Il y a deux types d'algorithmes de classification :

- Classification supervisée : on connaît les classes de certaines données (données d'entraînement) qui permettent de prédire la classe d'une nouvelle donnée. Exemples : k plus proches voisins, ID3.
- Classification non supervisée : Il n'y a pas de donnée d'entraînement et l'ensemble des classes possibles n'est pas connu à l'avance. Exemples : k-moyennes, classification hiérarchique ascendante.

I Algorithme des k-moyennes (k-*means*)

On note d une distance (par exemple la distance euclidienne) et k un entier.

Définition : Centre

Le centre (ou : isobarycentre) d'un ensemble de vecteurs $X = \{x_1, \dots, x_n\}$ est le vecteur

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Définition : Inertie

On veut trouver une partition $\mathcal{P}$ de X en k sous-ensembles $X_1, \dots, X_k$ (classes ou *clusters*) minimisant l'inertie :

$$I(\mathcal{P}) = \sum_{i=1}^{k} \sum_{x \in X_i} d(x, \overline{X_i})^2$$

Plus l'inertie est petite, plus les données sont proches au sein de leur classe et plus le partitionnement est bon.

Algorithme des k-moyennes

Entrée : Des données X, un entier k
Sortie : Une partition de X en k classes

Soient $c_1, ..., c_k$ des vecteurs (centres) choisis aléatoirement
Tant que les centres ont changé **:**
  Associer chaque donnée x à la classe X_i telle que $d(x, c_i)$ soit minimum
  Recalculer les centres des classes $c_i = \overline{X_i}$
Renvoyer $X_1, \dots, X_k$

Remarques :

- On peut choisir les centres initiaux aléatoirement dans R^p ou parmi X.
- k est le nombre de classes dans l'algorithme des k-moyennes alors que c'est le nombre de voisins dans l'algorithme des k plus proches voisins.
- Le problème de décision consistant à déterminer s'il existe une partition d'inertie inférieure à un seuil est NP-complet.

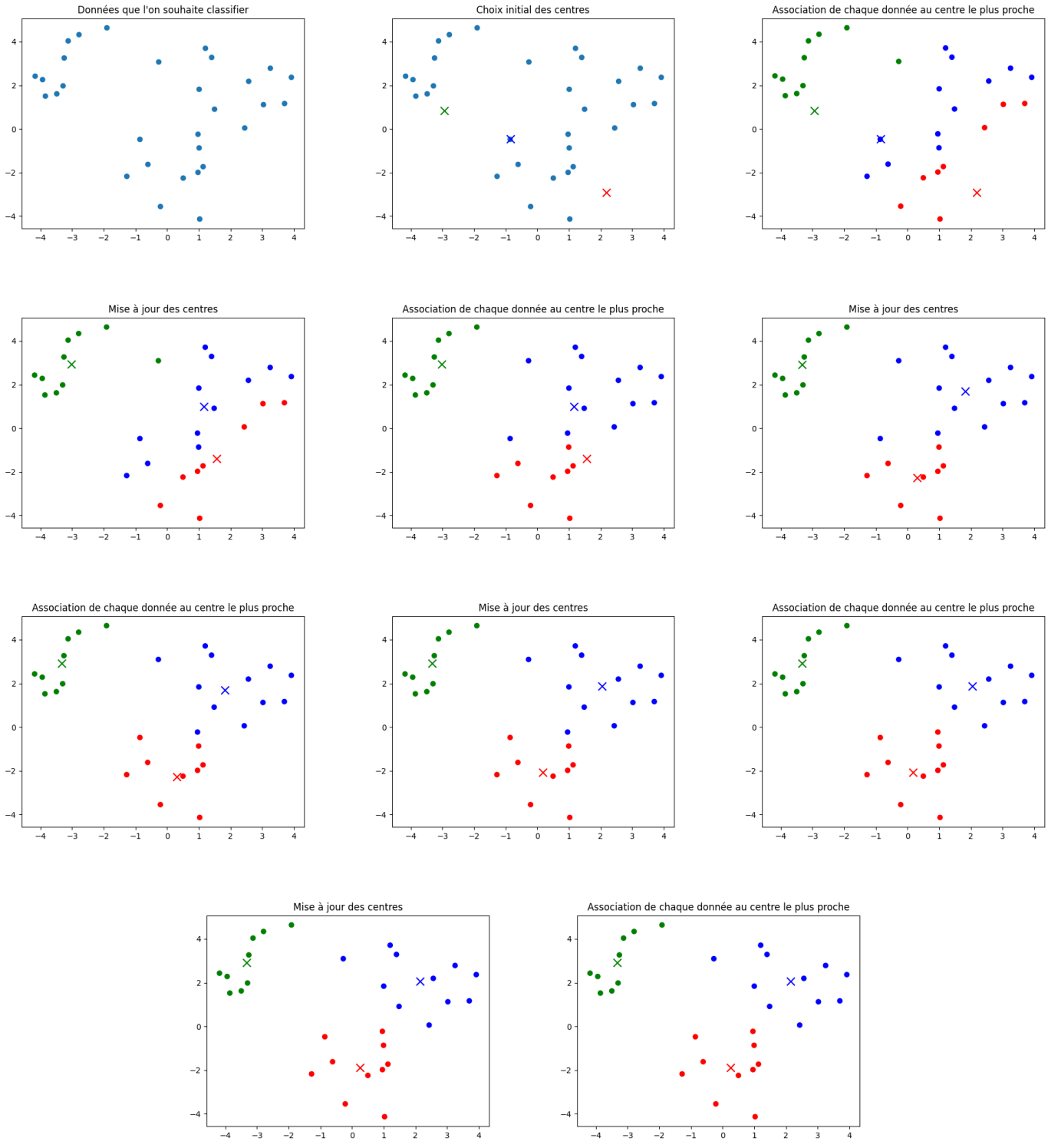
I.1 Terminaison (HP)

Théorème

L'algorithme des k-moyennes termine (pas de boucle infinie).

Preuve : Il existe un nombre fini de partitions de X en k classes, donc l'inertie I ne peut prendre qu'un nombre fini de valeurs. De plus, I diminuant strictement à chaque itération (c'est un variant) :

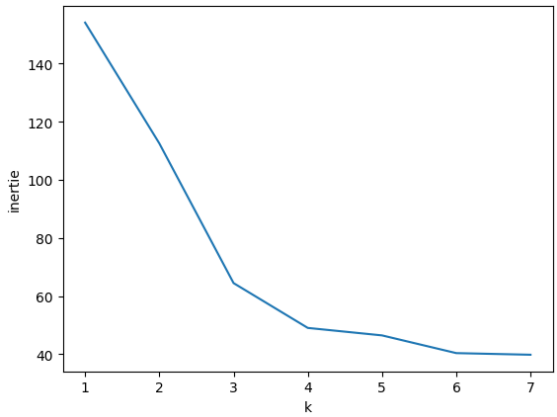
- Réassigner x de X_i à X_j si $d(x, c_i) > d(x, c_j)$ fait diminuer I.
- Recalculer les centres des classes fait diminuer I car $f : y \longmapsto \sum_{x \in X} d(x, y)^2$ est minimum pour $y = \overline{X}$.



Exemple d'exécution de l'algorithme des k-moyennes

I.2 Choisir k

On peut calculer l'inertie obtenue pour différentes valeurs de k. La méthode du coude consiste à choisir la plus grande valeur de k pour laquelle l'inertie diminue de façon significative.



On choisit $k = 3$ ou $k = 4$.

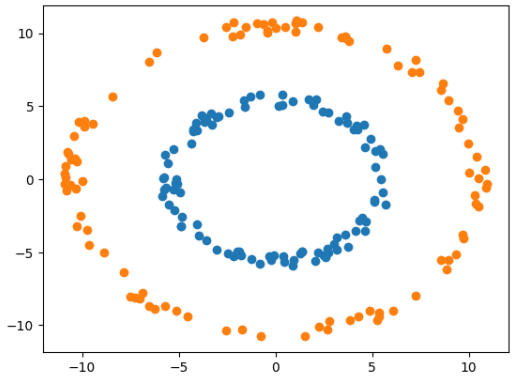
I.3 Non optimalité

L'algorithme des k-moyennes converge toujours vers un minimum local, mais pas forcément vers un minimum global de l'inertie.

Exercice 1.
Donner un exemple d'exécution de l'algorithme des k-moyennes qui ne donne pas une partition d'inertie minimum.

I.4 Limites

L'algorithme des k-moyennes ne marche que sur des données linéairement séparables (pouvant être séparées par un hyperplan).



L'algorithme des k-moyennes ne permettrait pas de classifier correctement ces données.

I.5 Interprétations

Les centres obtenus à la fin de l'algorithme donnent des informations sur les constituants des classes.

Centres obtenus avec $k = 10$ sur des chiffres manuscrits

II Classification hiérarchique ascendante

Classification hiérarchique ascendante

Entrée : Des données X
Sortie : Une partition de X en classes

Mettre chaque $x \in X$ dans une classe différente
Tant que nécessaire **:**
  Fusionner les deux classes les plus proches
Renvoyer Les classes obtenues

On peut choisir d'arrêter l'algorithme à un certain nombre de classes ou quand la distance minimum entre deux classes est supérieure à un certain seuil.

Exemples de distances entre classes A et B :

1. Distance minimum : $\min_{a \in A, b \in B} d(a, b)$.
2. Distance maximum : $\max_{a \in A, b \in B} d(a, b)$.
3. Distance moyenne : $\frac{1}{|A||B|} \sum_{a \in A, b \in B} d(a, b)$.

Exercice 2.
Appliquer l'algorithme de classification hiérarchique ascendante sur les données suivantes en dessinant le dendrogramme obtenu. On utilisera la distance 1.