

Clustering can be based on -

- i) Distance
- ii) Density
- iii) Structure

- \* Introduction to DBSCAN and Hierarchical Clustering
- \* DBSCAN Core Idea and key parameters (eps, min\_samples)
- \* Hierarchical Clustering & Linkage methods and Dendrograms
- \* Implementation of DBSCAN and Hierarchical Clustering
- \* Model Evaluation & Discussion

## Introduction to DBSCAN & Hierarchical Clustering

Why K-Means Alone is Not Enough

First, K Means required the number of clusters to be fixed in advance.

Second, K means assumed clusters are spherical

Third, K means does not handle noise well

DBSCAN:

Based on the idea of density

Ignores isolated points

Hierarchical:

Based on structure

## DBSCAN Core Idea & Key Parameters (eps, min\_samples)

DBSCAN: Density-Based spatial clustering of Application with Noise

↳ eps (epsilon)

↳ min\_samples (min pts)

### \* Understanding eps:

$A(0,0)$ ,  $B(0.1,0.1)$ ,  $C(0.2,0.1)$

$D(3,3)$

$$d(A,B) = \sqrt{(0.1-0)^2 + (0.1-0)^2} \\ = 0.14 < 0.3$$

$$d(A,C) = 0.22 < 0.3$$

$$d(A,D) = 4.24 \not< 0.3$$

Let,

$$\text{eps} \approx 0.3 \rightarrow (3)$$

### \* Understanding min\_samples:

Suppose,

min\_samples = 3

point A's eps-neighborhood: A, B, C

↳ It fulfils density condition as per min\_samples

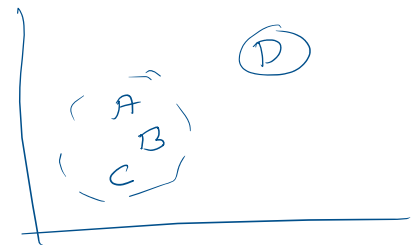
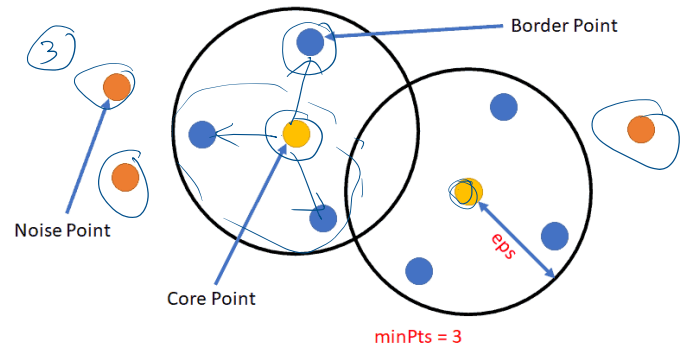
point D's eps-neighborhood: D

↳ It doesn't fulfil the condition

### \* Core, Border, Noise

Core point:

... is a point that has at least min\_samples



### Core point:

A core point is a point that has at least min-samples points within its eps neighborhood.

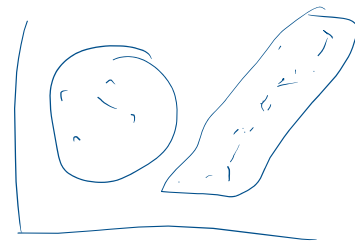
### Boarder point:

A boarder point lies within the eps neighborhood of a core point but does not itself have enough neighbors to be considered a core point.

### Noise point:

A noise point is neither a core point nor a boarder point.

- point A is a core point
- point B & C are also core points
- point D is noise point



## Hierarchical Clustering & Linkage Methods / Dendrograms

### Aglomerative clustering

#### Example:

$P_1 = (1,1)$ ,  $P_2 = (1,2)$ ,  $P_3 = (5,5)$ ,

$P_4 = (6,5)$

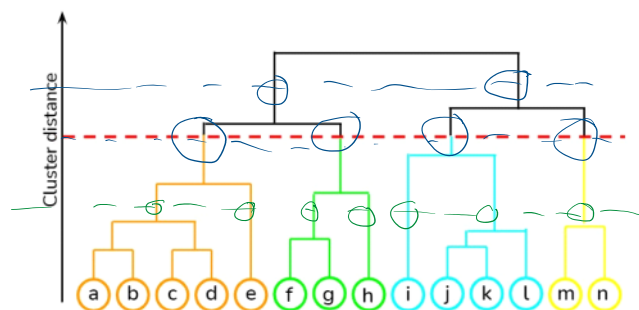
Step 1: → Each point is its own cluster  
Total number of clusters = 4  
 $\{P_1\}$ ,  $\{P_2\}$ ,  $\{P_3\}$ ,  $\{P_4\}$

#### Step 2:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

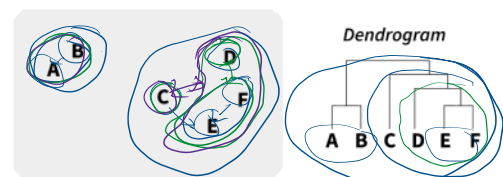
$$d(P_1, P_2) = 1$$

... 1.66

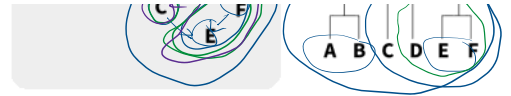


#### What is a Dendrogram?

A **dendrogram** is a tree that shows how clusters are merged step-by-step. We cut the dendrogram at a certain height to form final clusters.



$$\begin{aligned}
 d(P_1, P_2) &= 1 \checkmark \\
 d(P_1, P_3) &= 5.66 \\
 d(P_1, P_4) &= 6.40 \\
 d(P_2, P_3) &= 5 \\
 d(P_2, P_4) &= 5.83 \\
 d(P_3, P_4) &= 1 \checkmark
 \end{aligned}$$



Smallest distance:  
 $d(P_1, P_2) = 1$  and  $d(P_3, P_4) = 1$

Step-3: First merge:  
 $C_1 = \{P_1, P_2\}$      $C_2 = \{P_3, P_4\}$   
 $h \rightarrow 2$

Step-4: Distance between clusters (Linkage concepts)  
 we will use, average linkage

$$\begin{aligned}
 &d(P_1, P_3), d(P_1, P_4) \\
 &d(P_2, P_3), d(P_2, P_4)
 \end{aligned}$$

$$\frac{5.66 + 6.40 + 5 + 5.83}{4} = 5.72$$

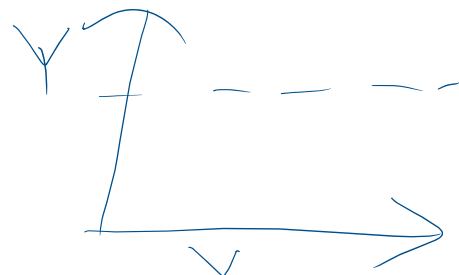
Distance between cluster  $C_1$  and  $C_2$

Step-5: Final merge:

$$\{P_1, P_2, P_3, P_4\}$$

What this process achieves:

- i) Records when clusters are merged
  - ii) Records at what distance they are merged
- ...date clusterings



- i) Records ..
- ii) Records at what distance they are ..
- iii) preserves all intermediate clusterings

