# Apprentissage non supervisé – part 2

Les limite des Kmeans et les alternatives possibles.



Rafik LACHAAL

# Les limites des kmeans

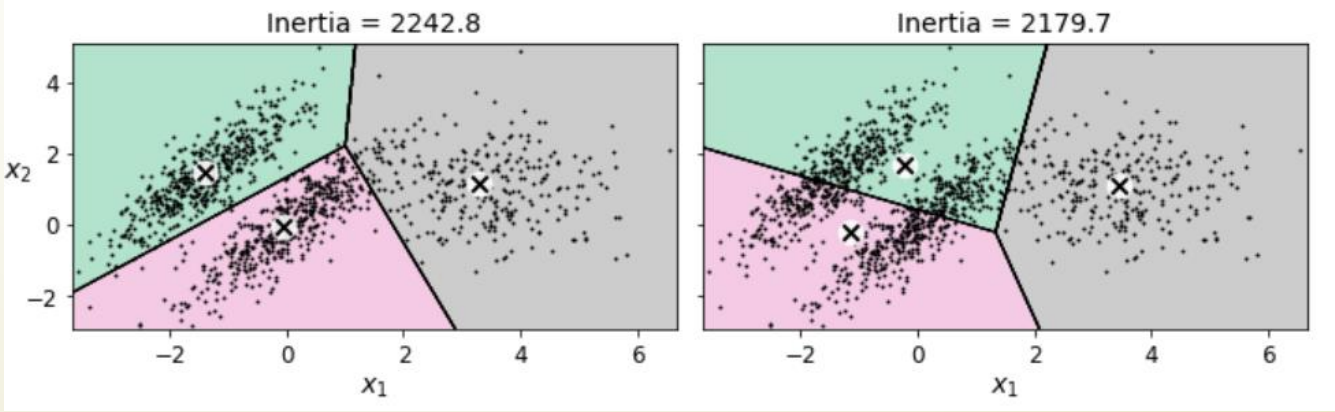


**Les avantages** que présente l'algorithme des kmeans :

- Très rapide
- S'adapte à de grande quantité de données

**Les limites** que présente l'algorithme des kmeans :

- Il faut exécuter l'algorithme plusieurs fois pour s'assurer d'avoir obtenue la solution optimal. Cela est dû au fait que l'or de la première itération les centrioles sont choisie de manière aléatoire.
- Il faut spécifier le nombre de partitions
- L'algorithme ne se comporte pas toujours bien lorsque les clusters, on des tailles, des densités et des formes non sphériques variées.

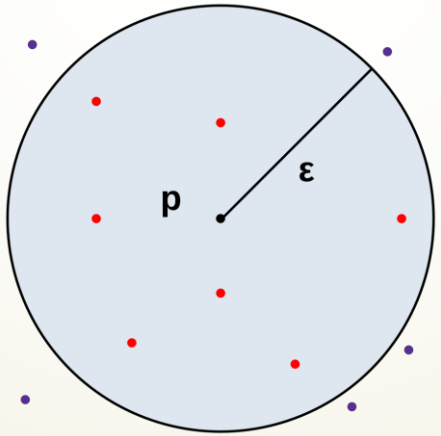


# Autre algorithme: le DBSCAN



- DBSCAN :
  - Basé sur l'estimation de la densité local
  - Permet d'identifier des clusters de forme arbitraire

- Principe de fonctionnement :
  1. Pour chaque observation, l'algorithme compte combien d'observations sont situé à moins d'une distance $\epsilon$ **(epsilon)** de celle-ci. Cette zone s'appelle le $\epsilon$**-voisinage** de l'observation.



# Le DBSCAN

2. SI cette observation à **au moins min_samples observations** (y compris elle-même) dans sont $\epsilon$-voisinage, alors on considère que c'est une **observation cœur** (core instance en anglais).

En d'autre terme c'est observations cœur sont celle situé dans des régions denses.

3. Toutes les observations au voisinage d'une observation cœur appartiennent au même cluster.

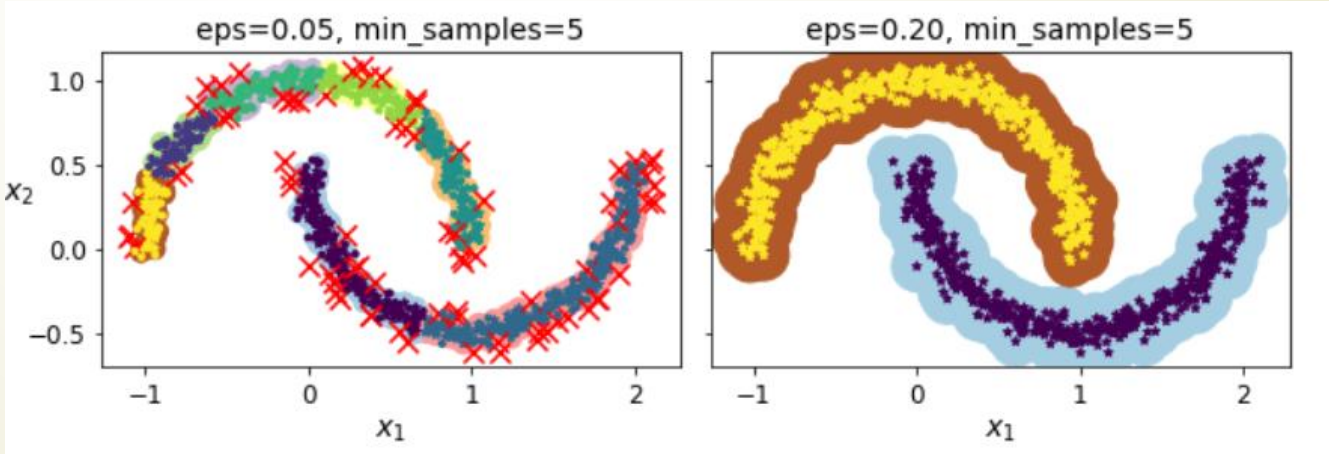
Il peut y avoir parmi elles d'autre observations cœur, par conséquent **une longue séquence d'observation cœur voisines constitue un cluster unique.**

4. Chaque observation qui n'est pas une observation cœur et qui n'en comporte pas une est **considéré comme une anomalie**.

# Le DBSCAN

Avantage :

- Fonctionne bien si tout les clusters sont suffisamment dense et s'il sont bien séparés par des zones de faible densité.
- N'est pas impacté par la forme des clusters et peut identifier des clusters de n'importe quel forme.
- Il n'est pas perturbé par les valeurs aberrantes.
- N'a que deux hyperparamètres et peux choisir ça métrique avec Sklearn.



# Le DBSCAN

Ses limites :

- Si la densité varie significativement d'un cluster à un autre, il peut lui être impossible de les identifier correctement.
- Sa complexité algorithmique est presque linéaire par rapport aux nombre de données.
- Si epsilon est grand sont utilisation avec Sklearn est consommatrice de mémoire.

# Autres algorithmes

- Partitionnement agglomératif
- BIRCH
- Mean-shift
- Propagation d'affinité
- Partitionnement spectrale

- Mélange Gaussiens