Expectation-Maximization Clustering

$$\mathcal{X} = \{x_i\}_{i=1}^{N} \qquad \text{likelihood} \Rightarrow L(\Phi \mid \mathcal{X}) = \left[\prod_{i=1}^{N} p(x_i \mid \Phi)\right]$$

$$\log L(\Phi \mid \mathcal{X}) = \sum_{i=1}^{N} \log \Bigg[\underbrace{\sum_{k=1}^{K} p(x_i \mid C_k) \cdot \Pr(C_k)}_{\text{mixture densities}} \Bigg]$$

two sets of random variables

Z = cluster memberships

Φ = parameters $[\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K, \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_K]$

E-STEP: $$E\left[L_c(\Phi \mid \mathcal{X}, Z) \mid \mathcal{X}, \Phi^{(t)}\right]$$

M-STEP: $$\Phi^{(t+1)} = \arg\max_{\Phi} E\left[L_c(\Phi \mid \mathcal{X}, Z) \mid \mathcal{X}, \Phi^{(t)}\right]$$

E-STEP:

$$h_{ik} = E\left[z_{ik} \mid x, \Phi^{(t)}\right] = \frac{p(x_i \mid C_k, \Phi^{(t)})\Pr(C_k)}{\sum_{c=1}^{K} \underbrace{p(x_i \mid C_c, \Phi^{(t)})}_{\text{multivariate Gaussian density}}\Pr(C_c)}$$

$$h_{ik} \geq 0 \quad \forall (i,k)$$

$$\sum_{k=1}^{K} h_{ik} = 1 \quad \forall i$$

K columns

$$H = \left[\begin{array}{ccc} & h_{1k} & \\ & h_{2k} & \\ & \vdots & \\ & h_{Nk} & \end{array}\right] \; N \text{ rows}$$

each row sums up to 1.

column k

M-STEP:

$$\hat{\Pr}^{(t+1)}(C_k) = \frac{\sum_{i=1}^{N} h_{ik}}{N} \quad \forall k$$

$$\hat{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik} \cdot x_i}{\sum_{i=1}^{N} h_{ik}} \quad \forall k$$

$$\hat{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik}\left(x_i - \hat{\mu}_k^{(t+1)}\right)\left(x_i - \hat{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} h_{ik}}$$

Hierarchical Clustering

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

COMPONENT #1: The distance function between pairs of data points

distance $\Rightarrow$ dissimilarity

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

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

squared distance

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

similarity = 1 similarity = 0 distance = 0 distance = $+\infty$

$$\boxed{\left(\sum_{d=1}^{D}(|x_{id} - x_{jd}|)^p\right)^{1/p}}$$

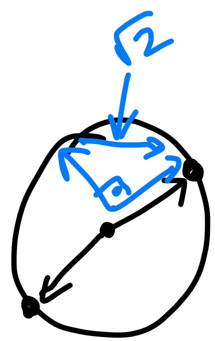
$$d(x_i, x_j) = \sum_{d=1}^{D} |x_{id} - x_{jd}|$$

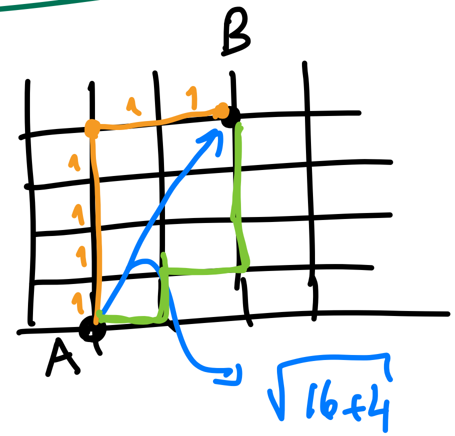
Manhattan Distance

Euclidean Distance

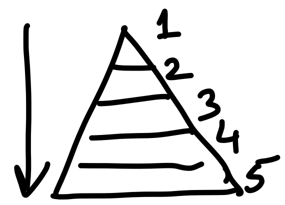
$$d(x_i, x_j) = \|x_i - x_j\|_2 = \sqrt{(x_i - x_j)^T (x_i - x_j)}$$

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





COMPONENT #2 : The direction to proceed



Divisive (top-to-bottom)

→ divides big clusters into smaller ones

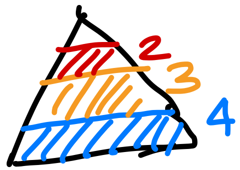
→ starts with "1" cluster

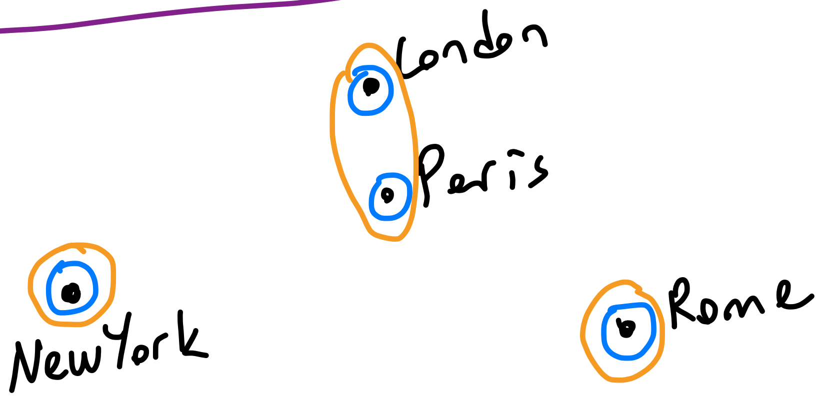
Agglomerative (bottom-to-top)

→ combines small clusters into bigger ones

→ starts with "N" clusters

COMPONENT #3: The distance function between groups of data points





distance ({London, Paris}, New York)

distance ({London, Paris}, Rome)

distance (New York, Rome)

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

midpoint

$\hookrightarrow$ cardinality

Single-Link Clustering: $d(G_A, G_B) = \min_{\substack{x_i \in G_A \\ x_j \in G_B}} d(x_i, x_j)$

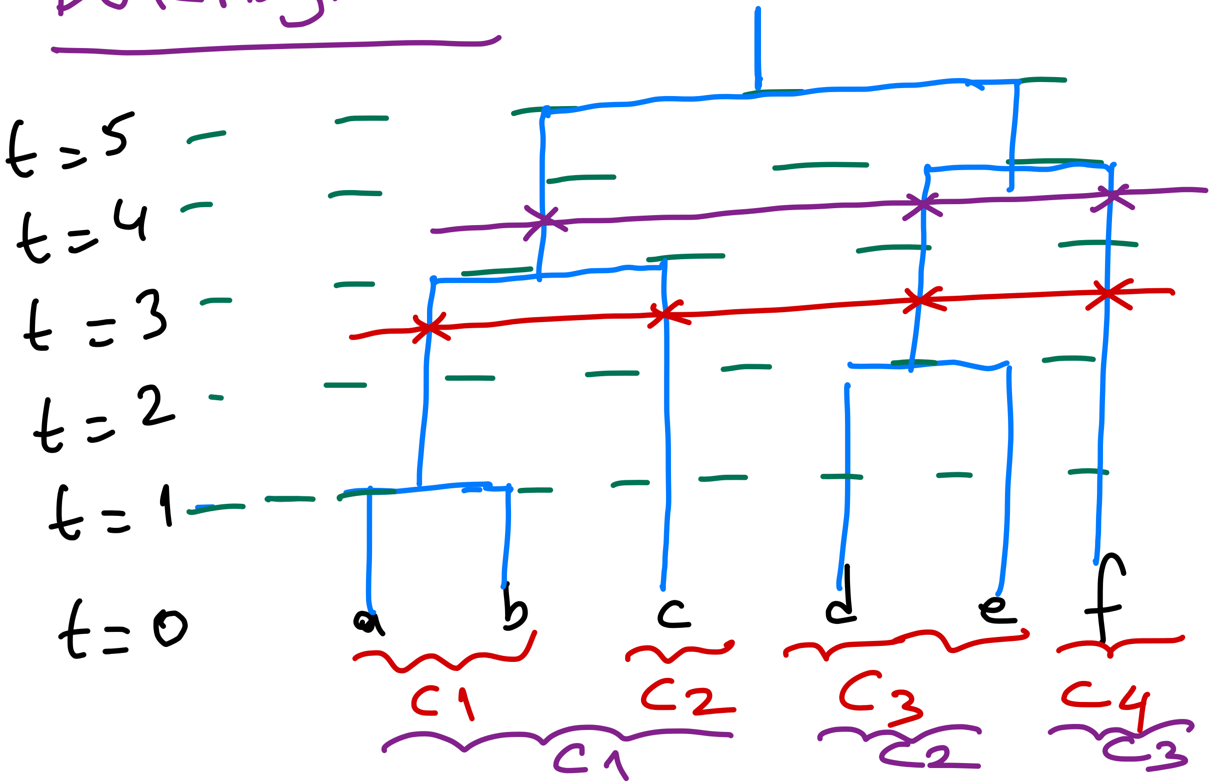
Complete-Link Clustering: $d(G_A, G_B) = \max_{\substack{x_i \in G_A \\ x_j \in G_B}} d(x_i, x_j)$

Average-Link Clustering: $d(G_A, G_B) = \dfrac{\sum_{x_i \in G_A} \sum_{x_j \in G_B} d(x_i, x_j)}{|G_A|\,|G_B|}$

$\vdots$

t = 0	6 clusters	a b c d e f
t = 1	5 clusters	ab c d e f
t = 2	4 clusters	ab c de f
t = 3	3 clusters	abc de f
t = 4	2 clusters	abc def
t = 5	1 cluster	abcdef

Dendrogram



K = 4 clusters

K = 3 clusters