

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

# Lecture 11: DB-Scan

## Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2023-2024

## 1 Introduction

## 2 Principle

## 3 Algorithm

## 4 Hyperparameters

## 5 Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Introduction

# Question

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Do you remember what the goal of clustering is ?  
Do you remember what algorithms we studied ?

# Introduction

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## DBScan

Density-based spatial clustering of applications with noise (DBSCAN) is a **density-based clustering non-parametric algorithm.**

# Introduction

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## DBScan

Density-based spatial clustering of applications with noise (DBSCAN) is a **density-based clustering non-parametric algorithm.**

Given a set of points in some space:

# Introduction

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## DBScan

Density-based spatial clustering of applications with noise (DBSCAN) is a **density-based clustering non-parametric algorithm.**

Given a set of points in some space:

- group together points that are closely packed together (many nearby neighbors)

# Introduction

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## DBScan

Density-based spatial clustering of applications with noise (DBSCAN) is a **density-based clustering non-parametric algorithm.**

Given a set of points in some space:

- group together points that are closely packed together (many nearby neighbors)
- mark as outliers point that lie alone in low-density regions (nearest neighbors are far away)

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

**Principle**

Algorithm

Hyperparameters

Advantages  
and limits

# Principle

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Given a radius  $\epsilon$ , a distance  $d$  and threshold number of points  $n_{points}$ , each individual in the dataset can be labelled as:

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Given a radius  $\epsilon$ , a distance  $d$  and threshold number of points  $n_{points}$ , each individual in the dataset can be labelled as:

- *Core points*: at least  $n_{points}$  are within distance  $\epsilon$

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Given a radius  $\epsilon$ , a distance  $d$  and threshold number of points  $n_{points}$ , each individual in the dataset can be labelled as:

- *Core points*: at least  $n_{points}$  are within distance  $\epsilon$
- *Directly reachable*:  $q$  is *directly reachable* if  $p$  is within distance  $\epsilon$  of  $q$

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Given a radius  $\epsilon$ , a distance  $d$  and threshold number of points  $n_{points}$ , each individual in the dataset can be labelled as:

- *Core points*: at least  $n_{points}$  are within distance  $\epsilon$
- *Directly reachable*:  $q$  is *directly reachable* if  $p$  is within distance  $\epsilon$  of  $q$
- *Reachable*:  $q$  is *reachable* from  $p$  if there is a path  $p_0, p_1, \dots, p_n, q$  where each  $p_i$  is directly reachable from  $p_{i-1}$ .

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Given a radius  $\epsilon$ , a distance  $d$  and threshold number of points  $n_{points}$ , each individual in the dataset can be labelled as:

- *Core points*: at least  $n_{points}$  are within distance  $\epsilon$
- *Directly reachable*:  $q$  is *directly reachable* if  $p$  is within distance  $\epsilon$  of  $q$
- *Reachable*:  $q$  is *reachable* from  $p$  if there is a path  $p_0, p_1, \dots, p_n, q$  where each  $p_i$  is directly reachable from  $p_{i-1}$ .
- *Outlier*: All points not reachable from any other points.

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Once each point has been properly labelled:

- A cluster is the all the points (core or non-core) reachable from a core point.
- Non-reachable points are not clusterized.

# Principle

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

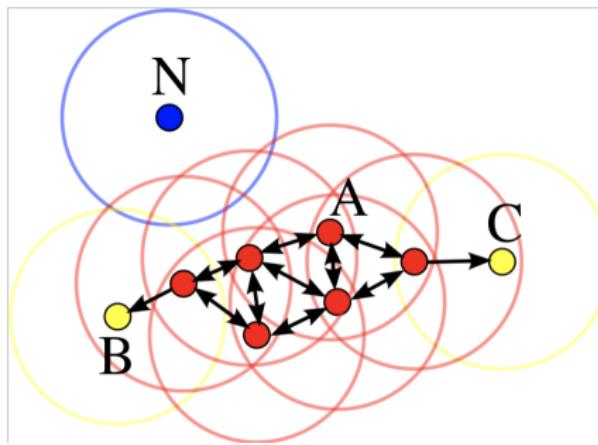
Principle

Algorithm

Hyperparameters

Advantages  
and limits

All red points are *core* points, yellow points are reachable and *N* is an outlier.



Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Algorithm

# Algorithm

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

- For every point in the dataset, find its  $\epsilon$  nearest neighbors and identify the core points with more than  $n_{points}$
- Find all the connected core points
- Assign each non-core point to a nearby cluster if the cluster is within  $\epsilon$  else assign it to noise.

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameter

Advantages  
and limits

# Hyperparameters

# Hyperparameters

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameter

Advantages  
and limits

## Question

Can you list the hyperparameters ?

# Hyperparameters

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameter

Advantages  
and limits

## Question

Can you list the hyperparameters ?

- Min points for individuals to be defined as core points
- $\epsilon$  neighborhood radius
- Distance between individuals

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages and limits

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

- Does not require to choose the number of clusters

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

- Does not require to choose the number of clusters
- Arbitrarily shaped clusters

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

- Does not require to choose the number of clusters
- Arbitrarily shaped clusters
- Notion of noise and robust to outliers

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

- Does not require to choose the number of clusters
- Arbitrarily shaped clusters
- Notion of noise and robust to outliers
- Only three hyperparameters that can be set by domain experts.

## Limits

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

- Does not require to choose the number of clusters
- Arbitrarily shaped clusters
- Notion of noise and robust to outliers
- Only three hyperparameters that can be set by domain experts.

## Limits

- Does not behave well for data with differing density across the parametric space

# Advantages and limits

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

## Advantages

- Does not require to choose the number of clusters
- Arbitrarily shaped clusters
- Notion of noise and robust to outliers
- Only three hyperparameters that can be set by domain experts.

## Limits

- Does not behave well for data with differing density across the parametric space
- Hyperparameter selection

# Questions

Lecture 11:  
DB-Scan

Sophie Robert

Introduction

Principle

Algorithm

Hyperparameters

Advantages  
and limits

Questions ?