

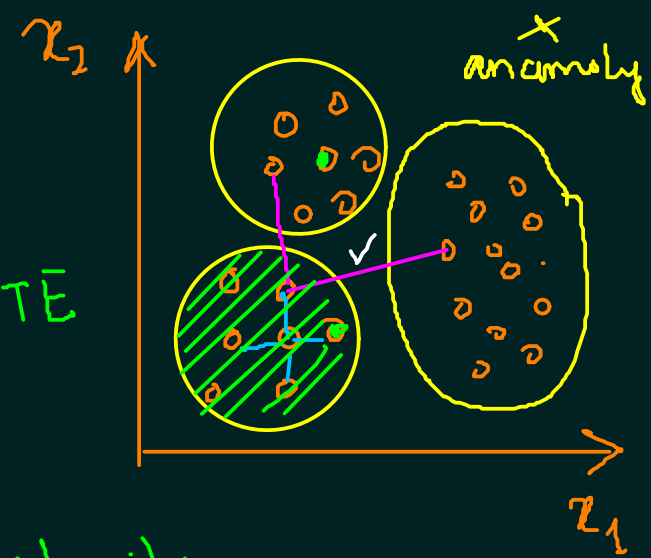
Supervised Learning

[Inp, Out] ← Training
 (label)

- Minimize Intra-cluster dist
 - Maximize Inter-cluster dist
- Goal → between clusters?

Unsupervised Learning

< data, ? > } TE
 Clustering



Similarity between points?

Distance between Points:

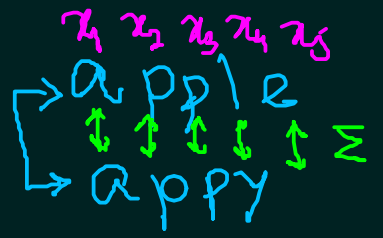
→ Symmetry: $D(a,b) = D(b,a)$



② Reflexivity: $D(a,a) = 0$

③ Triangular Inequality:

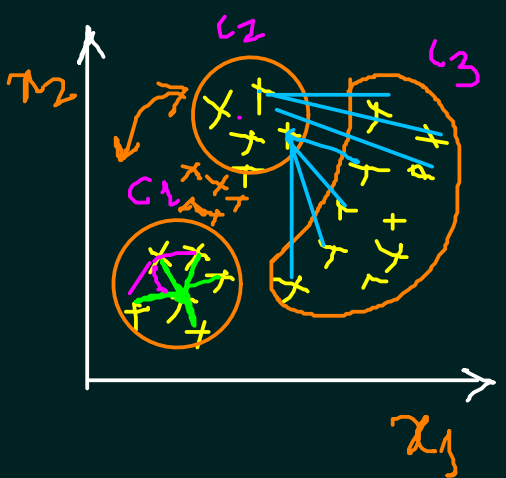
dist = \sum dist between letters



Similarity $\propto \frac{1}{dist}$



$$D(a,b) + D(b,c) \geq D(a,c)$$



→ Evaluation: Scatter coefficient

$$SC = \frac{\text{avg. intra-cluster dist}}{\text{avg. inter-cluster dist}}$$

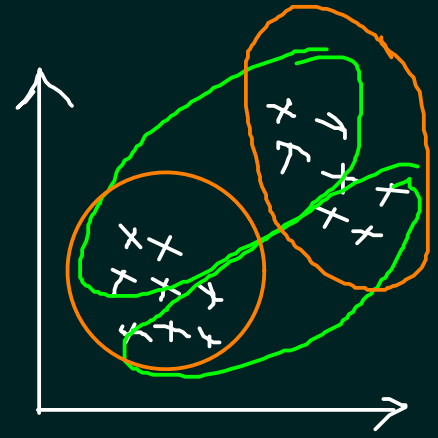
→ Computationally efficiency? }
 → # clusters - how many? } ??

Exp → k=3,4,2

3 ≤ k ≤ 8 ← domain knowledge

SC → Cohesion vs Separation

k=2

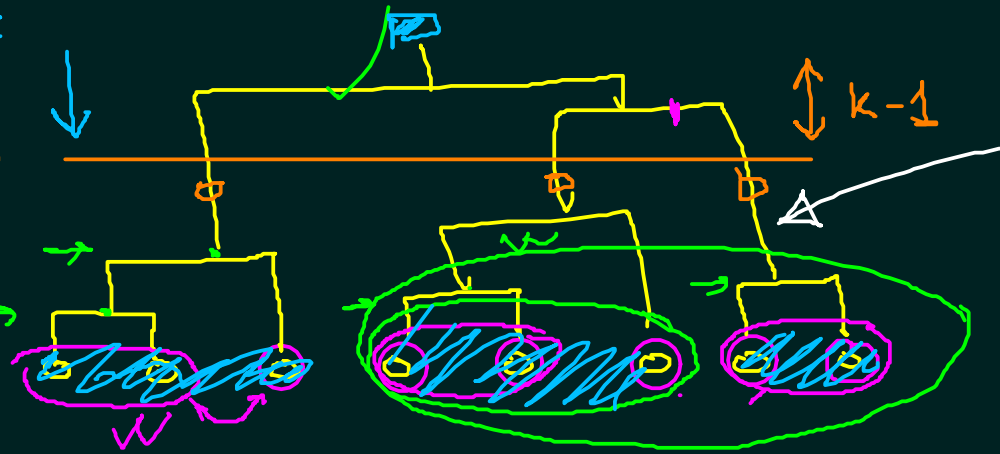


Clustering → Partitioning

Hierarchical clustering

Partitional clustering

n-1
k

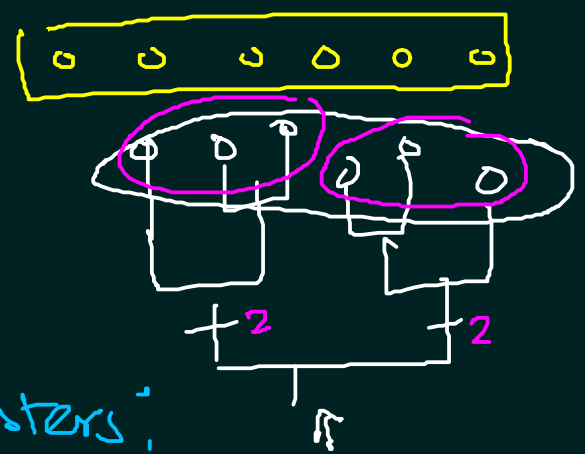
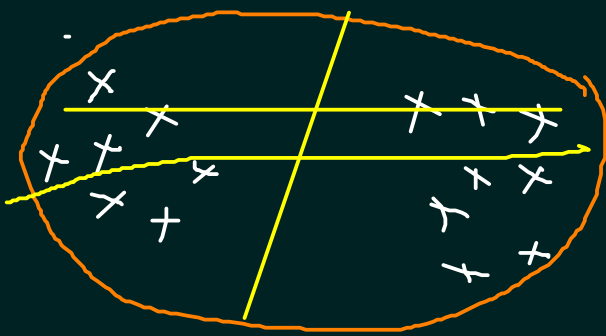


bottom up

dendrogram (tree)

→ dist(a,b)
 ? dist(C1, C2)

→ Agglomerative clustering



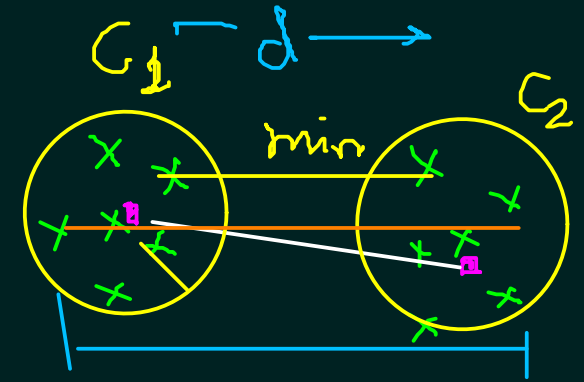
Top-Down
↓
Divisive Cluster

→ distance between clusters:

① Avg. distance: Centroids dist
 $C_i \leftrightarrow C_j$

② Min distance: dist between
closest of C_i & C_j

③ Max. distance: dist between
— farthest points of $C_i \leftrightarrow C_j$



↳ Cosine similarity
↳ euclidean dist

→ Manhattan dist

① → Avg. linkage clustering

② Single linkage clustering

③ Complete linkage clustering



→ Computational } Demerits
 → Noise in data }

Merit

↳ Can form
 [Arbitrary
 Shaped]



[Partitioned
 Clustering]

K-Means Clustering Algo:

d_1 d_2 ... d_n



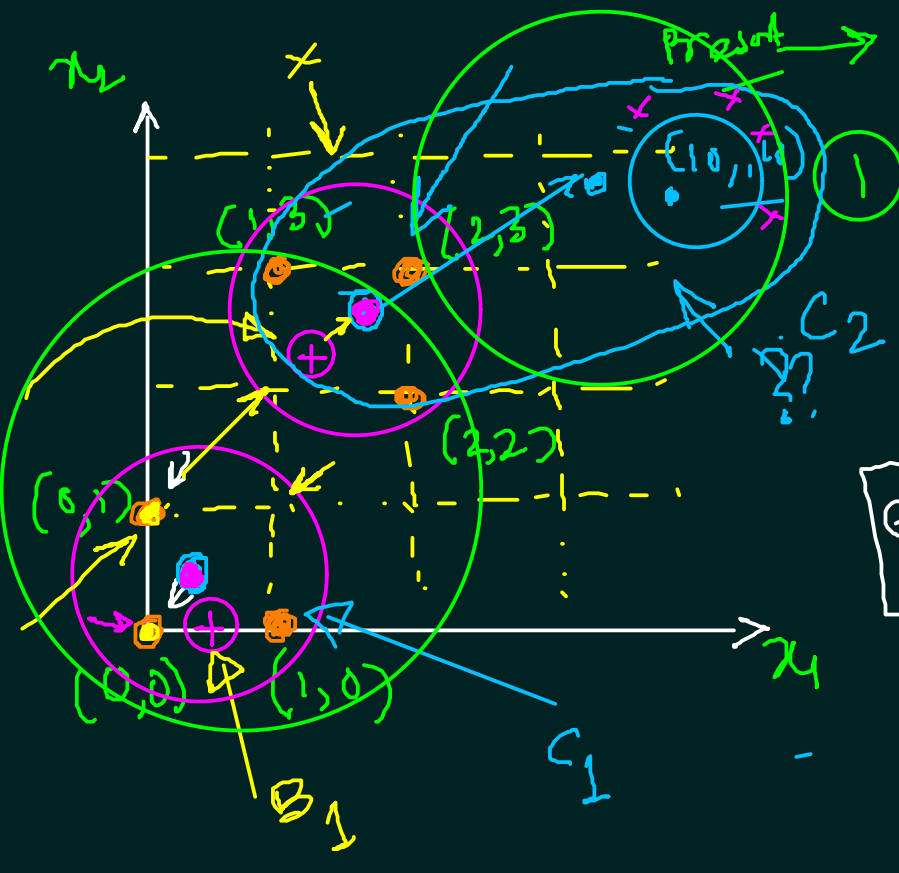
[until there is
 no change]

So on

m_1'

m_2'

m_3'



$$B_1 \rightarrow (0,0), (1,0) \quad \left. \begin{array}{l} \square \quad \square \quad \square \quad \square \\ \sqrt{O(kn)} \end{array} \right\} O(n)$$

$$B_2 \rightarrow (0,1), (1,3), (2,3), (2,2) \quad ?$$

[euclidean dist]

②

$$B_1 \rightarrow (0.5, 0) \quad (0,0), (1,0), (0,1)$$

$$B_2 \rightarrow (1.25, 2.25) \quad (1,3), (2,3), (2,2)$$

↓ EVALUATE ↓

③

$$B_1 \rightarrow (0.33, 0.33) \rightarrow [\text{-----}]$$

$$B_2 \rightarrow (1.66, 2.66) \rightarrow [\text{-----}]$$

→ $k = 3, 4$??

→ Initial Step?

↳ Arbitrary choice?

$$n-1 C_2 + n-2 C_2 + \dots + 2 C_2 = O(n^3)$$

→ Complexity? ✓