01-04
DBSCAN

Sommaire

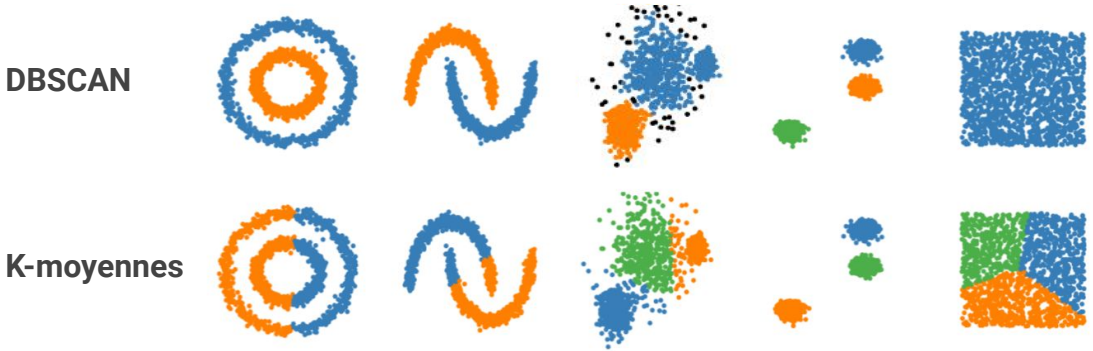
Sommaire

Rappel des principaux types de partitionnement

- Partitionnement basé sur
  - les centroïdes (K-moyennes, CURE, …)
  - la connectivité (hiérarchique, ...)
  - la distribution (BFR, …)
  - la densité (**DBSCAN**, OPTICS, …) 👈
  - les grilles
- Et d’autres ….

Introduction à DBSCAN

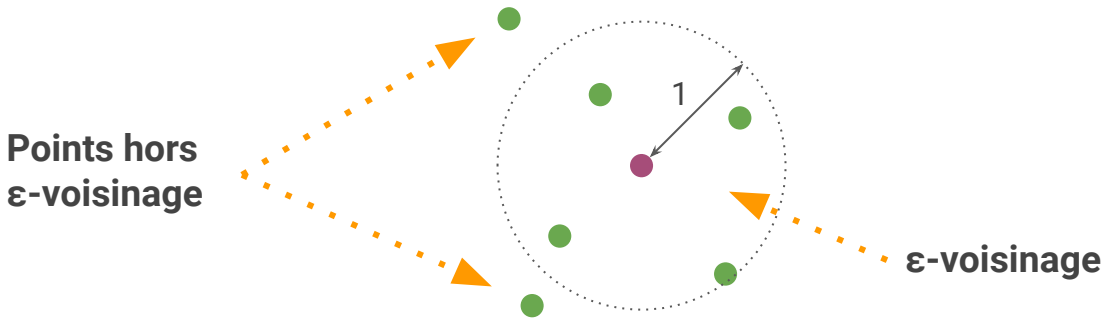
- **DBSCAN** → Density-Based Spatial Clustering of Applications with Noise
- Ne requiert pas le choix d'un nombre de clusters
- Permet de trouver des clusters de formes arbitraires, c.a.d. non nécessairement sphériques
- Robuste au bruit et données aberrantes (cf. détection d'anomalies)



Sommaire

Paramètres principaux

- DBSCAN requiert la spécification de deux paramètres:
  - **ε** → rayon maximum du voisinage
  - **MinPts** → nombre minimum de points dans un ε-voisinage d'un point donné (inclus)
- Example avec ε = 1

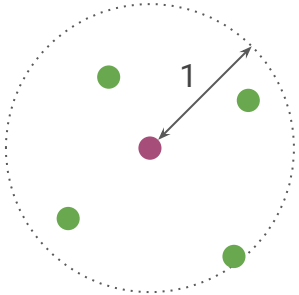


Terminologie

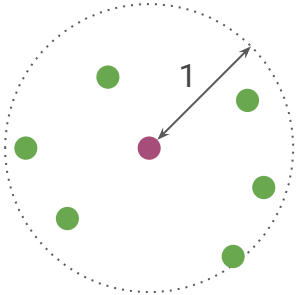
- **ε-voisinage dense**

  Un ε-voisinage est dit **dense** si le nombre de points est supérieur ou égal à MinPts

- Example avec ε = 1 et MinPts = 7



Non dense



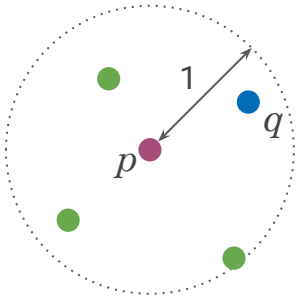
Dense

Terminologie

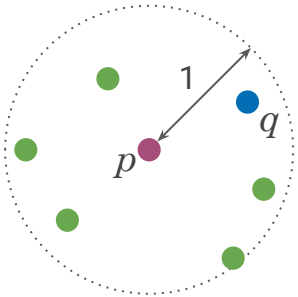
- **Point directement accessible par densité**

  Un point q est directement accessible par densité depuis un autre point p si l'ε-voisinage du point p est dense et si q appartient à l'ε-voisinage du point p

- Example avec ε = 1 et MinPts = 7



q n'est pas directement accessible par densité depuis p



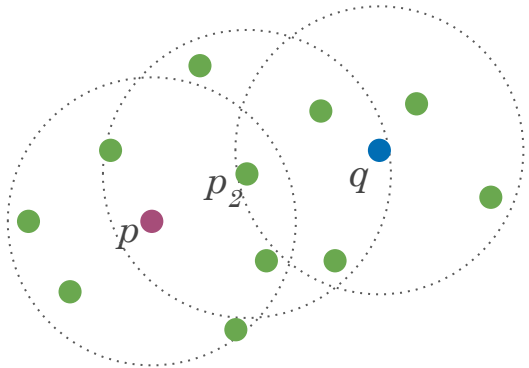
q est directement accessible par densité depuis p

Terminologie

- **Point accessible par densité**

  Un point q est accessible par densité depuis un autre point p s'il existe une séquence ordonnée de points (p_1, p_2, ... , p_n) telle que

  - $p_1 = p$
  - p_{i+1} est directement accessible par densité depuis p_i
  - $p_n = q$

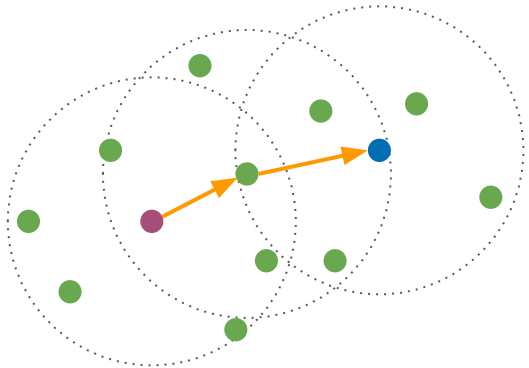


Terminologie

- **Point accessible par densité**

  Un point q est accessible par densité depuis un autre point p s'il existe une séquence ordonnée de points $(p_1, p_2, \ldots, p_n)$ telle que

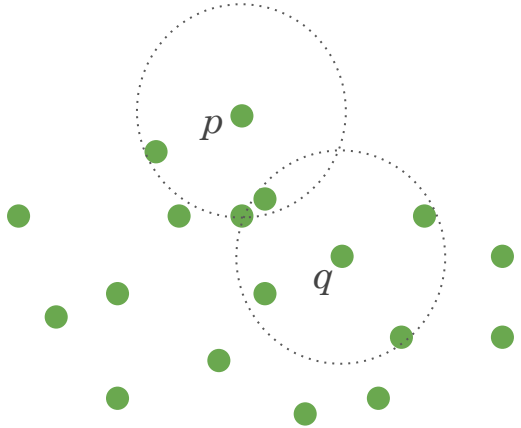
  - $p_1 = p$
  - p_{i+1} est directement accessible par densité depuis p_i
  - $p_n = q$



On parle aussi de **chemin fléché** de p vers q

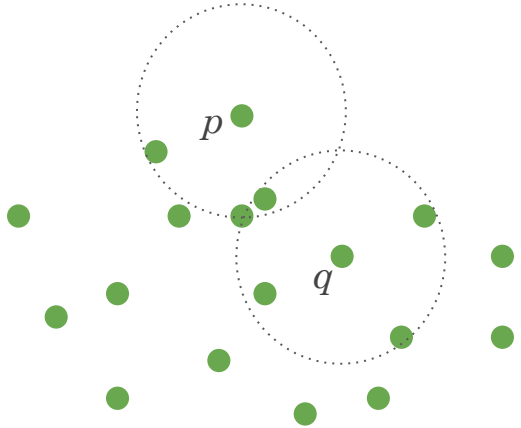
Question

- Soit les observations de la figure ci-dessous. En considérant MinPts = 5 et ε = 1, le point p est accessible par densité depuis le point q. Le point q est-il accessible par densité depuis le point p ?



Réponse

- Soit les observations de la figure ci-dessous. En considérant MinPts = 5 et ε = 1, le point p est accessible par densité depuis le point q. Le point q est-il accessible par densité depuis le point p ?



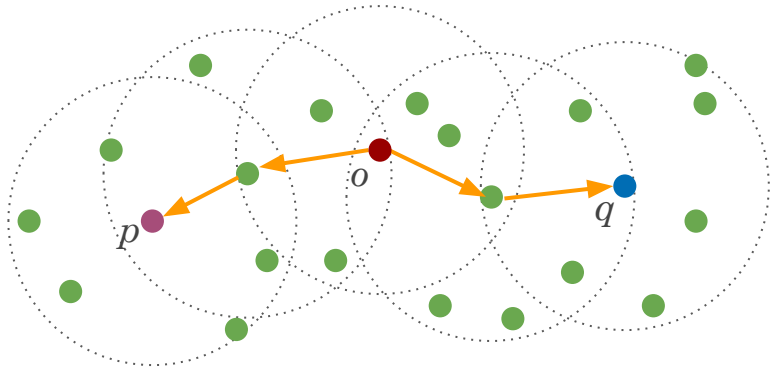
Non, car l'ε-voisinage de p n'est pas dense, donc aucun de ses points ne peut être directement accessible par densité

Terminologie

- **Point densément connecté**

  Un point q est densément connecté à un autre point p si

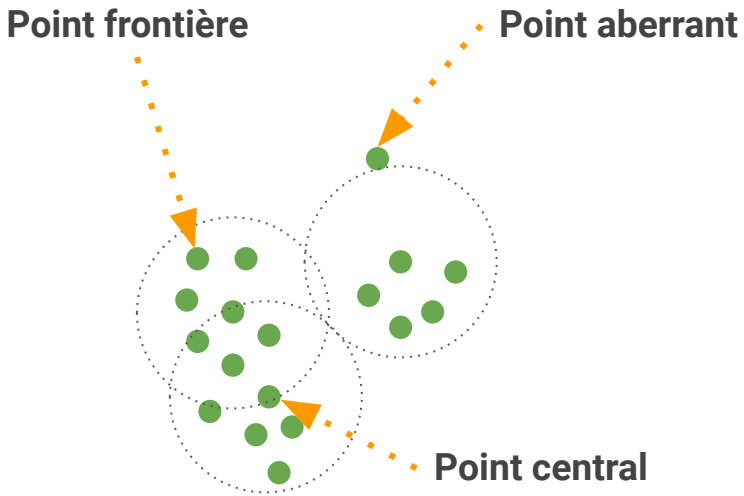
  - p est accessible par densité depuis un point o
  - q est accessible par densité depuis le point o



Terminologie

- Un point est dit **central** si son voisinage est dense
- Un point est dit **frontière** s'il n'est pas un point central et s'il appartient au voisinage d'un point central
- Enfin, un point est dit **aberrant** s'il n'est pas un point central et s'il n'appartient pas au voisinage d'un point central

- Example avec MinPts = 5



Sommaire

Algorithme DBSCAN

Sélectionner un point p quelconque

Trouver tous les points accessibles par densité depuis p (en fonction de epsilon et MinPts)

Si p est un point central, former un cluster

Si p est un point frontière, aucun point n’est accessible par densité depuis p. Passer au point suivant

Répéter jusqu’à ce que tous les points aient été traités

Sommaire

https://github.com/mswawola-cegep/420-a58-sf-gr-12060.git

01-04

Sommaire

Références

[1] Understanding DBSCAN Algorithm and Implementation from Scratch

[2] Cluster Analysis with DBSCAN : Density-based spatial clustering of applications with noise

[3] Comparing different clustering algorithms on toy datasets

[4] DBSCAN (Wikipedia, pour termes français)