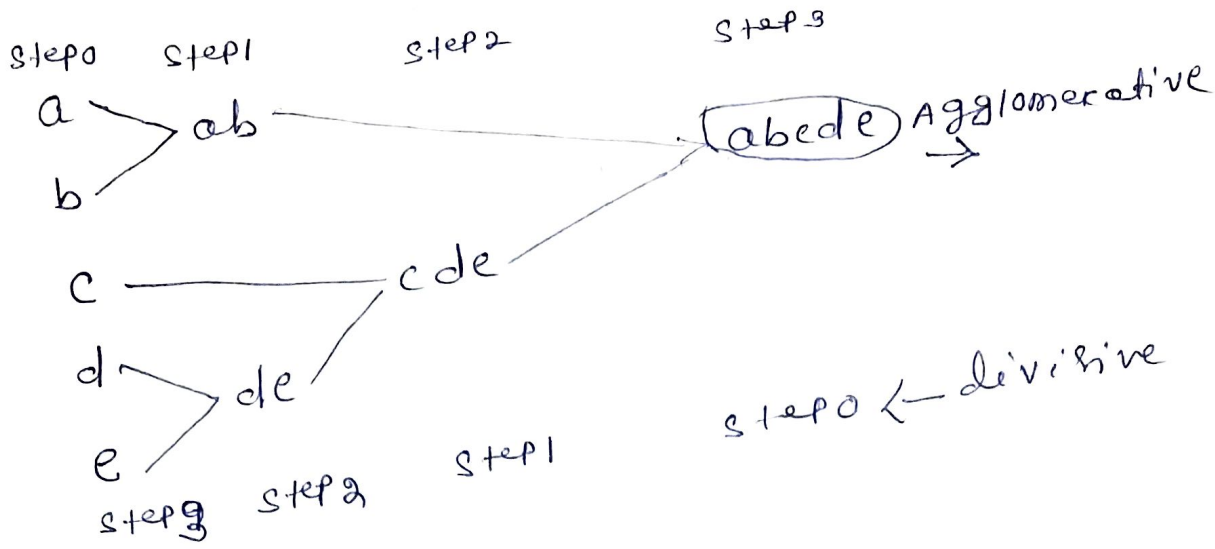
Hierarchical clustering (unsupervised)

1) Agglomerative (bottom-up)

2) Divisive (top-down)



Cluster distance measure

Ex

a	b	c	d	e
1	2	4	5	6

$C_1 = \{a, b\}$

$C_2 = \{c, d, e\}$

Distance Matrix.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Single link

$dist(C_1, C_2) = \min\{d(a,c), d(a,b), d(a,e), d(b,c), d(b,d), d(b,e)\}$

$= \min\{3, 4, 5, 2, 3, 4\} = 2$

Complete link

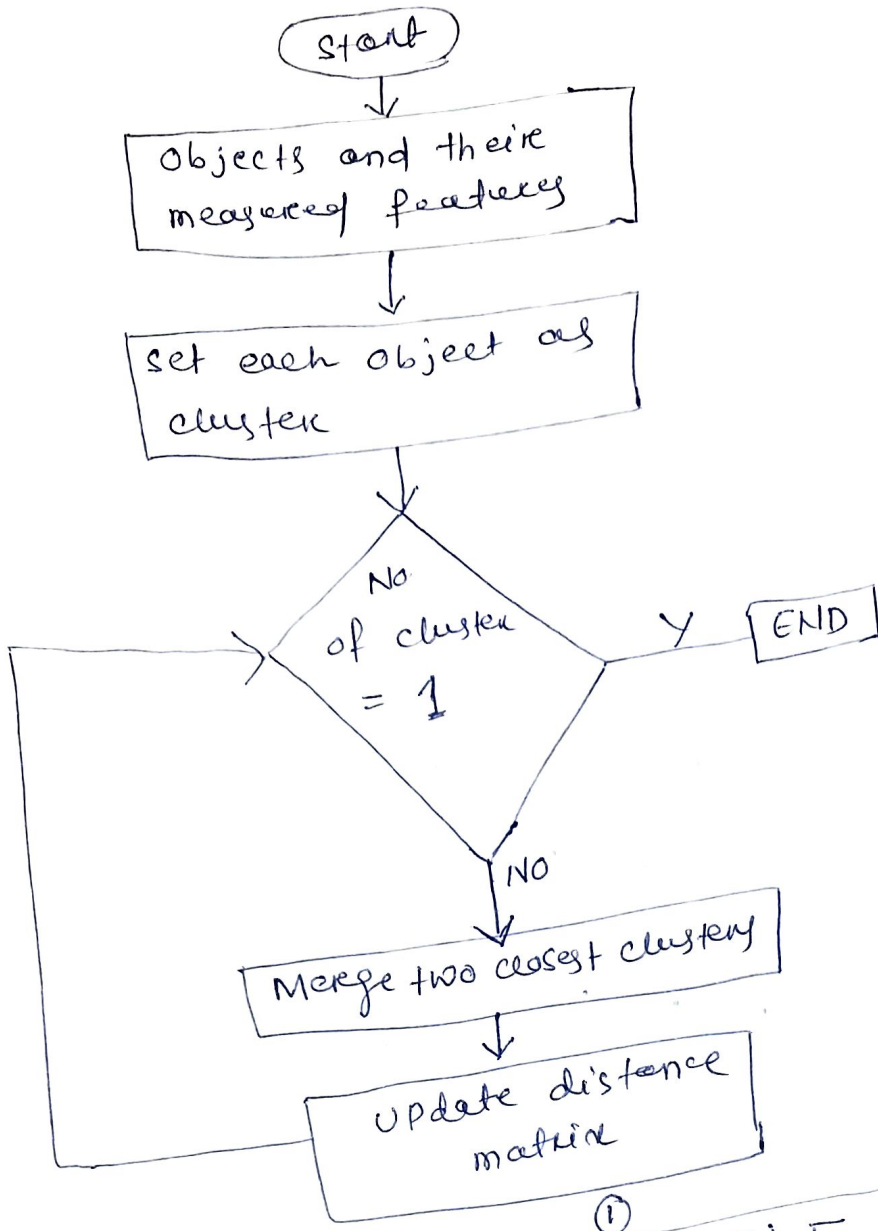
$dist(C_1, C_2) = \max\{d(a,c), d(a,b), d(a,c), d(b,c), d(b,d), d(b,e)\}$

$= \max\{3, 4, 5, 2, 3, 4\} = 5$

Average link $dist(C_1, C_2) = \frac{3+4+5+2+3+4}{6}$

$= \frac{21}{6} = 3.5$

Algorithm (Agglomerative)



Example

	x_1	x_2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

①

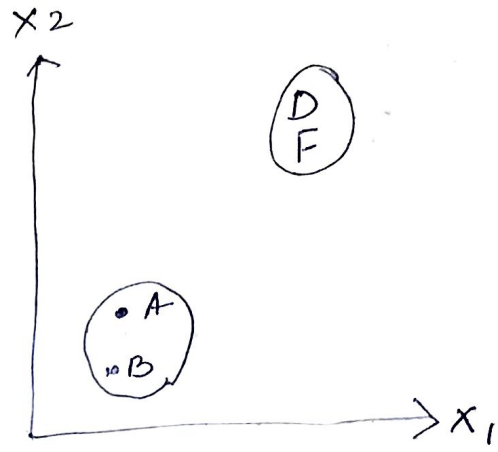
	A	B	C	D	E	F
A	0	0.71	5.66	3.61	4.24	3.2
B	171	0	4.95	2.92	3.54	2.5
C	5.66	4.95	0	2.24	1.41	2.5
D	3.61	2.92	2.24	0	1	0.5
E	4.24	3.54	1.41	1	0	1.12
F	3.2	2.50	2.5	0.5	1.12	0

$d_{AB} = \sqrt{(1-1.5)^2 + (1-1.5)^2} = 0.7071$

$d_{DF} = \sqrt{(3-3)^2 + (4-3.5)^2} = 0.5$

②

	A	B	C	DF	E
A	0	0.71	5.66	3.20	4.24
B	0.71	0	4.95	2.5	3.54
C	5.66	4.95	0	2.24	1.41
DF	3.20	2.50	2.24	0	1
E	4.24	3.54	1.41	1.00	0



single linkage

$d_{DF \to A} = \min(d_{DA}, d_{FA}) = \min(3.61, 3.20) = 3.20$

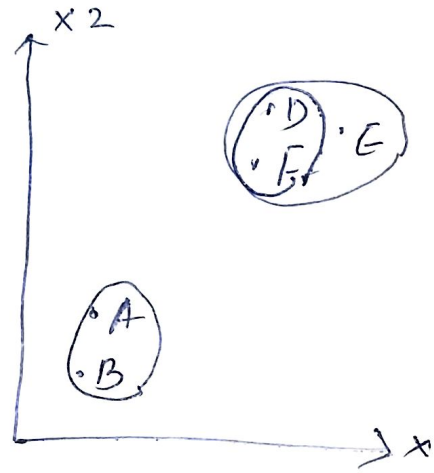
$d_{DF \to B} = \min(d_{DB}, d_{FB}) = \min(2.92, 2.50) = 2.50$

$d_{DF \to C} = \min(d_{DC}, d_{FC}) = \min(2.24, 2.50) = 2.24$

$d_{DF \to E} = \min(d_{DE}, d_{FE}) = \min(1, 1.12) = 1$

③

	AB	C	DF	E
AB	0	4.95	2.5	3.5
C	4.95	0	2.24	1.41
DF	2.5	2.24	0	1
E	3.54	1.41	1	0



$d_{(C \to AB)} = \min(d_{CA}, d_{CB}) = \min(5.66, 4.95) = 4.95$

$d_{DF \to AB} = \min(d_{DF \to A}, d_{DF \to B})$

$= \min(3.2, 2.5) = 2.5$

$d_{E \to AB} = \min(d_{EA}, d_{EB}) = \min(4.24, 3.54)$

$= 3.54$

④

	AB	C	DFE
AB	0	4.95	2.50
C	4.95	0	(1.41)
DFE	2.50	(1.41)	0

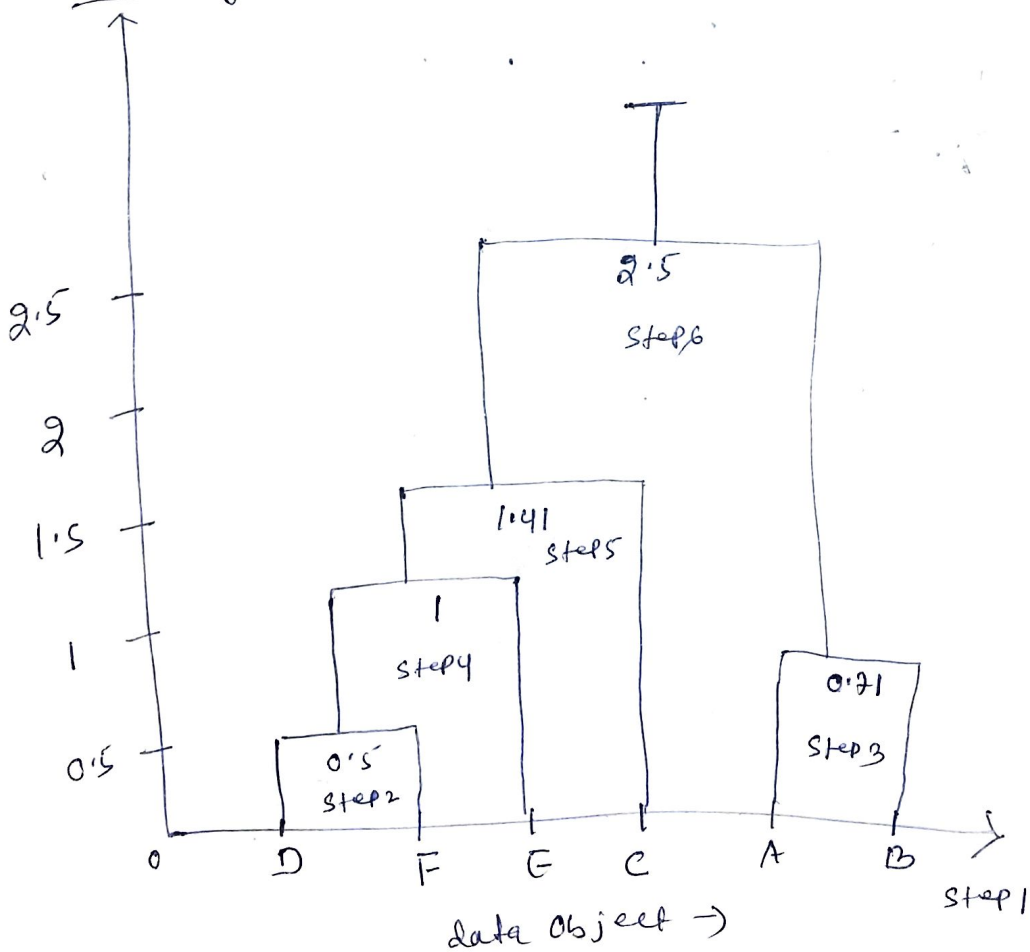
$d_{AB \to DFE}$
$= \min(d_{AB \to DF}, d_{AB \to E})$
$= \min(2.5, 3.5) = 2.5$

$d_{C \to DFE}$
$= \min(d_{C \to DF}, d_{C \to E})$
$= \min(2.24, 1.41) = 1.41$

	AB	DFEC
AB	0	(2.50)
DFEC	(2.50)	0

$d_{AB \to DFEC}$
$= \min(d_{AB \to DFE}, d_{AB \to C})$
$= \min(2.5, 4.95) = 2.5$

Dendrogram tree representation



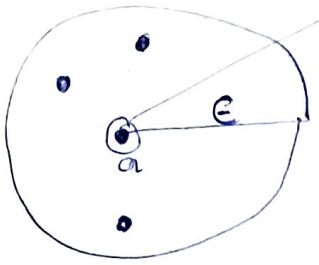
DBSCAN (Density based spatial clustering of applications with Noise)

Density → No. of points in a given area

Inputs - Epsilon (ϵ)

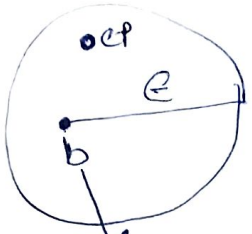
Minimum points = Let it be 3

ϵ, represents the radius to draw a circle for clustering.



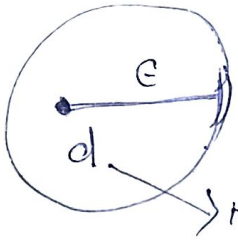
→ core point (CP)

[It can detect Noise or outliers easily.]

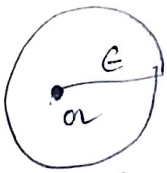


Border point (BP)

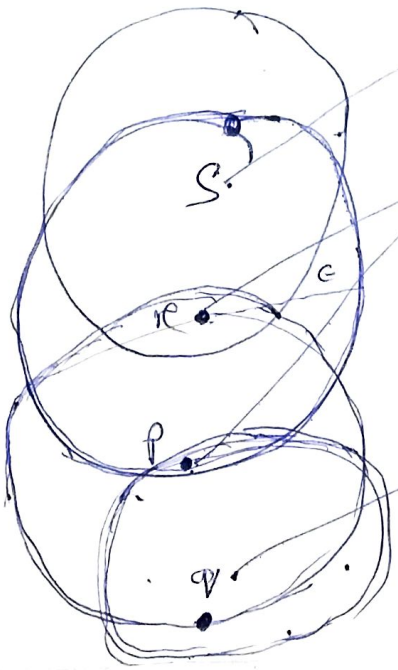
part of the cluster because it is within ϵ of a core point but does not meet the min-point criteria.



No corepoint
Not satisfying min-point criteria
→ Noise Point (NP)



Noise Point or outlier



→ Border point.

→ corepoint

→ Border point.

Ex

	B/N A	B/N B	B/N C	CP D	CP E	CP F
A	0	0.7	5.7	3.6	4.2	3.2
B	0.7	0	4.9	2.9	3.5	2.5
C	5.9	4.9	0	2.2	1.4 ✓	2.5
D	3.6	2.9	2.2	0 ✓	1	0.5
E	1.2	3.5	1.4	1 ✓	0 ✓	1.5
F	3.2	2.5	2.5	0.5 ✓	1.1 ✓	0 ✓

Let $\epsilon = 1.5$

min-point ≥ 3.

C - BP (Boarder point).
as closer to E a CP

B - Noise point

A - Noise point.