Stage LIESSE
Introduction à l'IA
B. Apprentissage Non Supervisé

Thomas Bonald

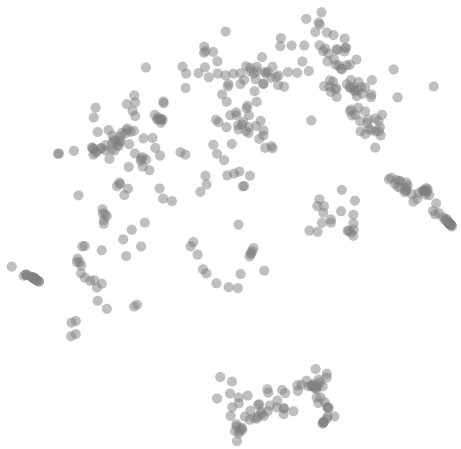
Avril 2021

Objectif

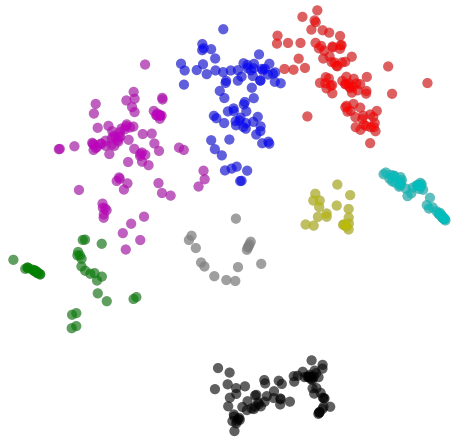
Comprendre la **structure** des données seules
(pas de classe ou de valeur connue).

En pratique, il s'agit de regrouper les données **proches** entre elles
$\rightarrow$ attribution d'une **classe**, qu'il faut ensuite **interpréter**.

Données

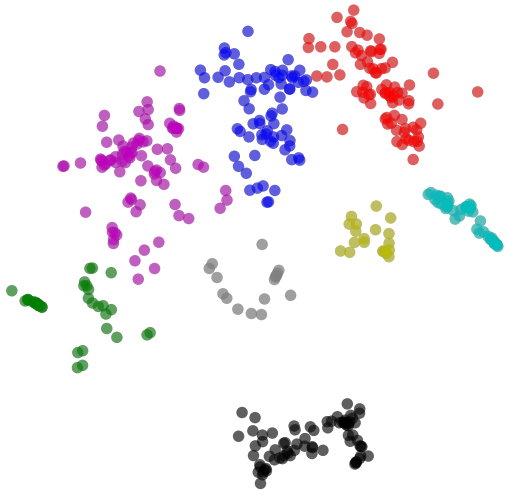


Partition

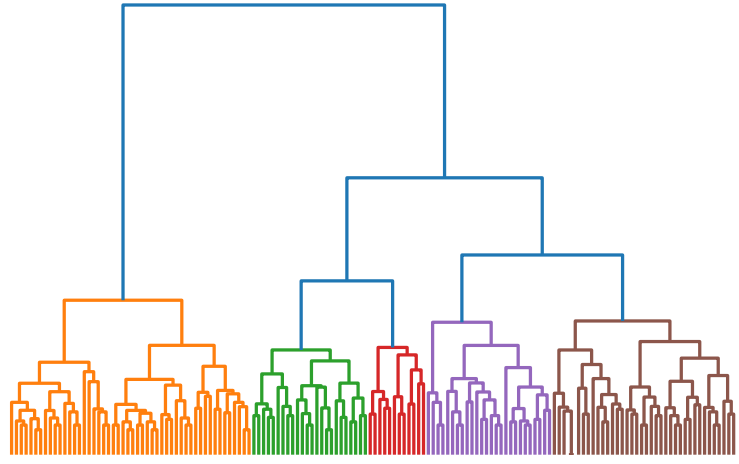


Plan

1. Partitionnement

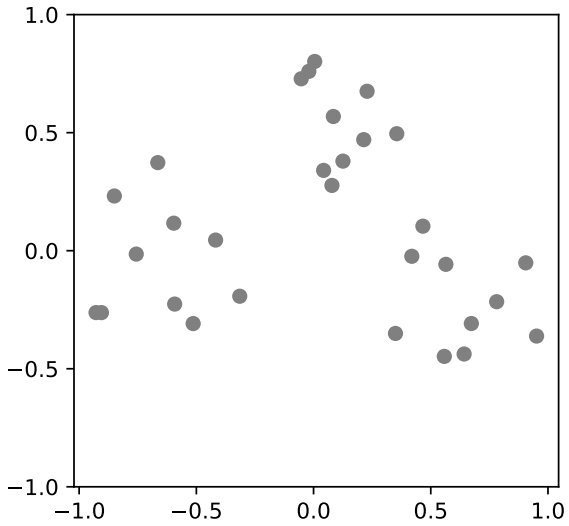


2. Structure hiérarchique



Partitionnement de données

- **Entrée :** $x_1, \ldots, x_n \in \mathbb{R}^d$
  k, nombre de clusters
- **Sortie :** $C_1, \ldots, C_k$, partition de $\{1, \ldots, n\}$



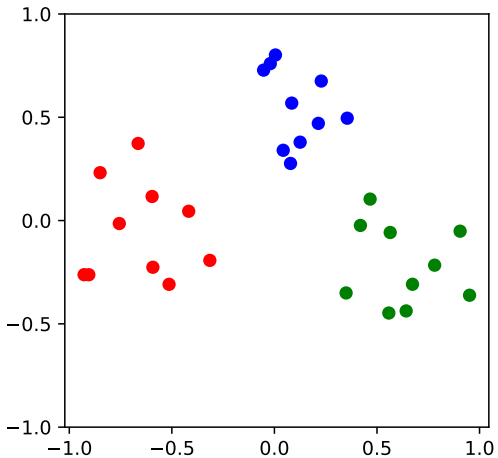
Fonction de coût

Somme des **moments d'inertie** :

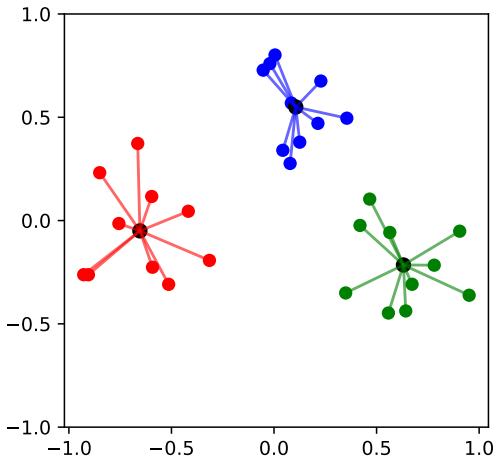
$$\sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2$$

où $\mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$ est le **barycentre** du cluster j

Partition



Coût

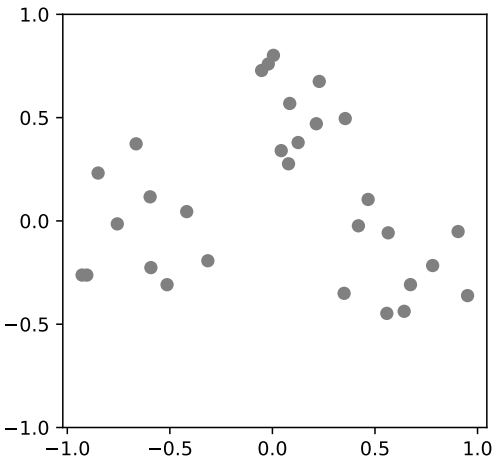


Un problème difficile

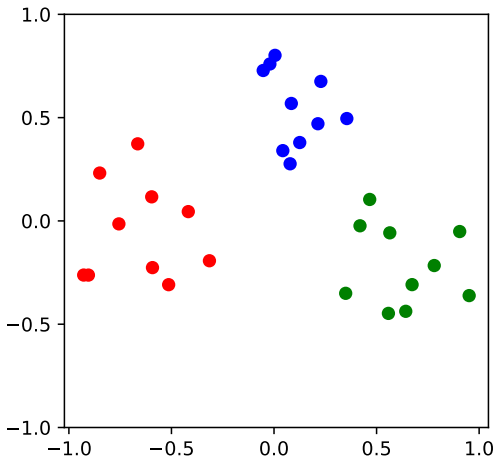
La recherche de la partition optimale est un problème **NP-difficile** :

$$\arg \min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2 \quad \text{avec } \mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$$

Données



Solution



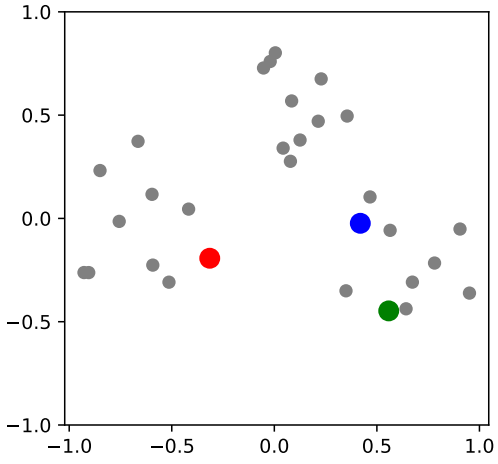
Algorithme des k moyennes

Algorithme **itératif** donnant une solution **approchée**.

Loyd 1957

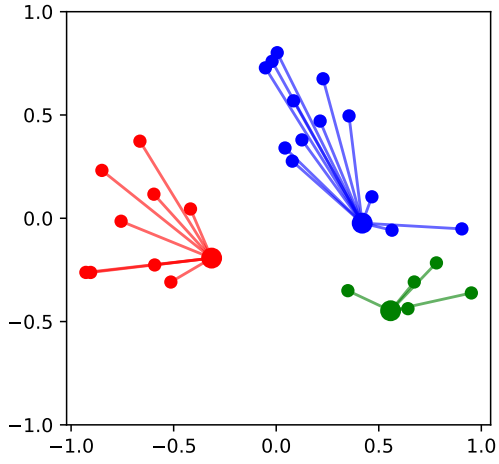
Initialisation

Choix aléatoire des centres



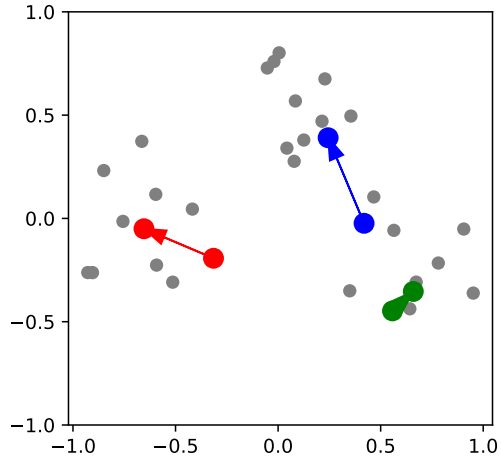
Première partition

Affectation au centre le plus proche



Mise à jour

Nouveaux centres des clusters



Application: Digits

1797 images de chiffres
8×8 pixels

Données

Clusters
$k = 10$

Apprentissage actif

1797 images de chiffres
8×8 pixels
Classification par **le plus proche voisin**
après étiquettage **manuel** de 20 échantillons

Choix aléatoire
Précision = 65%

Centres de clusters
Précision = 91%

Compression d'image

Image initiale
251 niveaux de gris

Image compressée
4 niveaux de gris

Image initiale
173 060 couleurs

Image compressée
16 couleurs

Segmentation d'image

Image initiale
251 niveaux de gris

Masque
1 niveau de gris

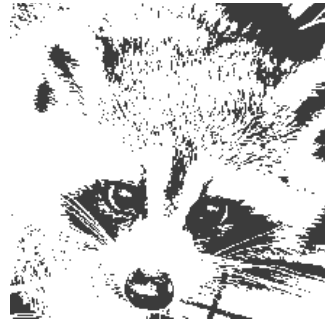
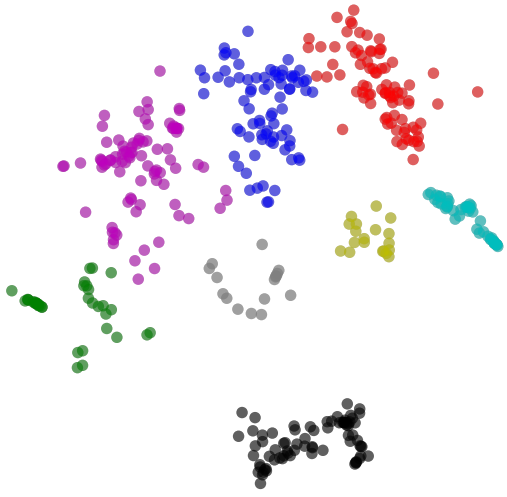


Image compressée
4 niveaux de gris

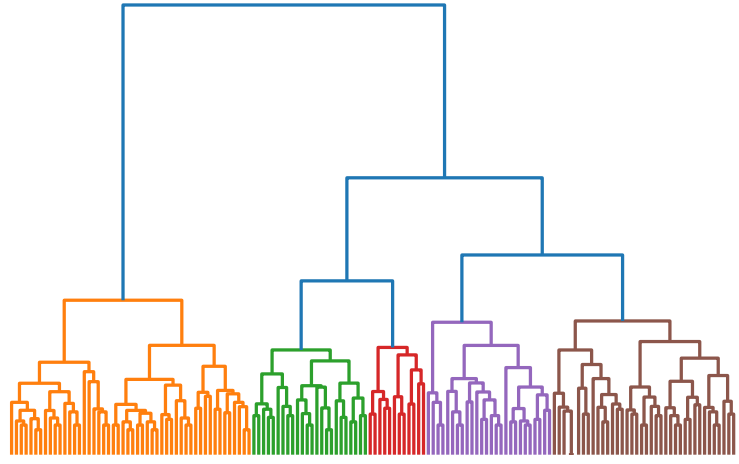
Image segmentée
10 segments

Plan

1. Partitionnement



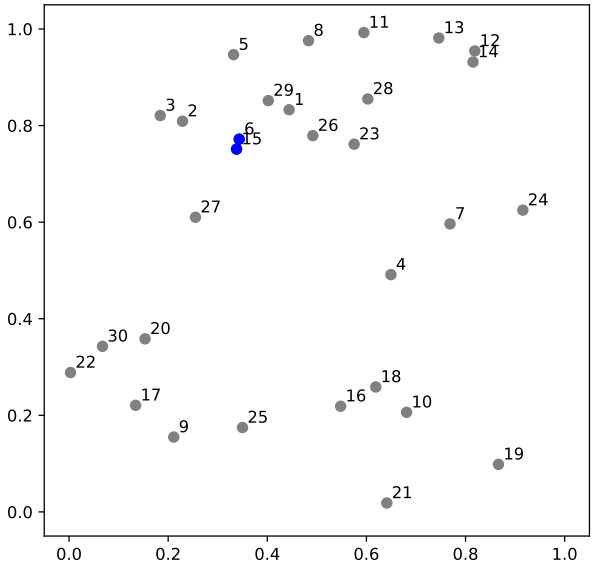
2. Structure hiérarchique



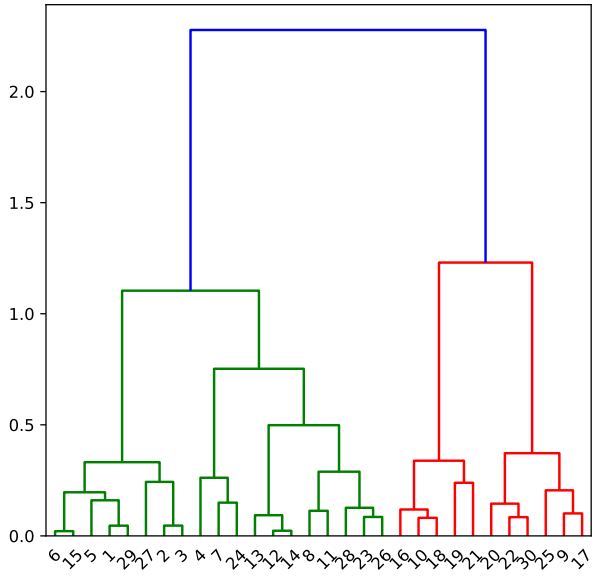
Structure hiérarchique

- **Entrée :** $x_1, \ldots, x_n \in \mathbb{R}^d$
- **Sortie :** Dendrogramme, arbre binaire de n feuilles

Données



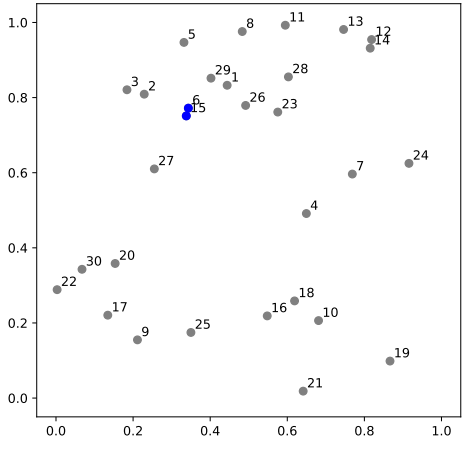
Dendrogramme



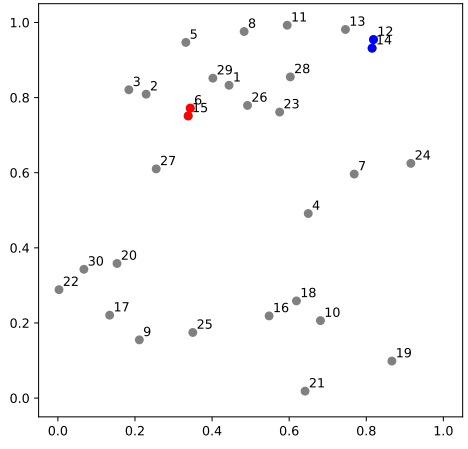
Algorithme agglomératif

Algorithme **glouton** regroupant les clusters les plus proches.

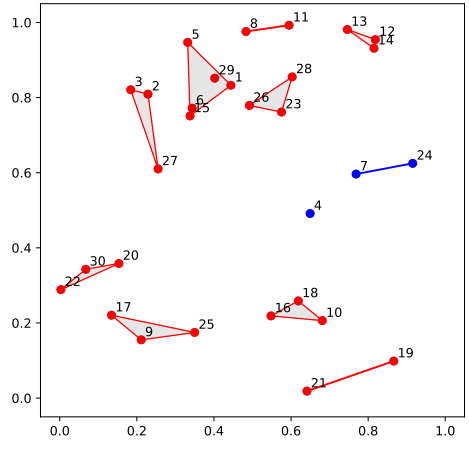
Étape 1



Étape 2



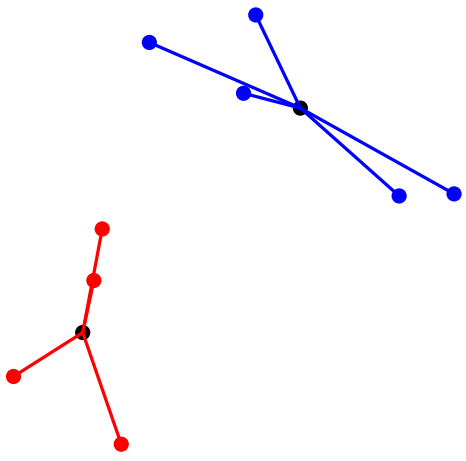
Étape 20



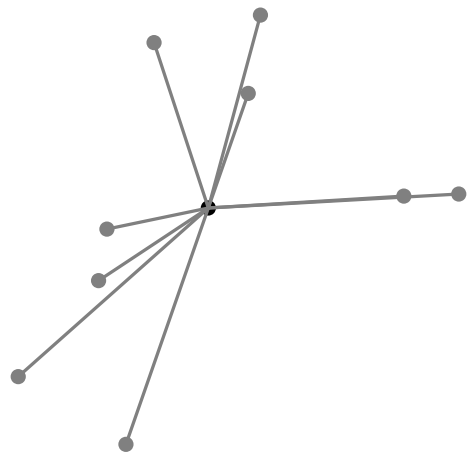
Méthode de Ward

Fusion des clusters minimisant l'augmentation du moment d'inertie
Ward 1963

Clusters séparés



Cluster fusionné



Équivalent à rechercher les clusters C_1, C_2 minimisant :

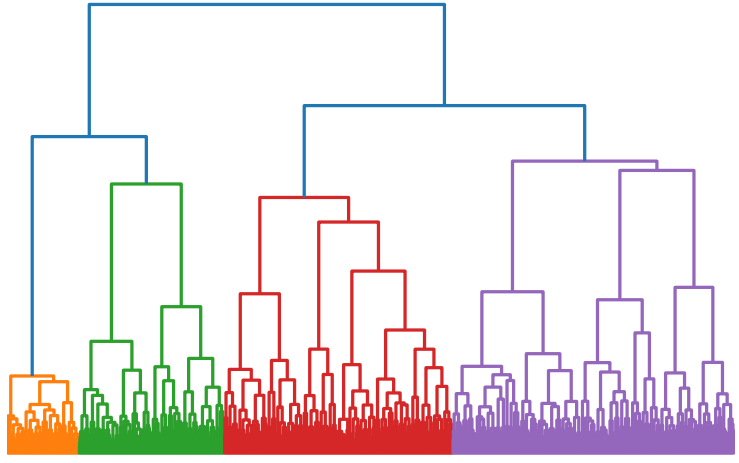
$$\frac{n_1 n_2}{n_1 + n_2} ||\mu_1 - \mu_2||^2 \quad \text{avec } n_j = |C_j| \quad \mu_j = \frac{1}{n_j} \sum_{i \in C_j} x_i$$

Application: Digits

1797 images de chiffres
8×8 pixels

Données

Dendrogramme
(méthode de Ward)



Application: Digits

1797 images de chiffres
8×8 pixels

Coupe du dendrogramme (12 clusters)

Centres des clusters

Dendrogramme