*Clustering

$X_{n \times d} \rightarrow X = \bigcup_{j=1}^{k} X^{(j)}$ → union of clusters

k → clusters → 1 to K

k-means algorithm

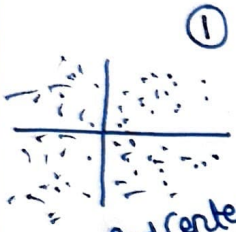
Find k disjoint clusters $X^{(1)}, \ldots X^{(k)}$ such that, if $\mu_1, \ldots \mu_j$ is the average point in cluster $X^{(j)}$, the following quantity is minimized

Minimize Distortion

$$\text{distortion} = \sum_{j=1}^{k} \sum_{x \in X^{(j)}} \|x - \mu_j\|^2$$

$\|x - \mu_j\|^2$ → distance (euclidean)

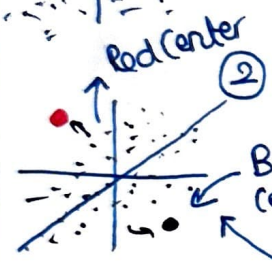
μ_j → avg in j^{th} cluster → mean of points in each cluster



Basic Algorithm Steps

1) Random Initialization: Find k random centers

$\mu_1, \ldots \mu_k$ → avg of clusters

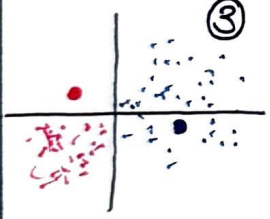


Repeat

2) Assignment Step: Assign each point to the nearest center

That makes K clusters

$X_1, \ldots X_k$

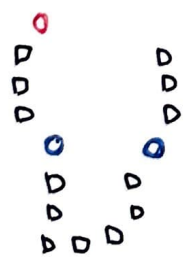


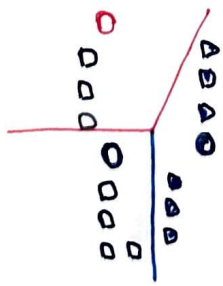
find new centers

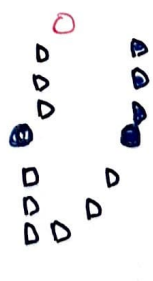
3) Update Step: Replace the previous centers with the average point in each cluster

$$\mu_j = \frac{1}{|X^{(j)}|} \sum_{x \in X^{(j)}} x$$

Repeat this until no change in cluster memberships







Voronoi diagram of space

Issues

→ The distortion objective is appropriate for spherical and coherent clusters

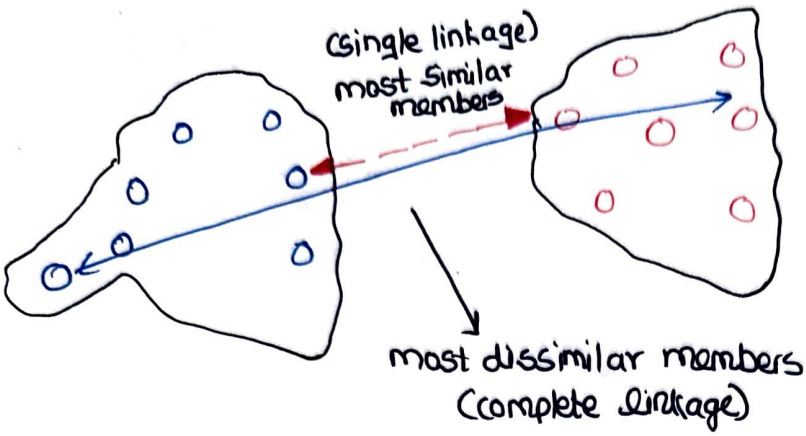
→ The algorithm is not guaranteed to find the best clusters for the given objective

multiple initialization

→ The algorithm can be very slow to converge, so in practice we have to implement stopping criteria

$2^{\sqrt{n}}$ steps

* Agglomerative Hierarchial Clustering



Distance between Clusters

2 options → one is most similar
other is most dissimilar

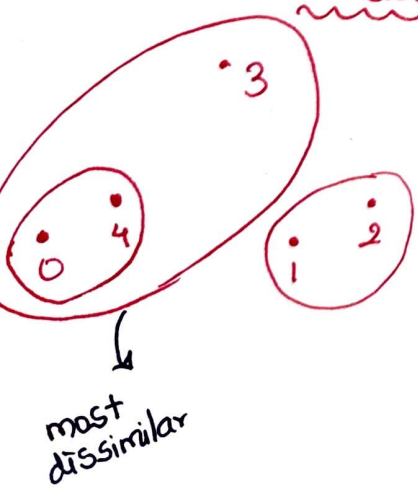
	ID_0	ID_1	ID_2	ID_3	ID_4
ID_0	0.00	4.97	5.51	5.89	3.85
ID_1	4.97	0.00	4.34	5.10	6.69
ID_2	5.51	4.34	0.00	7.24	8.31
ID_3	5.89	5.104	7.24	0.00	4.38
ID_4	3.83	6.69	8.31	4.38	0.00

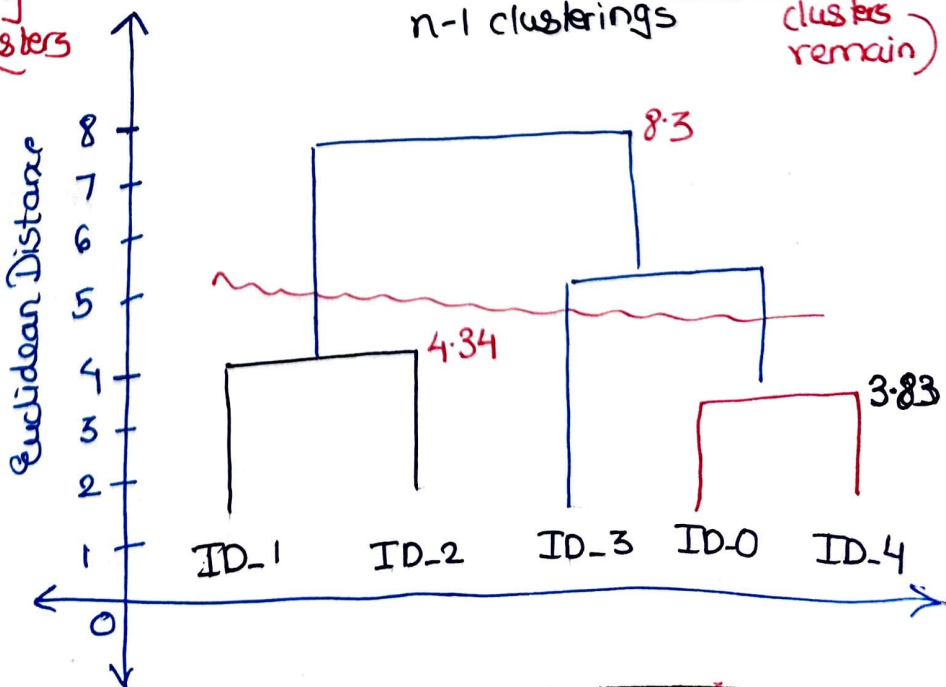
same → obvious

1) Start with individual clusters
2) Compute pairwise distances (single, complete linkage)
3) Merge closest clusters
4) Update the distances (between newly formed clusters)
5) Repeat (until one/ desired no of clusters remain)

n-1 clusterings

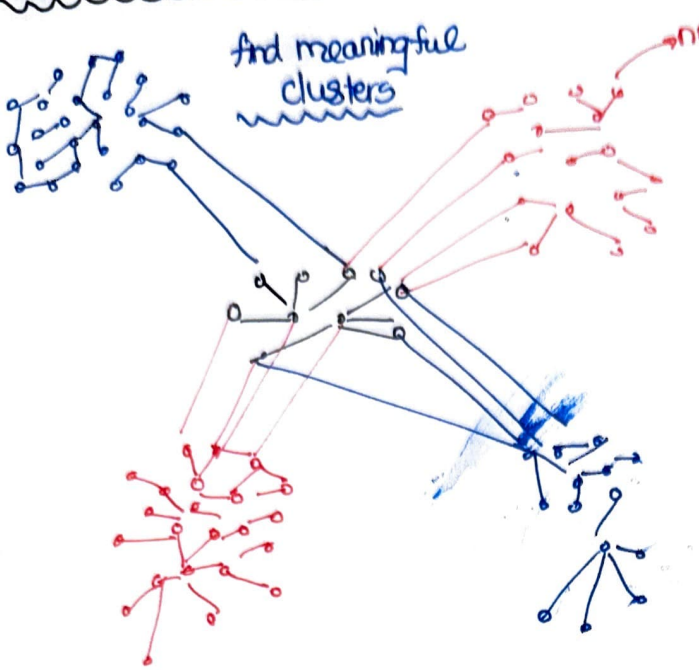
Repeatedly merging closest clusters





* How to Cluster Networks :-

Clustering Networks → Social media, Genome. etc



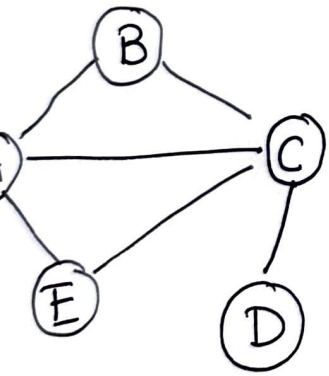
$\varphi(C) = \frac{\#\text{connections to other clusters}}{\#\text{connections inside cluster}}$

fee objective ↑

Fi Value

find clusters with small Fi value

Undirected Graph



Adjacency matrix

X	A	B	C	D	E
A	0	1	1	0	1
B	1	0	1	0	0
C	1	1	0	1	1
D	0	0	1	0	0
E	1	0	1	0	0

degrees

$D = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 1 \\ 2 \end{pmatrix}$

diagonal matrix.

$M = D^{-1} A$

k eigenvalues and eigenvectors.

λ_1, v_1

(λ_i, v_i)

k top eigenvalues) put into columns.

$Y = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}$ n × k

matrix of points