Clustering (Grouping together similar entities)

$(\mathbb{X}, d(\ ,\ ))$

$\mathbb{X} = \{x_1, x_2, \ldots, x_N\}$

$d : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}^+$

$x_i \in \mathbb{R}^n$

- Clustering Hypothesis

  Objects which are similar must be grouped together in same cluster

  Objects which are not similar should be assigned to different clusters.

$\mathbb{X} = \{x_1, \ldots, x_N\}$

$P = \{X_1, X_2, \ldots, X_K\} \longrightarrow$ Partition of $\mathbb{X}$

$\forall i,j \; X_j \subseteq \mathbb{X}, \quad X_i \cap X_j = \phi, \quad \bigcup_{i=1}^{k} X_i = \mathbb{X}$

Take home quiz.

Q. $N = 4$ how many distinct Partions of $\mathbb{X}$?

$x_1, x_2, x_3, x_4 \longrightarrow P_1 = \{\{\}, \{x_1, x_2, x_3, x_4\}\}$

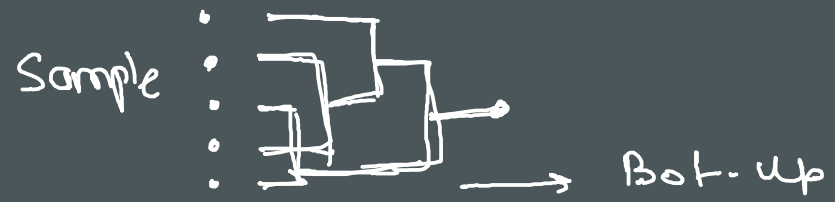
$P_2 = \{\{x_1\}, \{x_2, x_3, x_4\}\}$ $P_3 = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\}$

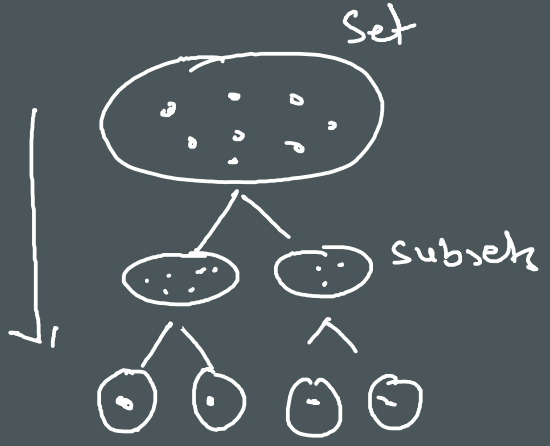
$\{x_1, x_2, x_3, x_4\}$

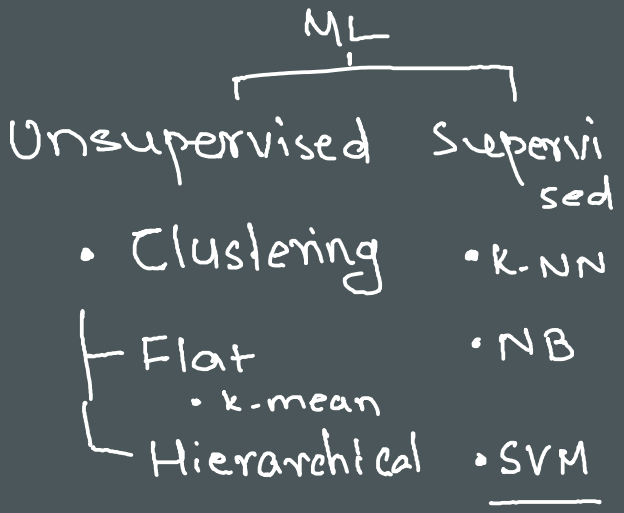
$\{x_1\}$ $\{x_2\}$ $\{x_3\}$ $\{x_4\}$

Hierarchical Clustering

- Top-down
- Bottom - up



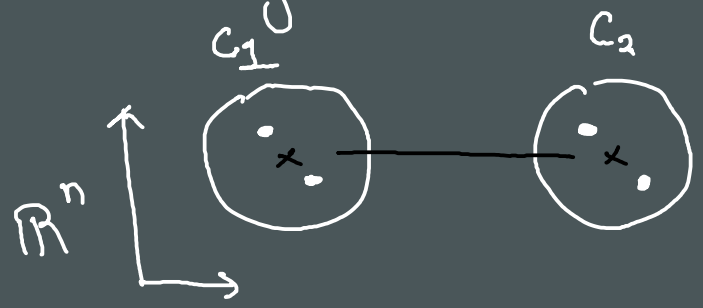




- Comparision

- Hierarchical Agglomerative Clustering (HAC) BU↑

→ Merge — Similarity



$(X, d(,))$ ⟹ Metric Space

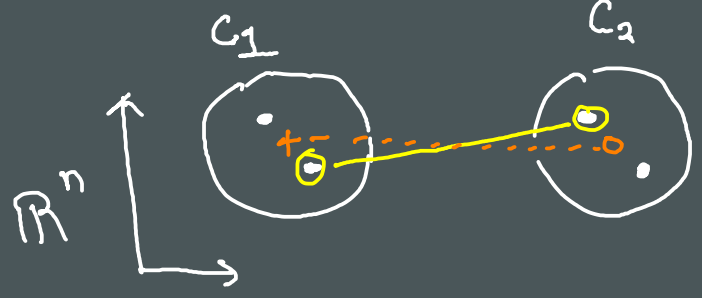
1. $d \geq 0$
2. $d(x,y) = 0 \Rightarrow x = y$
3. $d(x,z) \leq d(x,y) + d(y,z)$ ← triangle inequality

→ distance metric in $\mathbb{R}^n$ $\quad \| t_1 - t_2 \| \quad \langle t_1 \mid t_2 \rangle$

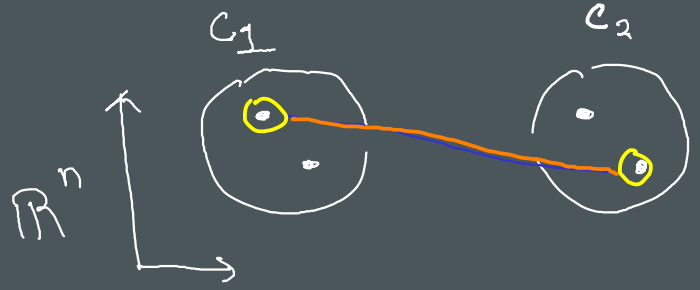
$X = \{x_1, \ldots, x_N\}$ $\quad d(x_i, x_j)$

(i) Centroids $\quad sim(\bar{C}_1, \bar{C}_2)$ $\quad \bar{C}_1, \bar{C}_2 \in \mathbb{R}^n$

(ii) Single - link



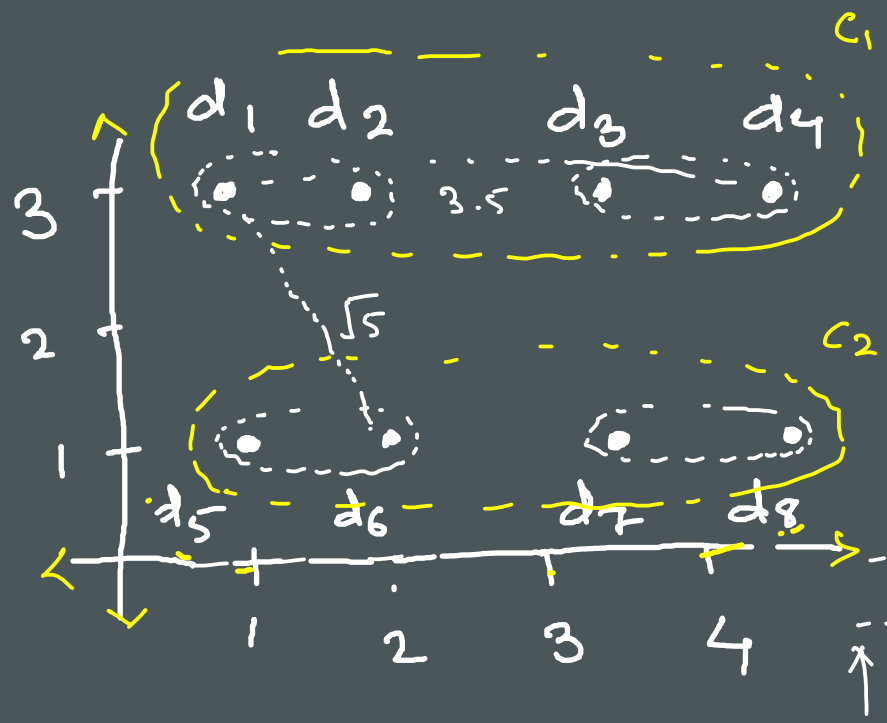
(iii) Complete- link

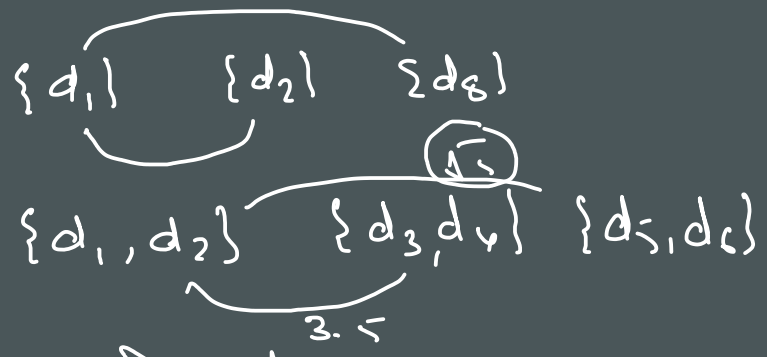


$$d(C_1, C_2) = \min_{\substack{x_1 \in C_1 \\ x_2 \in C_2}} d(x_1, x_2)$$

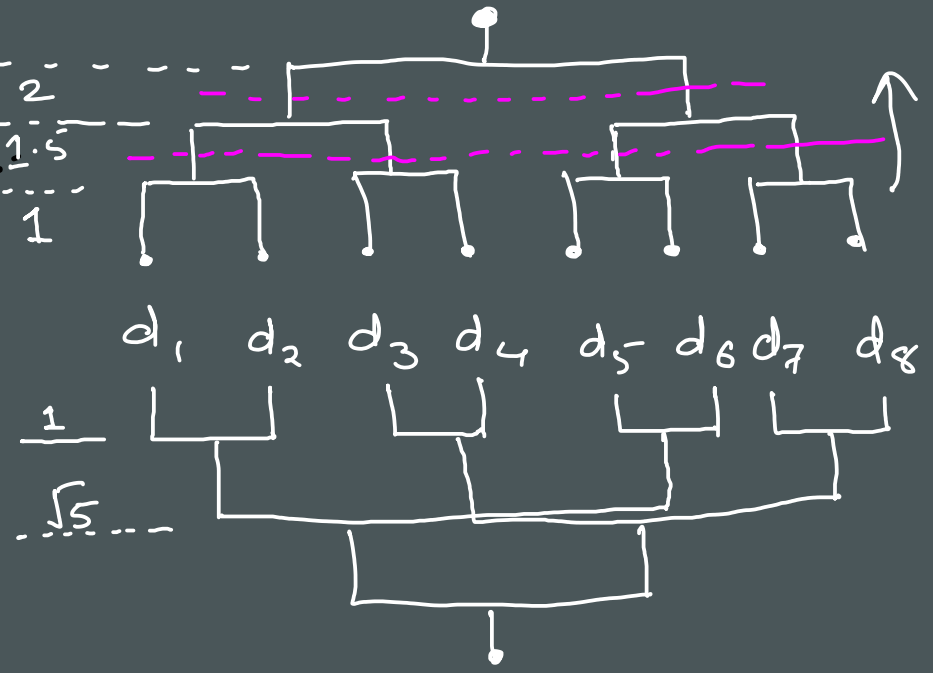
$$d(C_1, C_2) = \max_{\substack{x_1 \in C_1, \\ x_2 \in C_2}} d(x_1, x_2)$$

: "dendrogram"





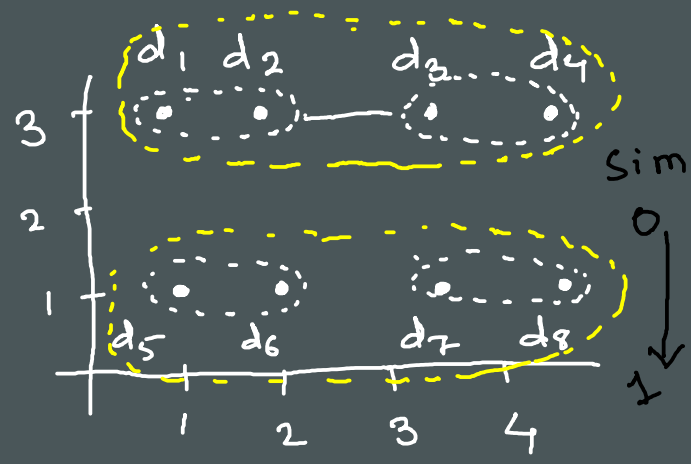
Dendrogram

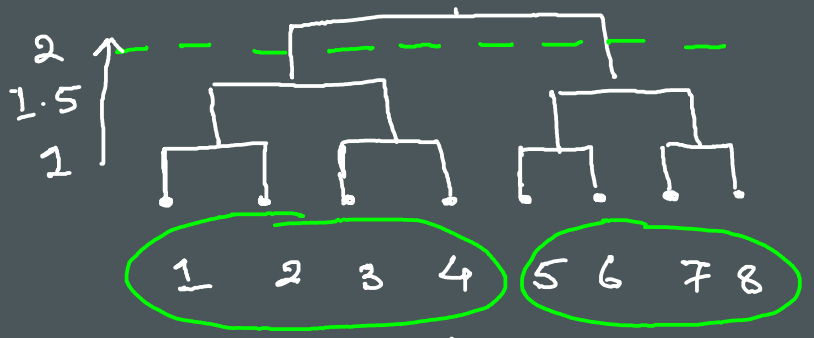


$$d(c_i, c_j) = \min_{d_i \in c_i,\, d_j \in c_j} d(d_i, d_j)$$

$$d(c_i, c_j) = \max_{d_i \in c_i,\, d_j \in c_j} d(d_i, d_j)$$

K-means





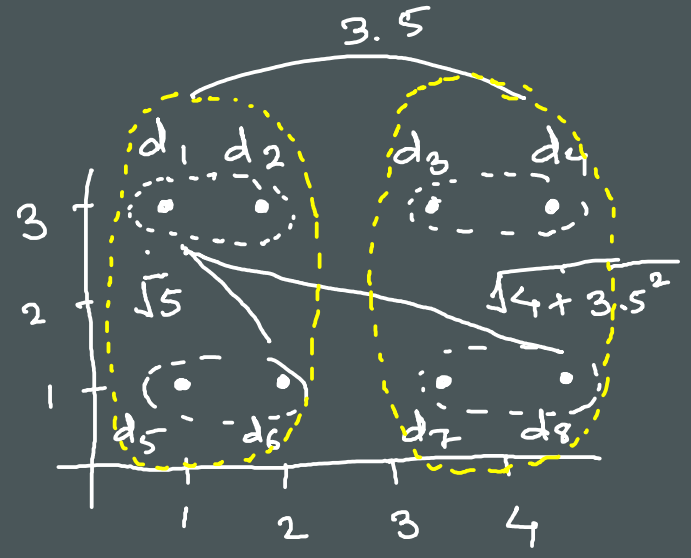
Dendrogram

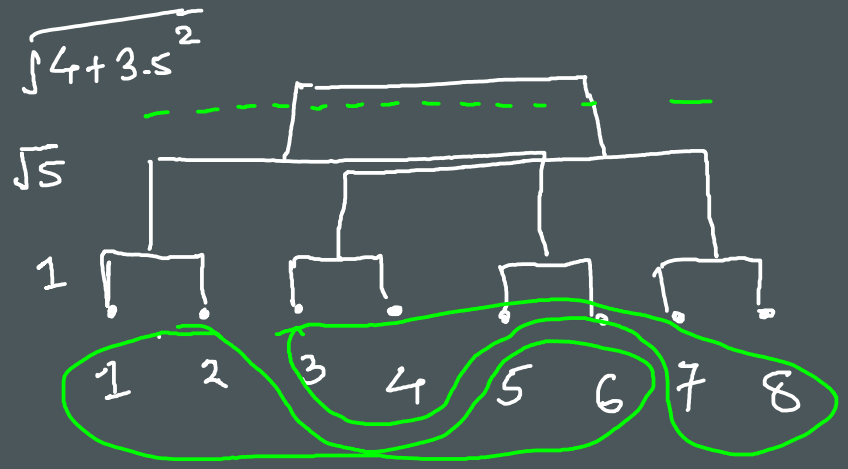
Single-Link

min $d(d_i, d_j)$

$d_i \in c_i$

$d_j \in c_j$





Complete-Link

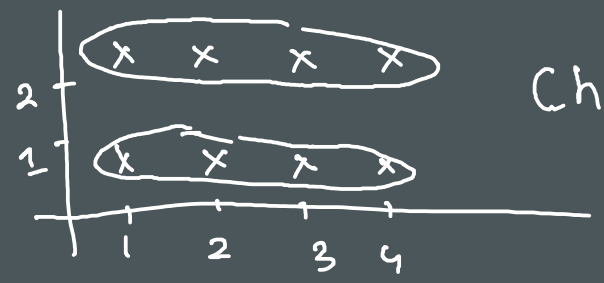
max $d(d_i, d_j)$

$d_i \in c_i$

$d_j \in c_j$

- Reduce the assessment of cluster quality to a single similarity b/w a pair of docs.
  - two most similar
  - two most dissimilar

Single link



Chaining!

- Time Complexity ?

N - $O(N^2)$ ~ $O(N^3)$

clist Merg

~ $O(N^2 \log N)$

- Complete link - susceptible to outliers

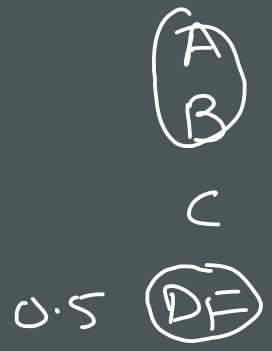
1+2ε, 4, 5+2ε, 6, 7-ε

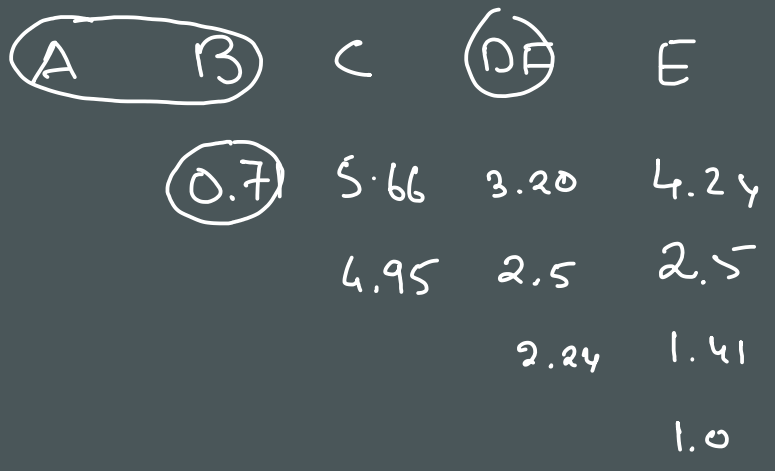
(ε>0)

Exercise

Explain Hierarchical Agglomerative Clustering in detail with three different 'merge' operations. We are given an input distance matrix of size 6 by 6 in the following Table. Entries in this matrix are calculated based on the geometry of feature vectors corresponding to the six observed document vectors. For the given distance matrix, sketch dendrograms and compare clustering results for at least two of the 'merge' strategies.

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B		0.00	4.95	2.92	3.54	2.50
C			0.00	2.24	1.41	2.50
D				0.00	1.00	0.50
E					0.00	1.12
F						0.00





single link

– min over all pairwise dist.

iterate similarly till you get one cluster!

Take home

Q : Why names < single link / Complete link ? (Graph Theory)