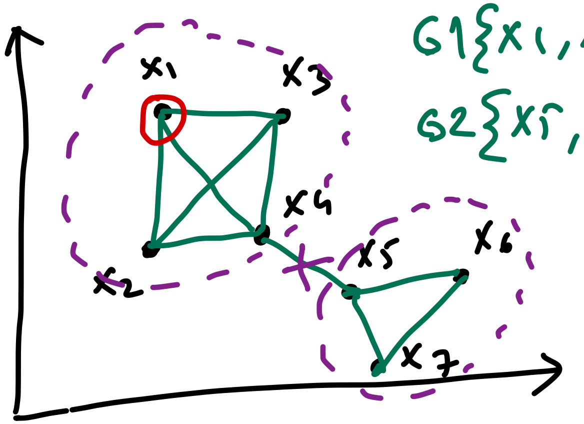
Spectral Clustering

- define local neighborhoods
- if the distance between x_i & x_j is smaller than a threshold, they are neighbors

threshold

$$b_{ij} = \begin{cases} 1 & \text{if } \|x_i - x_j\|_2 < \delta \\ 0 & \text{otherwise} \end{cases}$$



G1 $\{x_1, x_2, x_3, x_4$

G2 $\{x_5, x_6, x_7$

$$b_{ij} = \begin{cases} \exp\left[-\dfrac{\|x_i - x_j\|_2^2}{2\sigma^2}\right] & \text{if } \|x_i - x_j\|_2 < \delta \\ 0 & \text{otherwise} \end{cases}$$

$$b_{ii} = 0 \quad \forall i$$

$$d_{ii} = \sum_{j \neq i} b_{ij} \quad \forall i$$

$$d_{ij} = 0 \quad \forall (i \neq j)$$

$$B = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

connectivity or adjacency matrix

$$D = \begin{bmatrix} 3 & & & & & & \\ & 3 & & & & 0 & \\ & & 3 & & & & \\ & & & 4 & & & \\ & 0 & & & 3 & & \\ & & & & & 2 & \\ & & & & & & 2 \end{bmatrix}$$

Laplacian Matrix :

$$\underset{N\times N}{L} = \underset{N\times N}{D} - \underset{N\times N}{B}$$

↳ each row (column) sums up to 0.

Normalization

$$L_{RANDOM-WALK} = D^{-1} \cdot L = D^{-1} \cdot (D - B) = \boxed{I - D^{-1} \cdot B}$$

$$L_{SYMMETRIC} = D^{-1/2} \cdot L \, D^{-1/2} = D^{-1/2} \cdot (D - B) \cdot D^{-1/2} = \boxed{I - D^{-1/2} \cdot B \cdot D^{-1/2}}$$

SPECTRAL CLUSTERING

STEP #1: Find the eigenvectors of normalized L matrix.

STEP #2: Pick R smallest eigenvectors

STEP #3: Construct Z matrix as follows:

$$Z = \begin{bmatrix} v_1 & v_2 & \cdots & v_R \end{bmatrix}_{N\times R}$$

STEP #4: Run k-means clustering algorithm on Z matrix to find K clusters.

PARAMETERS

δ: threshold

R: # of eigenvectors to be included

K: # of clusters to be found.

Hierarchical Clustering

- finding groups such that instances (data points) in a group are $\boxed{\text{more similar}}$ to each other than instances in different groups.

closer

COMPONENT #1 : The Distance Function Between Data Points

distance $\Rightarrow$ dissimilarity

distance $\uparrow$ $\Rightarrow$ similarity $\downarrow$

distance $\downarrow$ $\Rightarrow$ similarity $\uparrow$

$$k(x_i, x_j) = \exp\left[-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right]$$

0 dissimilar

1 similar

0 similar

$+\infty$ dissimilar

0	0	0	1
1	1	0	1

Euclidean Distance

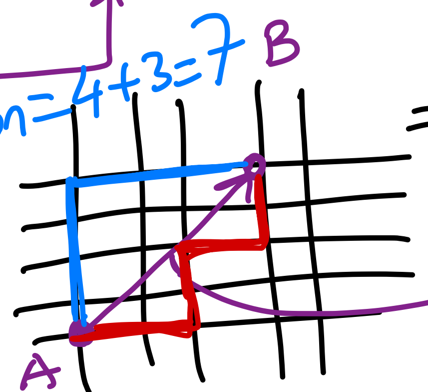
$$d(x_i, x_j) = \|x_i - x_j\|_2$$

$$= \sqrt{\sum_{d=1}^{D}(x_{id} - x_{jd})^2}$$

$$= \sqrt{x_i^T x_i - 2x_i^T x_j + x_j^T x_j}$$

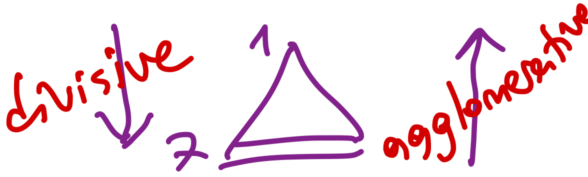
Manhattan Distance

$$d(x_i, x_j) = \sum_{d=1}^{D} |x_{id} - x_{jd}|$$



Euclidean distance $= \sqrt{4^2+3^2} = 5$

COMPONENT #2: The Direction to Proceed.



Agglomerative (bottom-to-top)

⇒ Combines small clusters into bigger ones

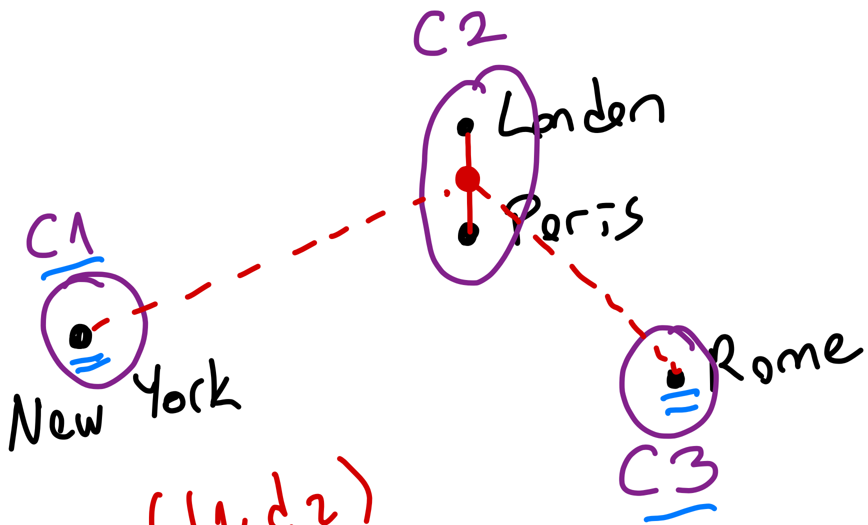
⇒ starts with "N" clusters

Divisive (top-to-bottom)

⇒ divides big clusters into smaller ones

⇒ starts with "1" cluster

COMPONENT #3: The Distance Function Between Groups of Data Points.



Distance (C1, C2)

Distance ((C2, C3)

Distance ((C1, C3)

d(NY, Paris) = d_1

d(NY, London) = d_2

Distance({New York}, {London, Paris})

Distance({London, Paris}, {Rome})

ave(d1, d2)

min(d1, d2)

max(d1, d2)

median(d1, d2)

Centroid Clustering

$$d(G_A, G_B) = \left\| \frac{\sum_{x_i \in G_A} x_i}{\underbrace{|G_A|}_{\text{cardinality}}} - \frac{\sum_{x_j \in G_B} x_j}{|G_B|} \right\|_2$$

Single-Link Clustering

$$d(G_A, G_B) = \min_{\substack{x_i \in G_A \\ x_j \in G_B}} d(x_i, x_j)$$

Complete-Link Clustering

$$d(G_A, G_B) = \max_{\substack{x_i \in G_A \\ x_j \in G_B}} d(x_i, x_j)$$

Average-Link Clustering

$$d(G_A, G_B) = \frac{\sum_{x_i \in G_A} \sum_{x_j \in G_B} d(x_i, x_j)}{|G_A|\,|G_B|}$$

⋮

$t=0$ 6 clusters

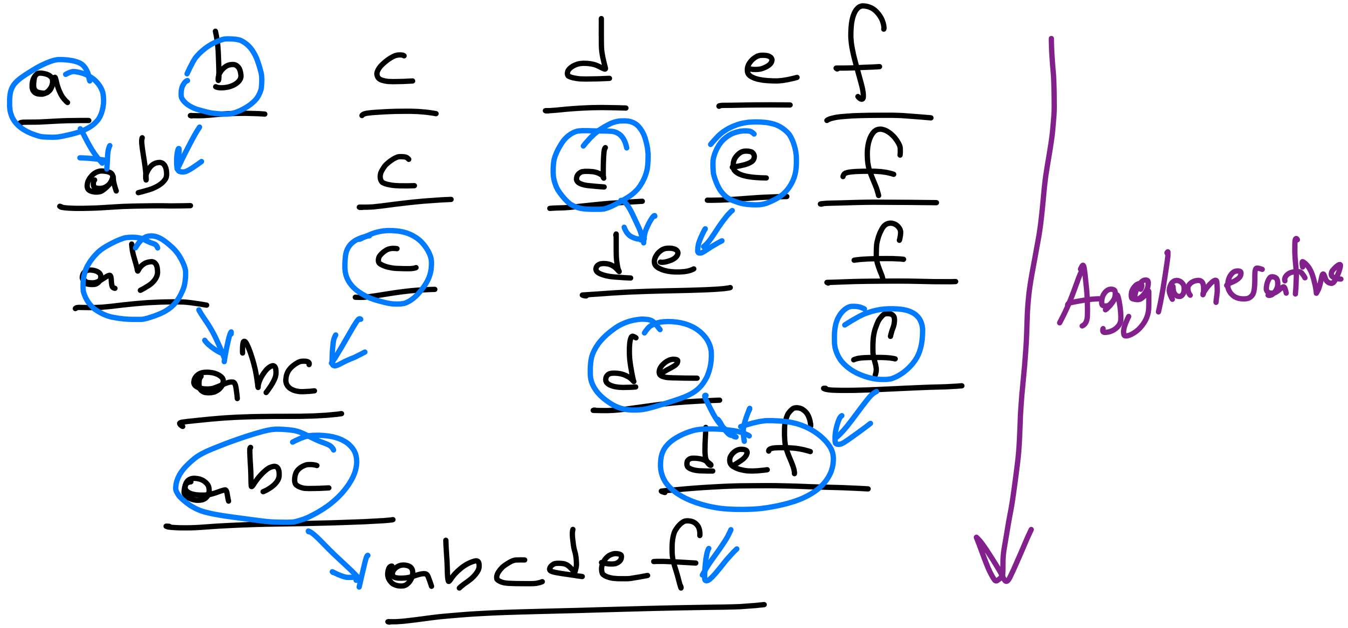
$t=1$ 5 clusters

$t=2$ 4 clusters

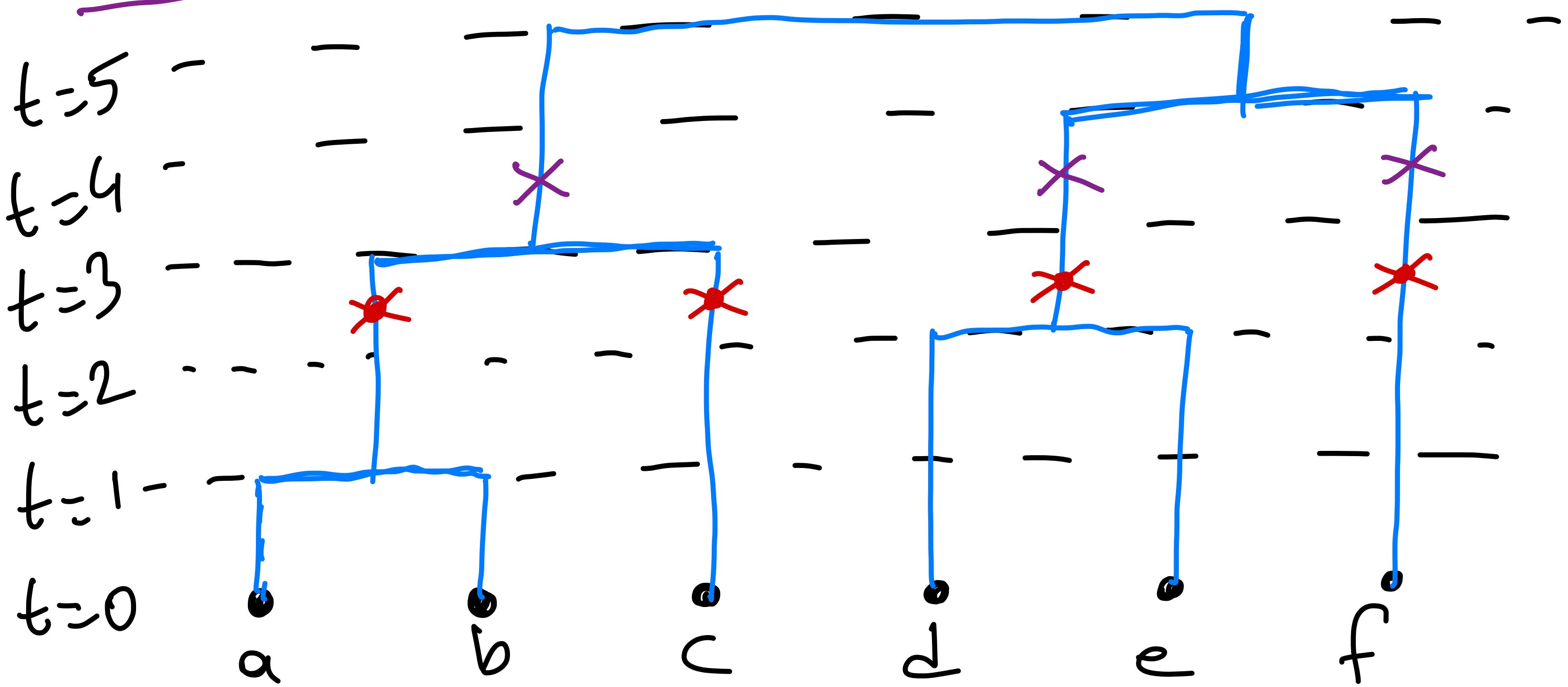
$t=3$ 3 clusters

$t=4$ 2 clusters

$t=5$ 1 cluster



Dendrogram



$K=3$ clusters

$C_1 = \{a, b, c\}$

$C_2 = \{d, e\}$

$C_3 = \{f\}$

$K=4$ clusters

$C_1 = \{a, b\}$

$C_2 = \{c\}$

$C_3 = \{d, e\}$

$C_4 = \{f\}$