1
OB-6	2	4	2
OB-7	1	1	2
OB-8	2	1	1

K-means

- k-means = regrouper les données en **k** groupes d'individus les plus semblables possibles
  - k petit => grands groupes
  - k grand => petits groupes

K-means

- Avantages
  - Classification rapide d'une nouvelle instance
- Inconvénients
  - Nombre de groupe k ?
  - Initialisation idéale des centroïdes pour une convergence rapide ?
  - Valeurs d'attributs non numériques