

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 Dendograms
- * 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:

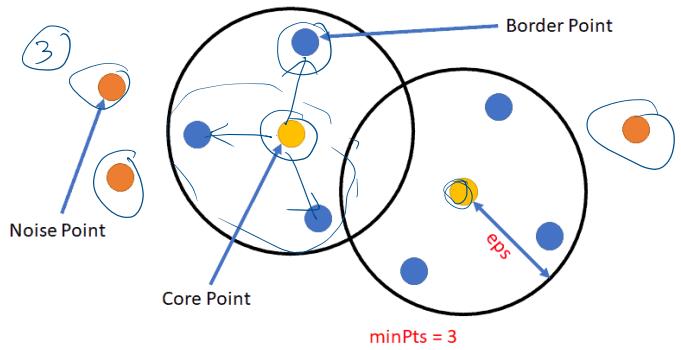
$A(0,0), B(0.1, 0.1), C(0.2, 0.1)$

$D(3,3)$

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

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

$$d(A, D) = 4.25 > 0.3$$



Let's say $\text{eps} = 0.3$ $\Rightarrow 3$

* Understanding min-samples:

Suppose,

min-samples = 3

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

point A's eps-neighborhood: A, B, C
point D's eps-neighborhood: D
→ It fulfills density condition as per min-samples

point D's eps-neighborhood: D
→ It doesn't fulfill 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 ϵ s neighborhood.

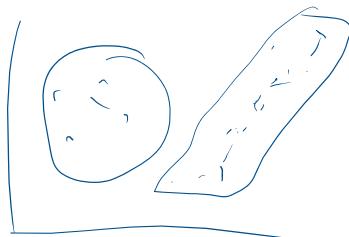
Border point:

A border point lies within the ϵ s 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 border point.

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



Hierarchical Clustering & Linkage Methods / Dendrograms

A agglomerative clustering

Example:

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

$$P_4 = (6,5)$$

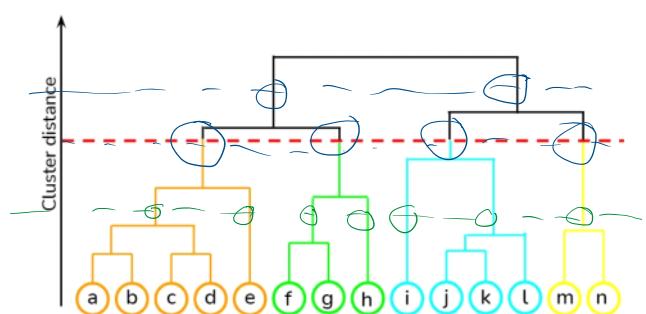
Step 1: → Each point is its own cluster
Total number of clusters = 4

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

Step 2:

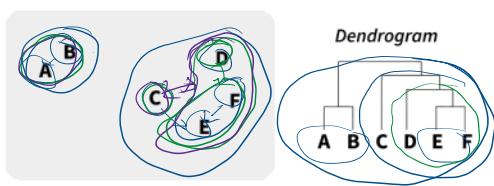
$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

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



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)
Distance between clusters

We will use, average linkage

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

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

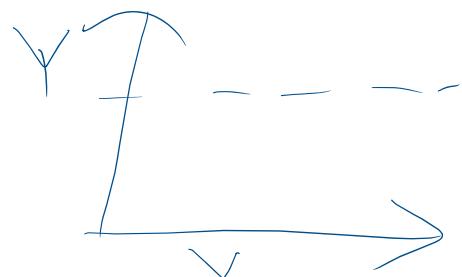
Distance between cluster C_1 and C_2

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
- iii) Records intermediate clusterings



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

