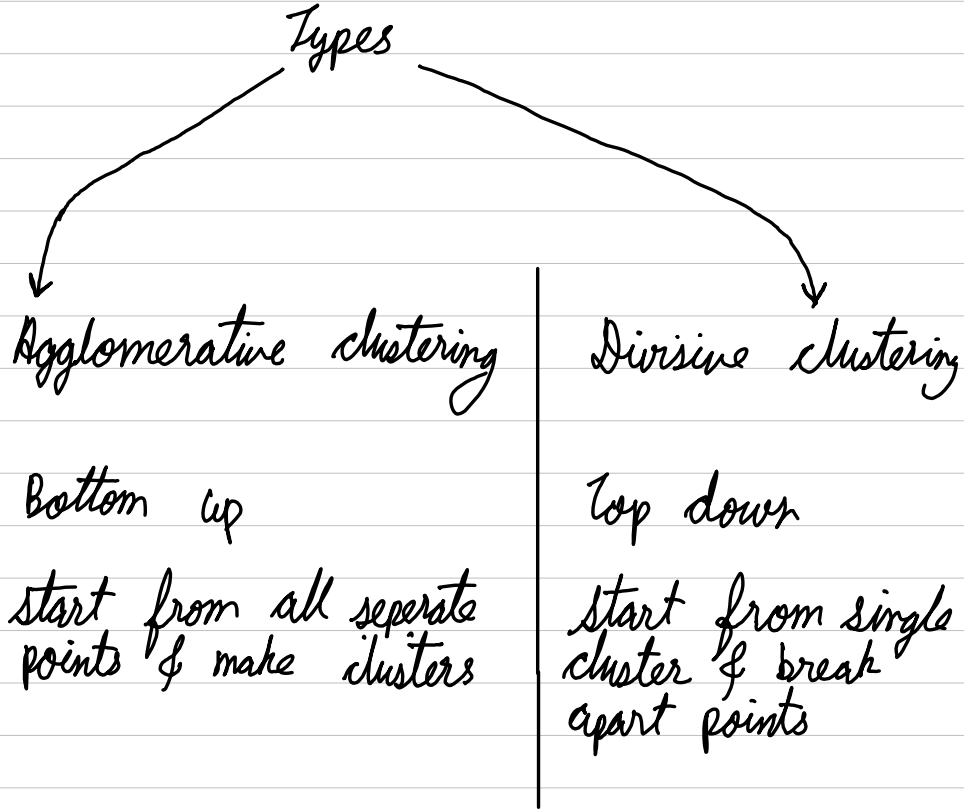
Heirchical Clustering

Unsupervised clustering algorithm

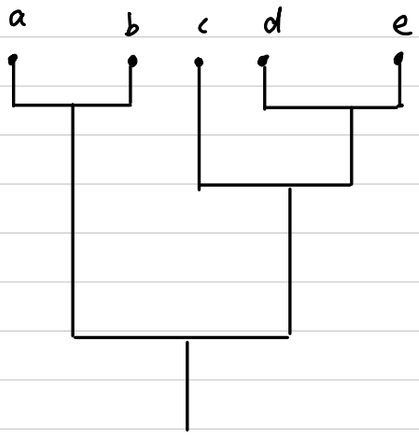
It works via grouping data into a tree of clusters

steps → 1. Identify clusters close together

2. Merge the clusters



Hierchical Clustering is based on dendograms



Points that are close together form clusters together

The graph is called dendogram

At first level, the group formed is

(a, b) c (d, e)

The similarity is based on distance creiterion

For that, all points are compared

In next step, c is checked.
It is more close with cluster (d, e) so it is merged.

At level 2

(a, b) , (c, d, e)

At level 3

(a, b, c, d, e)

At last level, all points are in single cluster

The process is heirarchical hence called hierchical clustering

(Q1) Solve using herchical clustering

	X	Y
P1	0.4	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

→

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.234	0				
P3	0.22	0.14	0			
P4	0.37	0.19	0.13	0		
P5	0.34	0.143	0.27	0.284	0	
P6	0.23	0.24	0.1	0.219	0.38	0

P3, P6 have smallest value

∴ (P3, P6) will be first cluster

So we combine P_3 P_6

	P_1	P_2	$(P_3\ P_6)$	P_4	P_5
P_1	0				
P_2	0.234	0			
$(P_3\ P_6)$	0.22	0.14	0		
P_4	0.37	0.19	0.13	0	
P_5	0.34	0.143	0.28	0.284	0

$$d((A,B),C) = Min\{d(A,C), d(B,C)\}$$

Less value is considered. Average value is not taken

eg $Min((P_3P_6), P_4) = Min(0.13, 0.22)$
$= 0.13$
etc ...

$Min((P_3P_6), P_5) = Min(0.28, 0.34) = 0.28$

For old values, dont update

Minimum Value of the table is 0.13

Combine P_4 with P_3 P_6 $(P_4, (P_3, P_6))$

	P_1	P_2	$(P_4(P_3,P_6))$	P_5
P_1	0			
P_2	0.23	0		
$(P_4,(P_3P_6))$	0.22	0.14	0	
P_5	0.34	0.14	0.23	0

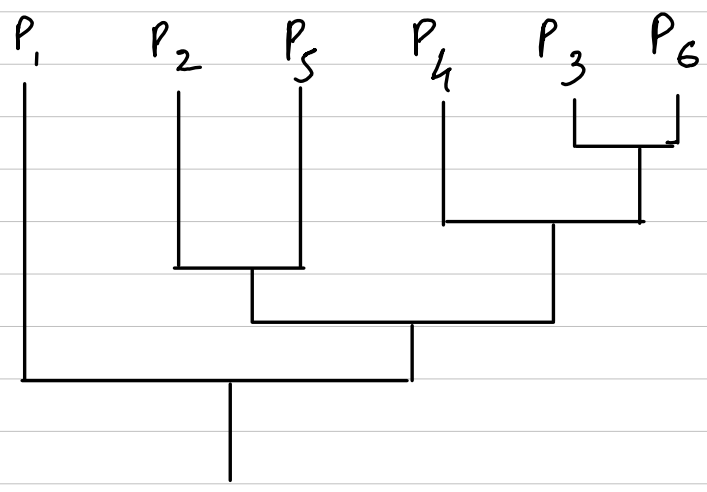
There are two Minimum values.
lets consider P_2 P_5 combine

	P_1	$(P_2\ P_5)$	$(P_4(P_3,P_6))$
P_1	0		
$(P_2\ P_5)$	0.23	0	
$(P_4(P_3P_6))$	0.22	0.14	0

Now combine clusters $(P_2\ P_5)$ and $(P_4(P_3, P_6))$

	P_1	$((P_2P_5)(P_4,(P_3,P_6)))$
P_1	0	
$((P_2P_5)(P_4,(P_3P_6)))$	0.22	0

Lastly combine P_1 to make the last cluster



steps for Hierchical Clustering

step 1: Make adjustancy table

step 2: Combine points with smallest value

step 3: Make new adjustancy table where distance of point from cluster is minimum of all distances of point from the points in the cluster

step 4: Repeat steps 2 & 3 N times

step 5: Make Dendogram

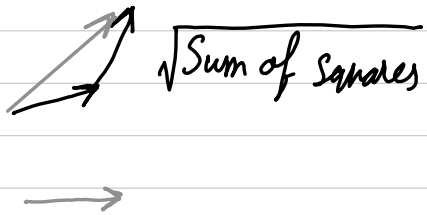
Types of distance

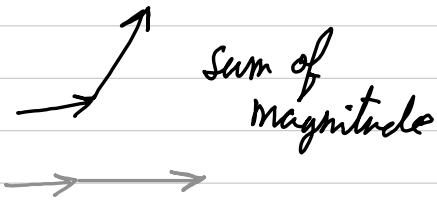
Euclidian

$\sqrt{\sum x^2}$

Manhatten

$\sum ||x||$





In last example we have used the euclidian distance

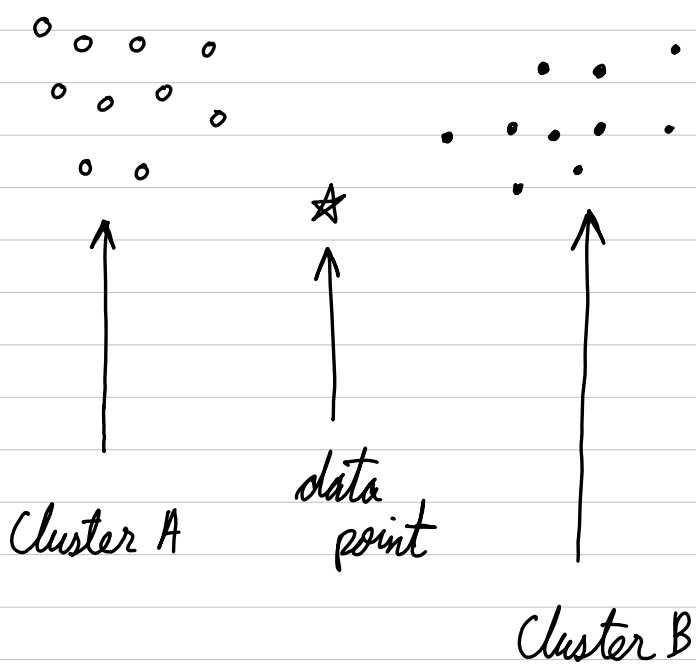
Both distances will give different clusters

Closeness Measures

We need to define what "closeness" means

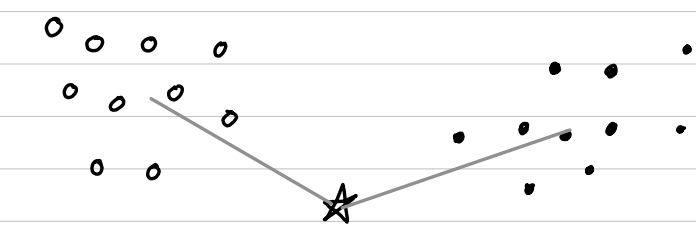
In last example, we used closeness as distance.

Other Methods also can be used



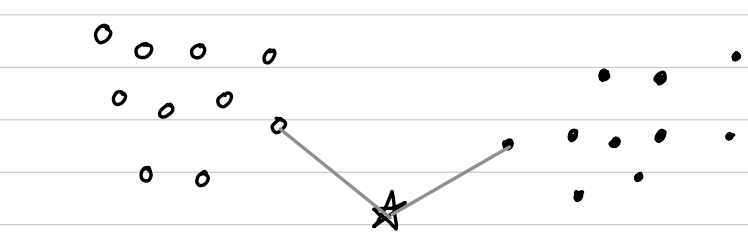
① Centroid Method

Distance of data from a cluster is distance of data from the average of all data points (ie centroid)



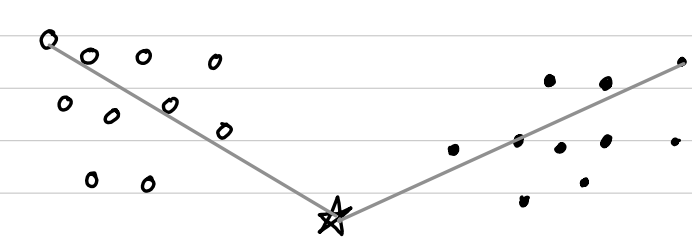
② Single linkage

Closest point to cluster

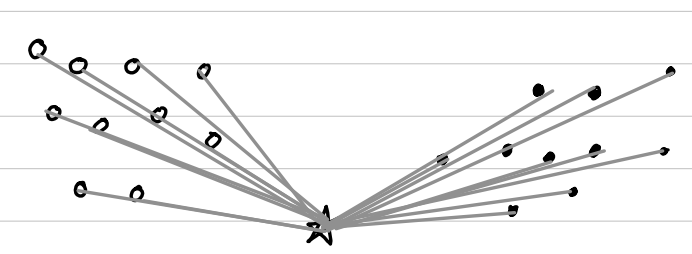


③ Complete linkage

Furthest point from cluster



④ Average linkage



Average of all the points

Advantages

Handels non convex clusters

Handels clusters of different sizes & densities

Handel Missing data & noise

Reveals the hierarchical structure that can be used to understand relationships between data

Deterministic results → No need for initial seed

Drawbacks

Need to define stopping criterion

High computational cost & memory requirement

Time complexity $O(n^3)$

Too slow even for medium data sets