Intelligence Artificielle et Apprentissage

Cours 4 : clustering par densité

Adrien Revault d'Allonnes

ara@up8.edu

Université Paris 8 – Vincennes à Saint-Denis

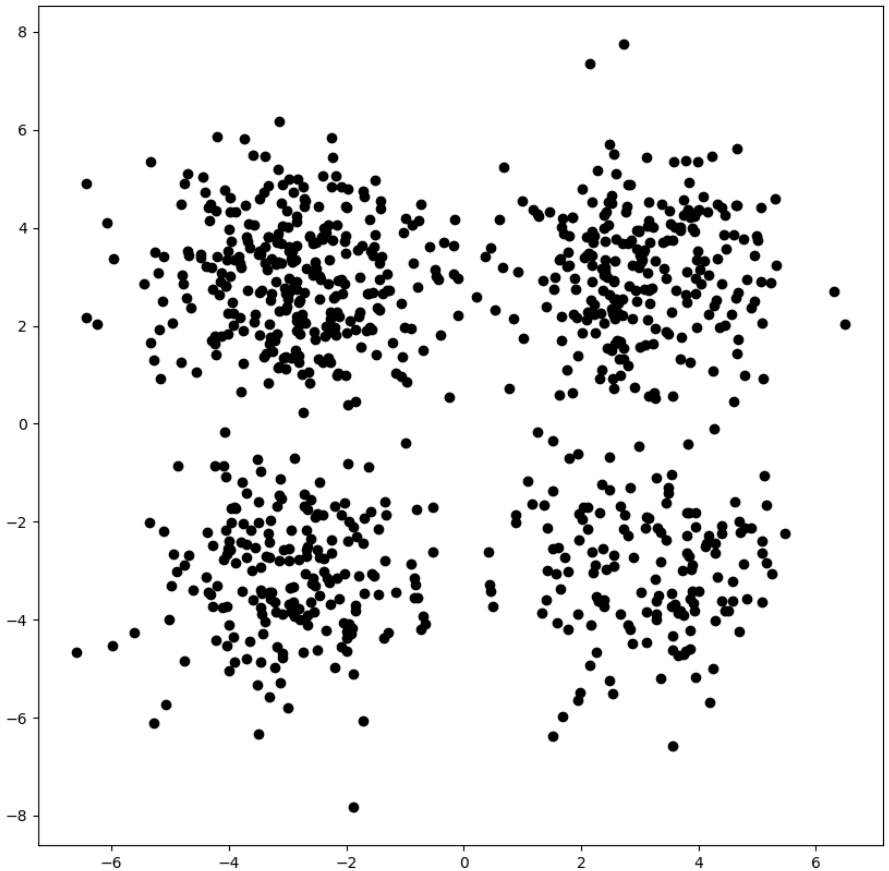
IAA – S2 – 2024

Aux l'épisodes précédents

- k-moyennes
- Clustering hiérarchique

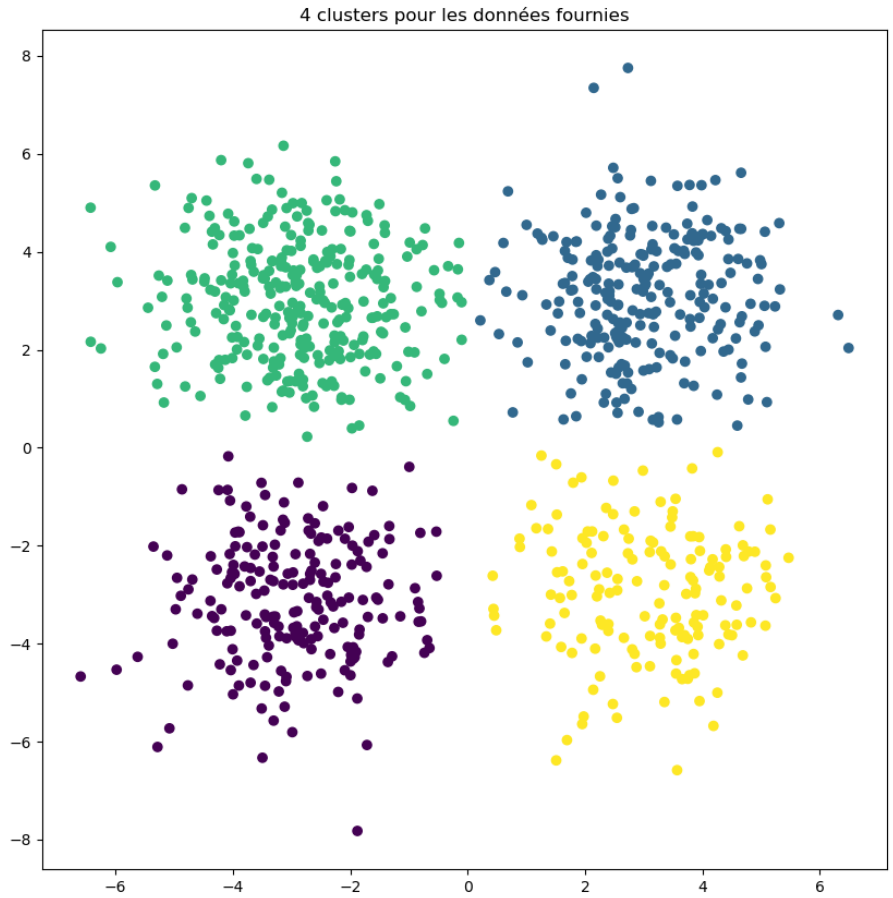
Aux l'épisodes précédents

- k-moyennes
- Clustering hiérarchique



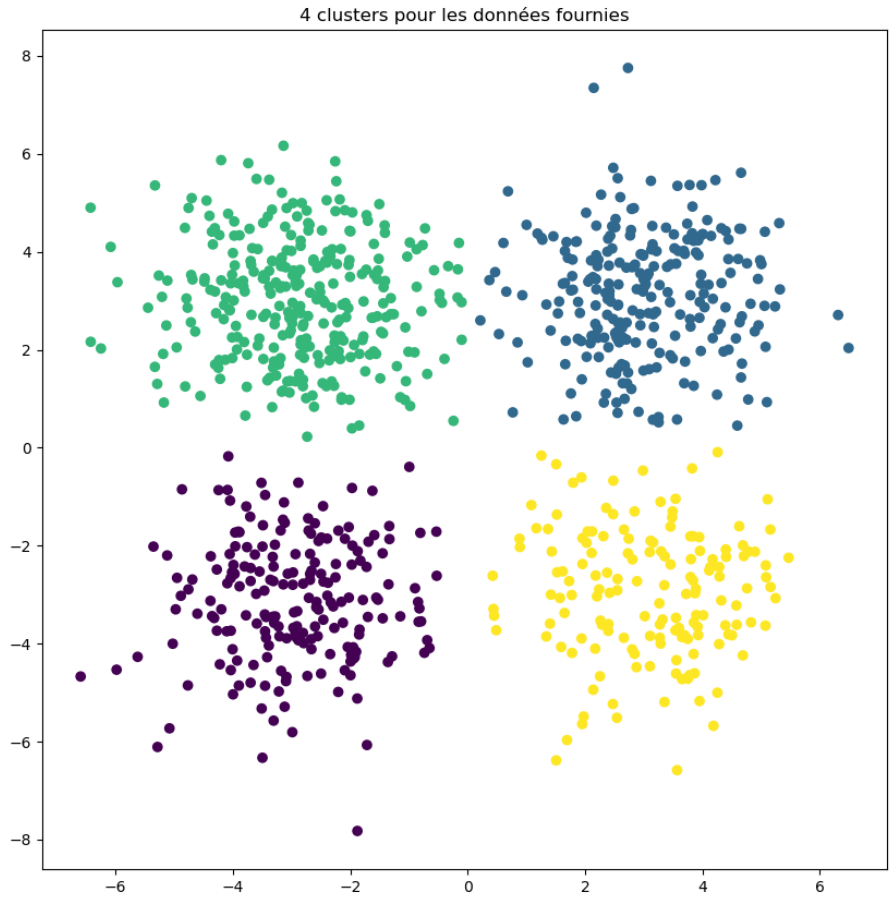
Aux l'épisodes précédents

- k-moyennes
- Clustering hiérarchique



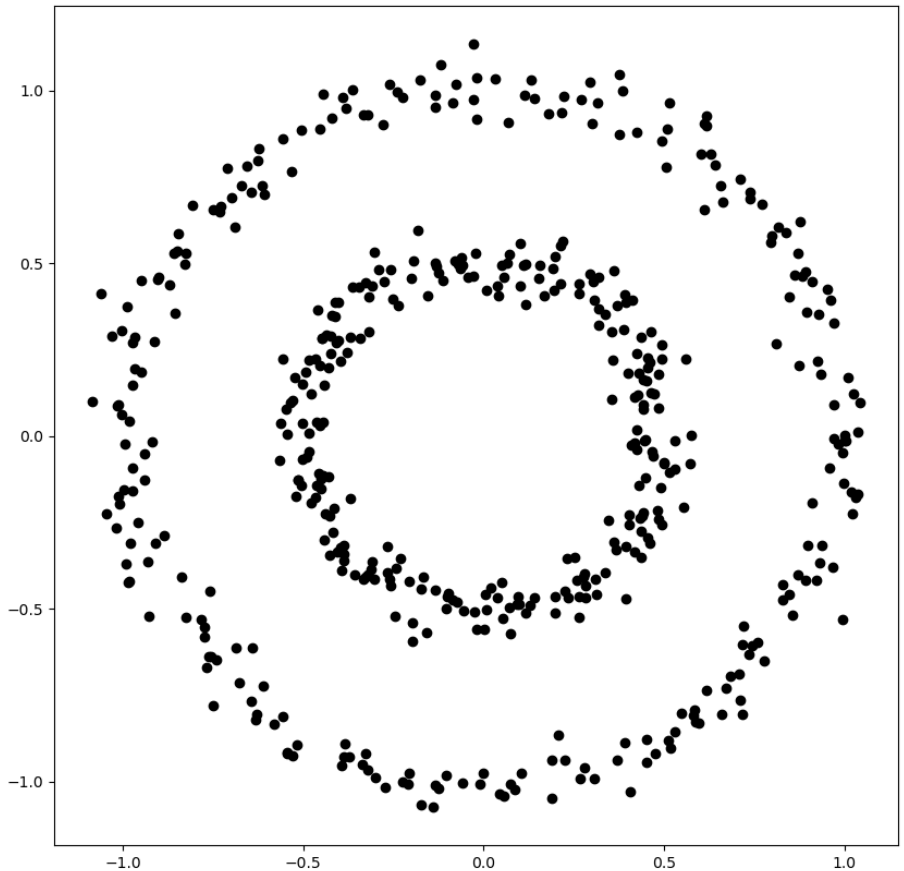
Aux l'épisodes précédents

- k-**moyennes** et sur d'autres données ?
- Clustering hiérarchique



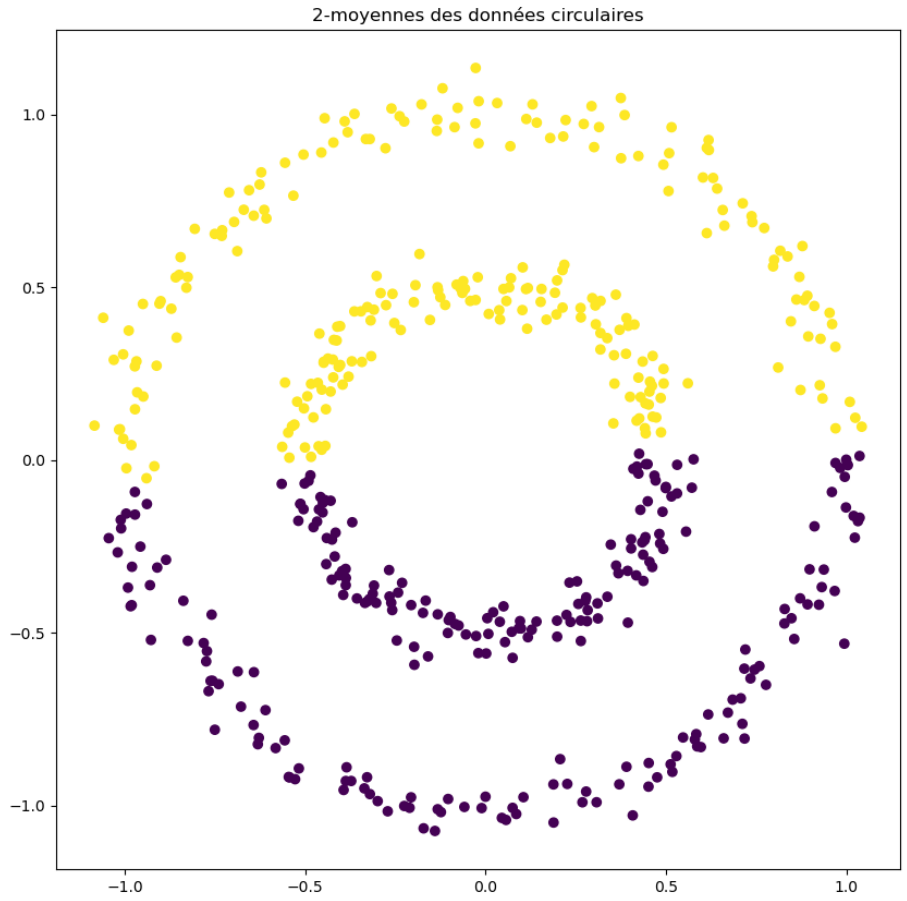
Aux l'épisodes précédents

- k-**moyennes** et sur d'autres données ?
- Clustering hiérarchique



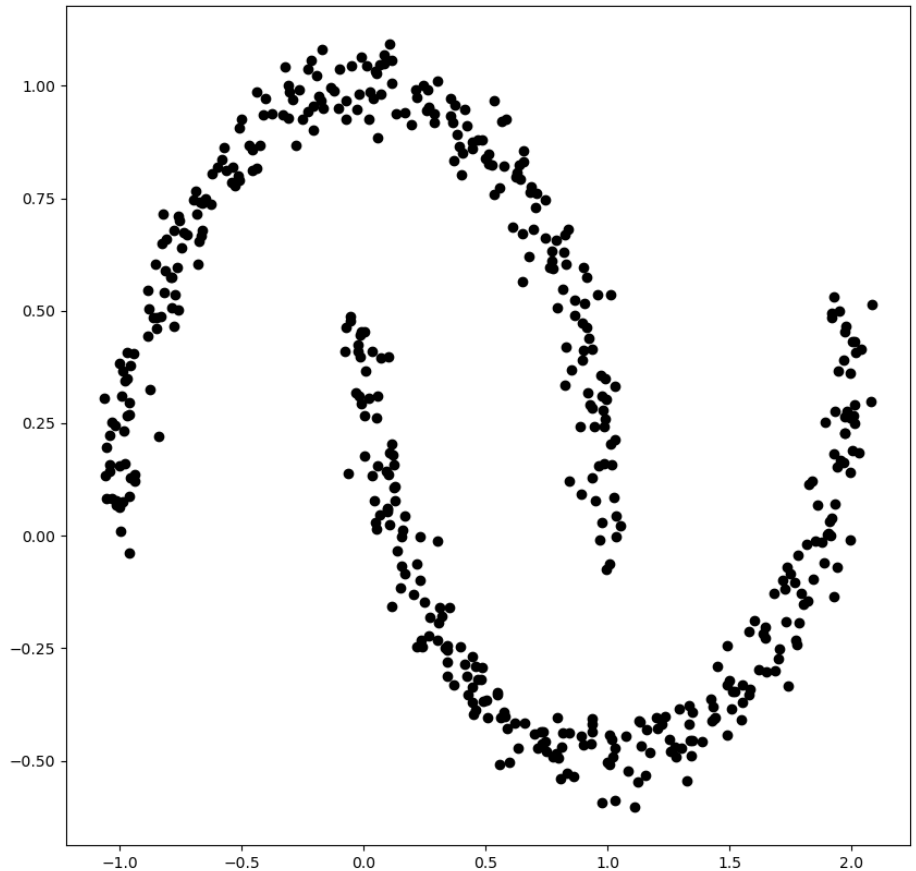
Aux l'épisodes précédents

- k-**moyennes** et sur d'autres données ?
- Clustering hiérarchique



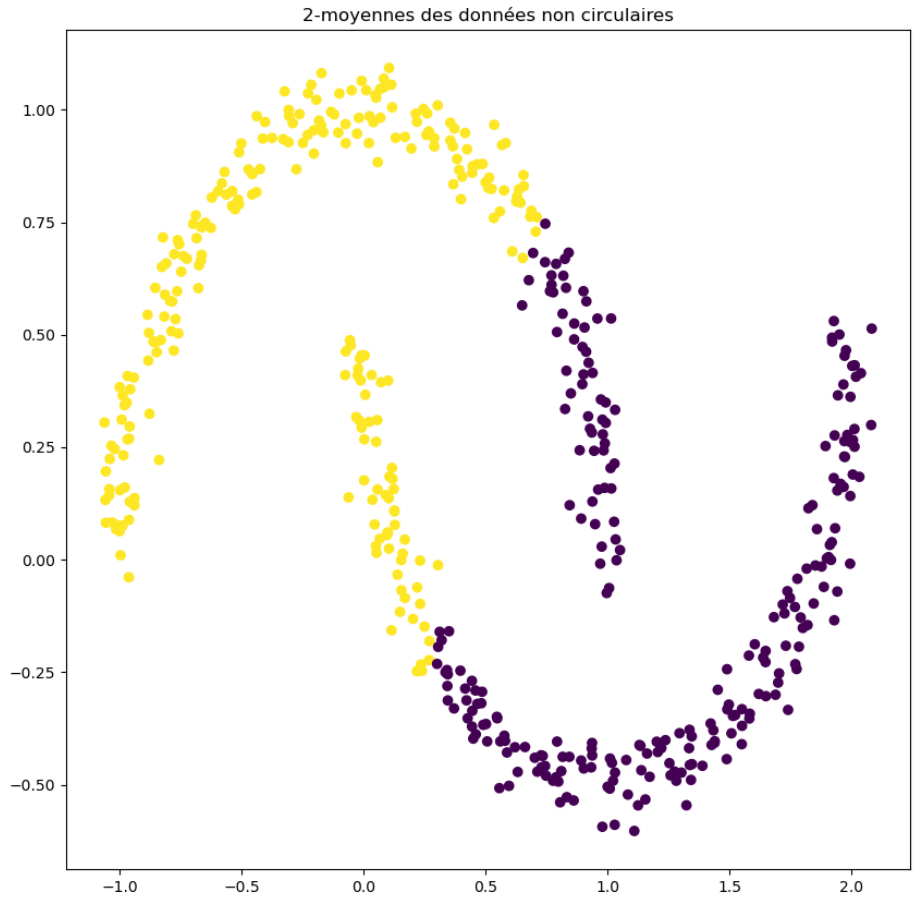
Aux l'épisodes précédents

- k-**moyennes** et sur d'autres données ?
- Clustering hiérarchique



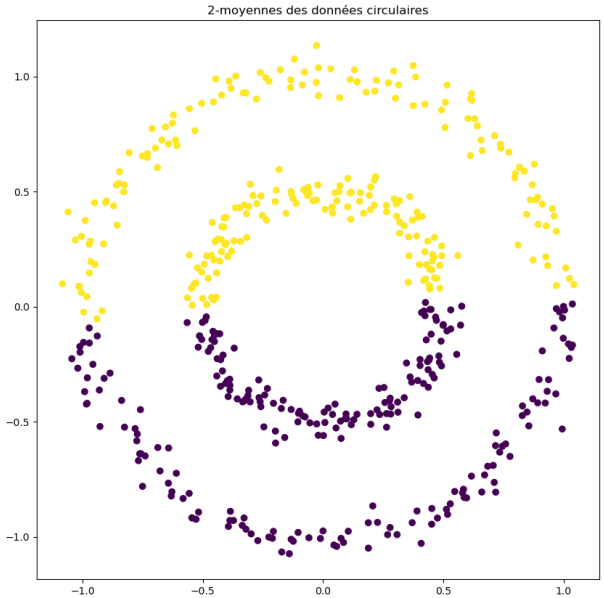
Aux l'épisodes précédents

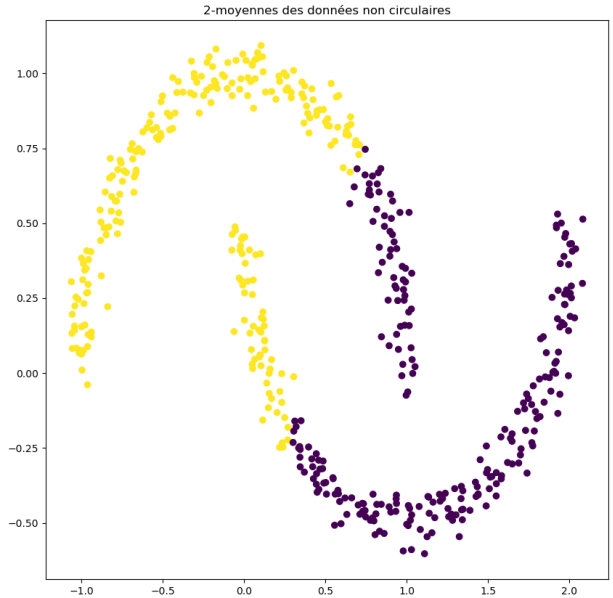
- k-**moyennes** et sur d'autres données ?
- Clustering hiérarchique



Un autre point de vue sur le clustering

- Problèmes des k-moyennes :
  - clusters non sphériques
  - clusters de tailles différentes
  - clusters de densités différentes
  - exceptions
  - clusters vides
  - fixer k





Un autre point de vue sur le clustering

- Problèmes des k-moyennes :
  - clusters non sphériques
  - clusters de tailles différentes
  - clusters de densités différentes
  - exceptions
  - clusters vides
  - fixer k

$\Rightarrow$ Clustering par densité

Un autre point de vue sur le clustering

- Problèmes des k-moyennes :
  - clusters non sphériques
  - clusters de tailles différentes
  - clusters de densités différentes
  - exceptions
  - clusters vides
  - fixer k

$\Rightarrow$ Clustering par densité

  - un cluster = une zone où la densité dépasse un seuil

Un autre point de vue sur le clustering

- Problèmes des k-moyennes :
  - clusters non sphériques
  - clusters de tailles différentes
  - clusters de densités différentes
  - exceptions
  - clusters vides
  - fixer k

$\Rightarrow$ Clustering par densité

- un cluster = une zone où la densité dépasse un seuil

\+ clusters de formes arbitraires

\+ gère le bruit

\+ un seul parcours des données

Un autre point de vue sur le clustering

- Problèmes des k-moyennes :
  - clusters non sphériques
  - clusters de tailles différentes
  - clusters de densités différentes
  - exceptions
  - clusters vides
  - fixer k

$\Rightarrow$ Clustering par densité

- un cluster = une zone où la densité dépasse un seuil

\+ clusters de formes arbitraires

\+ gère le bruit

\+ un seul parcours des données

\− paramètre de densité pour terminer

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

$\Rightarrow$ Densité d'un point = nb. points à $\leq \varepsilon$ du point

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

$\Rightarrow$ Densité d'un point = nb. points à $\leq \varepsilon$ du point

- ε-voisinage d'un point p : $\mathcal{V}_\varepsilon(p) = \{q \in D | d(p,q) \leq \varepsilon\}$

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

$\Rightarrow$ Densité d'un point = nb. points à $\leq \varepsilon$ du point

- ε-voisinage d'un point p : $\mathcal{V}_\varepsilon(p) = \{q \in D | d(p,q) \leq \varepsilon\}$
- Densité de p=$|\mathcal{V}_\varepsilon(p)|$

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

$\Rightarrow$ Densité d'un point = nb. points à $\leq \varepsilon$ du point

- ε-voisinage d'un point p : $\mathcal{V}_\varepsilon(p) = \{q \in D | d(p,q) \leq \varepsilon\}$
- Densité de p=$|\mathcal{V}_\varepsilon(p)|$
  - $\mathcal{V}_\varepsilon(c) \geq$`MinPts` : **point central** *(central point)*

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

$\Rightarrow$ Densité d'un point = nb. points à $\leq \varepsilon$ du point

- ε-voisinage d'un point p : $\mathcal{V}_\varepsilon(p) = \{q \in D | d(p,q) \leq \varepsilon\}$
- Densité de p=$|\mathcal{V}_\varepsilon(p)|$
  - $\mathcal{V}_\varepsilon(c) \geq$`MinPts` : point central *(central point)*
  - $\mathcal{V}_\varepsilon(f) <$`MinPts` et $d(f,c) \leq \varepsilon$: **point frontière** *(border point)*

DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise*
- Un cluster = une zone où la densité dépasse un seuil

Â¿ Qu'est-ce que la densité ?

« Qualité de ce qui est dense, de ce qui est fait d'éléments nombreux et serrés, contient beaucoup de matière par rapport à l'espace occupé. » `https://cnrtl.fr/definition/densité`

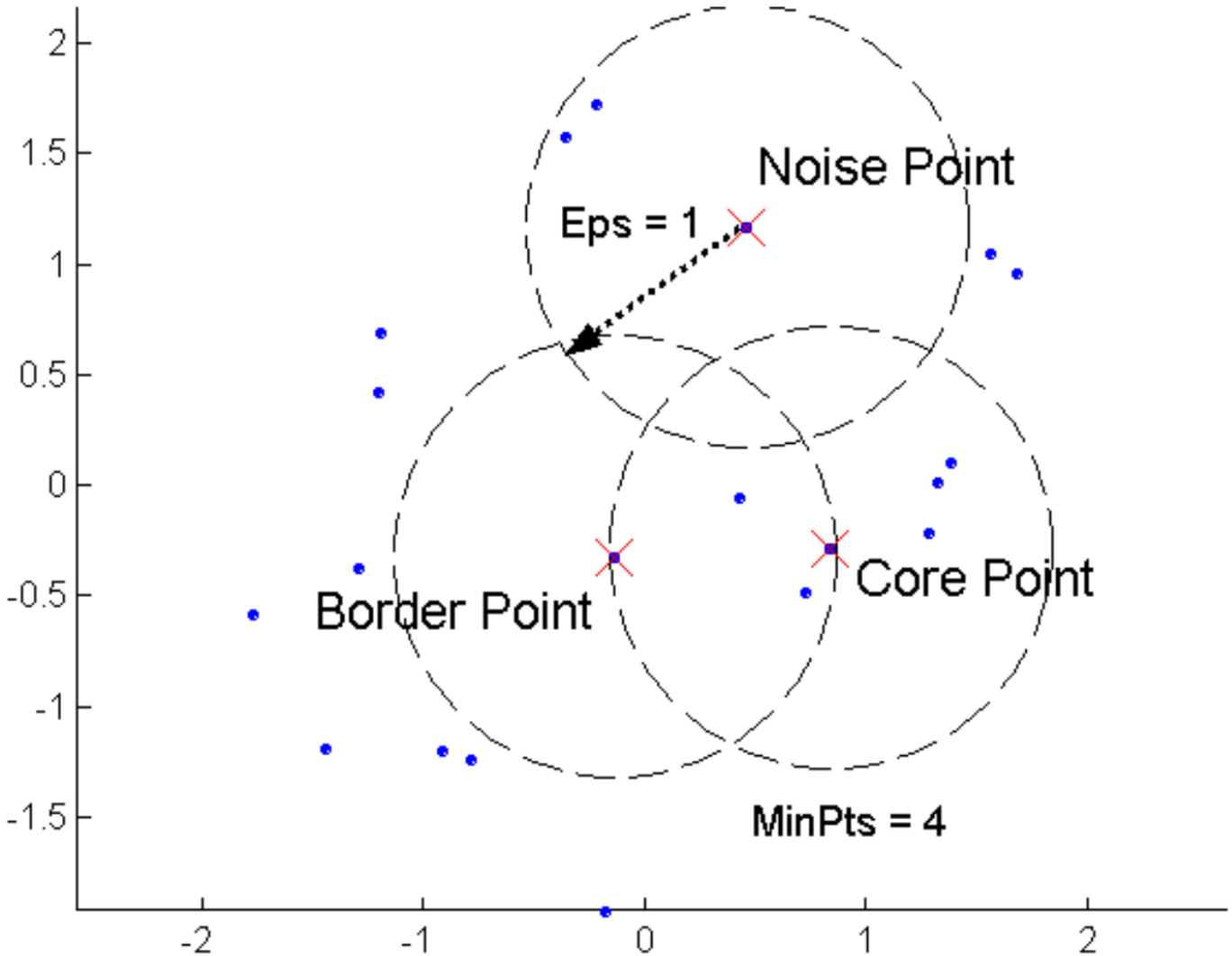
$\Rightarrow$ Beaucoup de trucs dans un petit espace

- Deux paramètres :
  - ε (ou `eps`) : le rayon du voisinage
  - `MinPts` : nombre minimal de voisins dans ε

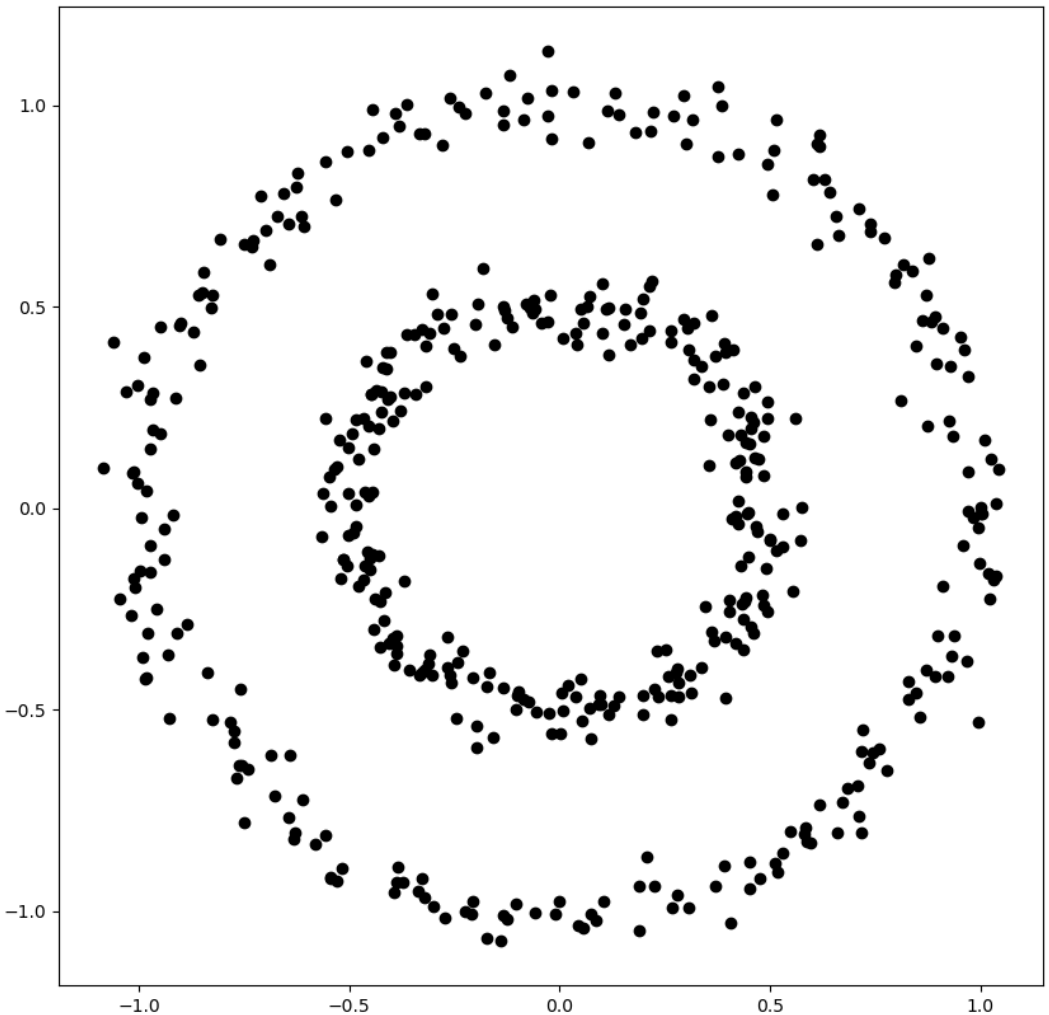
$\Rightarrow$ Densité d'un point = nb. points à $\leq \varepsilon$ du point

- ε-voisinage d'un point p : $\mathcal{V}_\varepsilon(p) = \{q \in D | d(p,q) \leq \varepsilon\}$
- Densité de p=$|\mathcal{V}_\varepsilon(p)|$
  - $\mathcal{V}_\varepsilon(c) \geq$`MinPts` : point central *(central point)*
  - $\mathcal{V}_\varepsilon(f) <$`MinPts` et $d(f,c) \leq \varepsilon$: point frontière *(border point)*
  - $\mathcal{V}_\varepsilon(a) <$`MinPts` et $d(a,c) > \varepsilon$: **point aberrant** *(noise point)*

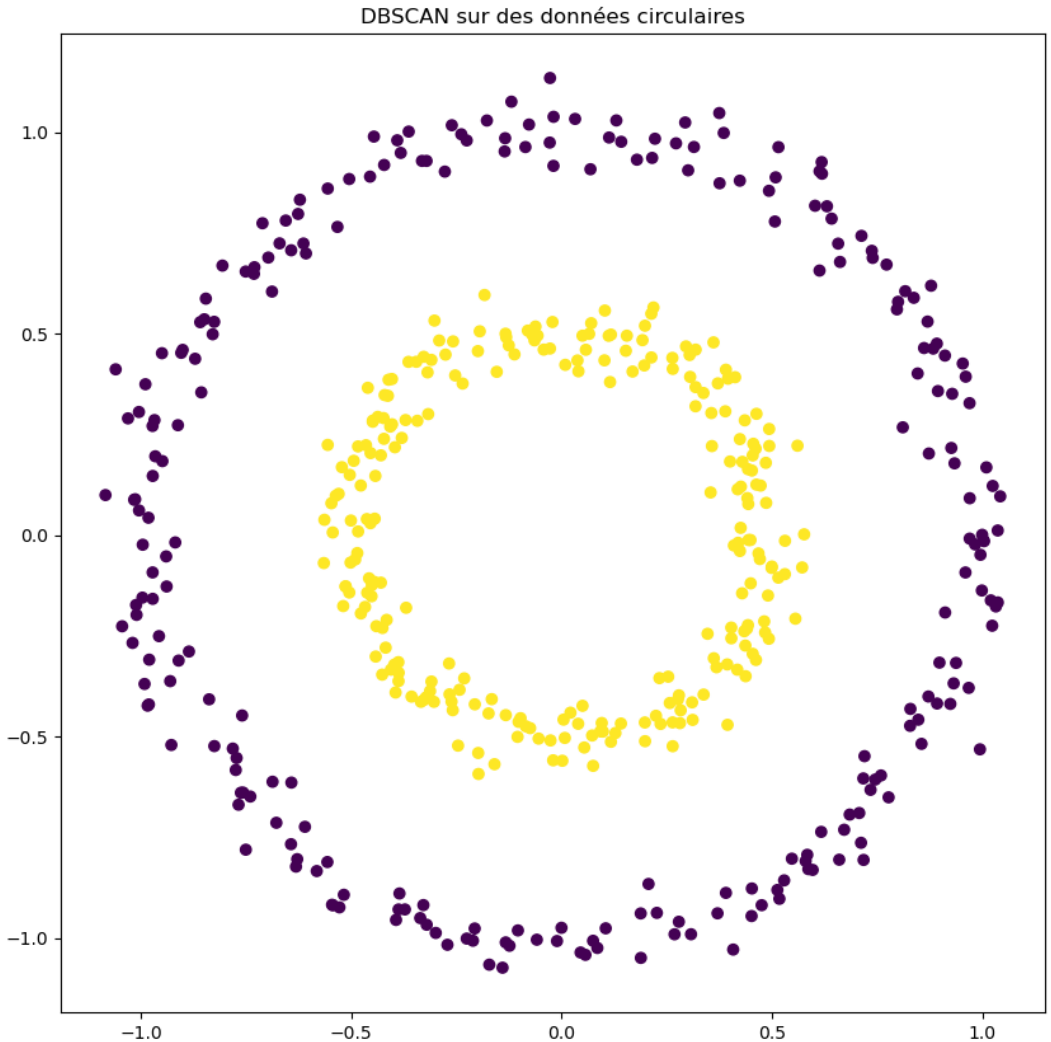
Visuellement



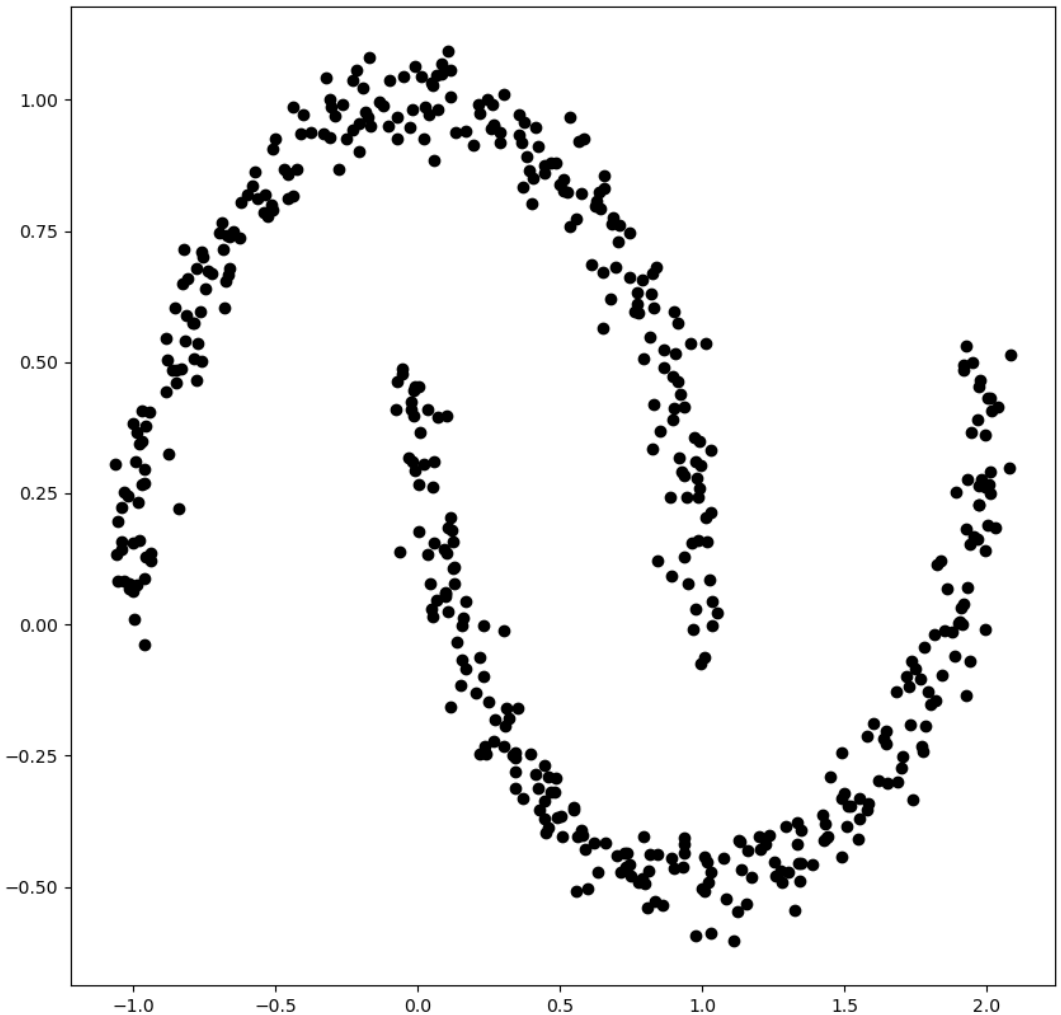
Ce que ça peut faire



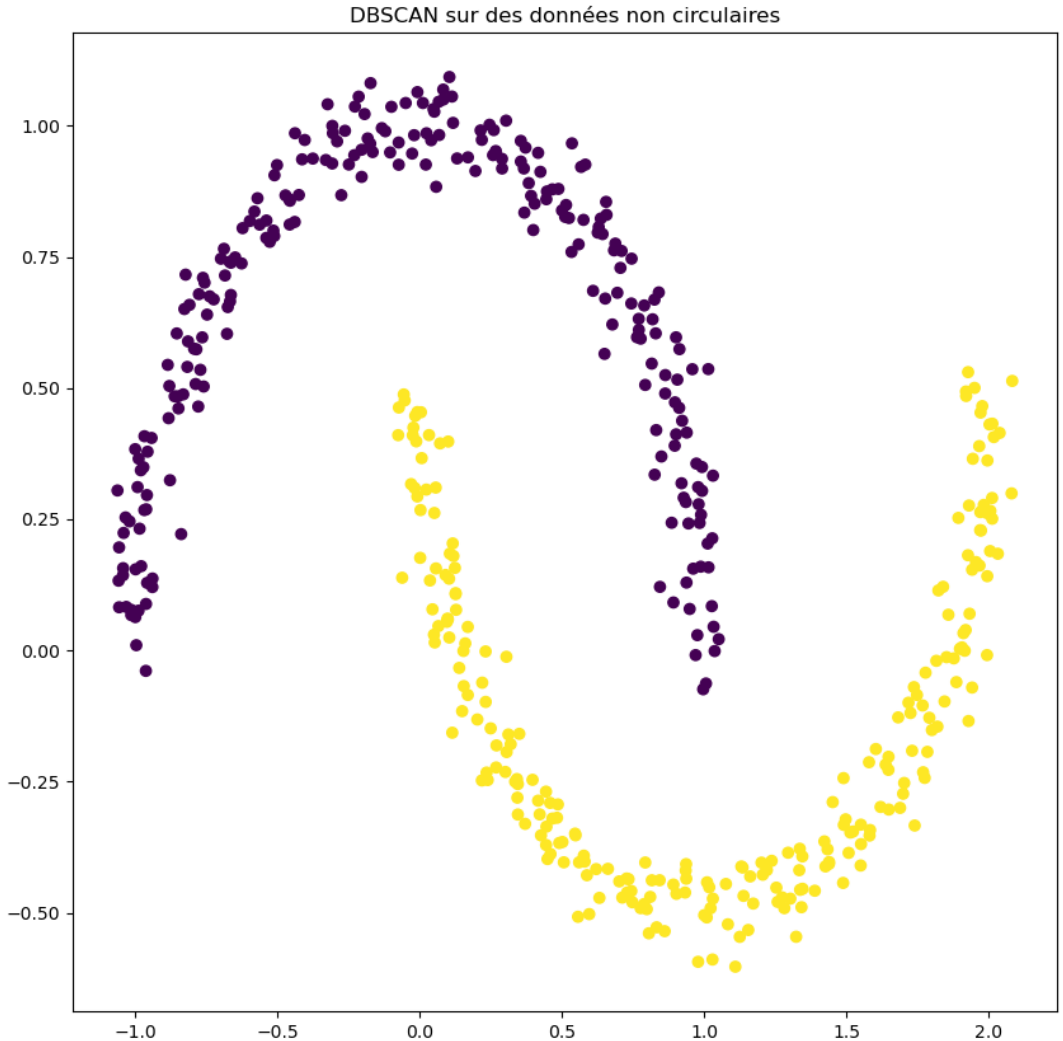
Ce que ça peut faire



Ce que ça peut faire



Ce que ça peut faire



Accessibilité par densité

- q est **directement accessible par densité** depuis p si :
  - $\mathcal{V}_\varepsilon(p)$ est dense (i.e. p est un point central)
  - $q \in \mathcal{V}_\varepsilon(p)$
- q est **accessible par densité** depuis p
  s'il existe une séquence $\{p_1, \ldots, p_n\}$ telle que :
  - $p_1, = p$
  - p_{i+1} est directement accessible par densité depuis p_i
  - $p_n = q$
- q est **densément connecté** à p si $\exists o \in D$
  - p est accessible par densité depuis o
  - q est accessible par densité depuis o

DBSCAN, l'algo

```
DBSCAN(D, eps, MinPts)
   C = 0
   pour chaque point P non visité des données D
      marquer P comme visité
      PtsVoisins = epsilonVoisinage(D, P, eps)
      si tailleDe(PtsVoisins) < MinPts
         marquer P comme BRUIT
      sinon
         C++
         étendreCluster(D, P, PtsVoisins, C, eps, MinPts)
étendreCluster(D, P, PtsVoisins, C, eps, MinPts)
   ajouter P au cluster C
   pour chaque point P' de PtsVoisins
      si P' n'a pas été visité
         marquer P' comme visité
         PtsVoisins' = epsilonVoisinage(D, P', eps)
         si tailleDe(PtsVoisins') >= MinPts
            PtsVoisins = PtsVoisins U PtsVoisins'
      si P' n'est membre d'aucun cluster
         ajouter P' au cluster C
epsilonVoisinage(D, P, eps)
   retourner tous les points de D qui sont à une distance inférieure à eps de P
```

Illustration