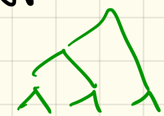
CLUSTERING

Hierarchical Agglomerative Clustering

1000s of types of clustering!

- HAC 
- Assignment-based clustering
- Spectral

"When data is easily clusterable, most clustering algorithms work quickly and well.

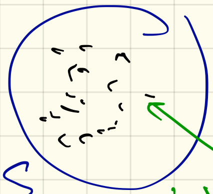
when data is not easily clusterable, then no algorithm can find good clusters."

clustering

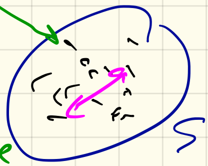
$S_1, S_2, S_3 \subset X$

$S_1 \cup S_2 \cup S_3 = X$

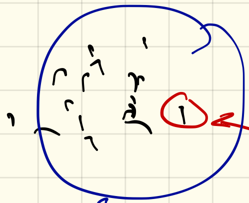
$S_i \cap S_j = \emptyset$



S_1



S_2



S_3

$x \in X$

distance

$d: X \times X \to \mathbb{R}$

intra dist

$\forall x_i \in S_i \quad x_j \in S_j$

$i \neq j$

$d(x_i, x_j)$ large

inter dist small

$\forall x_1, x_2 \in S_j$

$d(x_1, x_2)$ small

common data

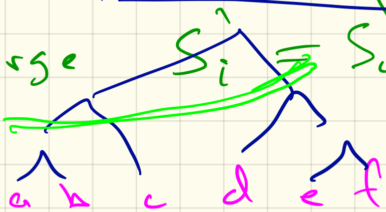
Hier Agg Clust

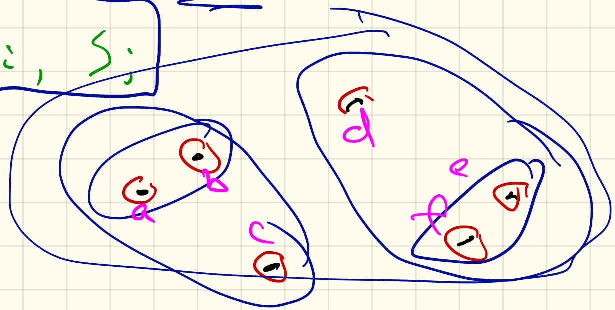
Input X, dist $\boxed{d.}$ $d: X \times X \to \mathbb{R}$

Algo: If two points (or clusters) are close enough, put them in the same cluster. Repeat.

0. Each $x_i \in X \to$ put in separate cluster S_i

1. while (two clusters are close enough)
    - 1a. Find $\boxed{\text{closest pair } S_i, S_j}$
    - 1b. Merge $S_i = S_i \cup S_j$





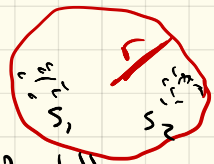
Distance between pair clusters S_1, S_2

- find "center" c_1 of S_1, c_2 of S_2

  $$D(S_1, S_2) = d(c_1, c_2)$$

  - c_i = average(S_i) — c_i = random(S_i)
  - $c_i = \underset{x \in S_i}{\arg\min} \, \|x\|$
  - c_i = median(S_i) $= \underset{c \in \cancel{X}\, S_i}{\arg\min} \sum_{x \in S_i} d(x, c)$

- $D(S_1, S_2)$ = radius (min enclosing ball$(S_1 \cup S_2)$)



- average, max

  $$D(S_1, S_2) = \underset{x_1 \in S_1,\, x_2 \in S_2}{\boxed{\min}} d(x_1, x_2)$$

  "single link"

- Build generative model
  compare likelihoods $L(S_1), L(S_2), L(S_1 \cup S_2)$

two moons

Which Variant?

- proper (gives) clustering: "smallest error"

  ↳ Dist D linked to error eval.

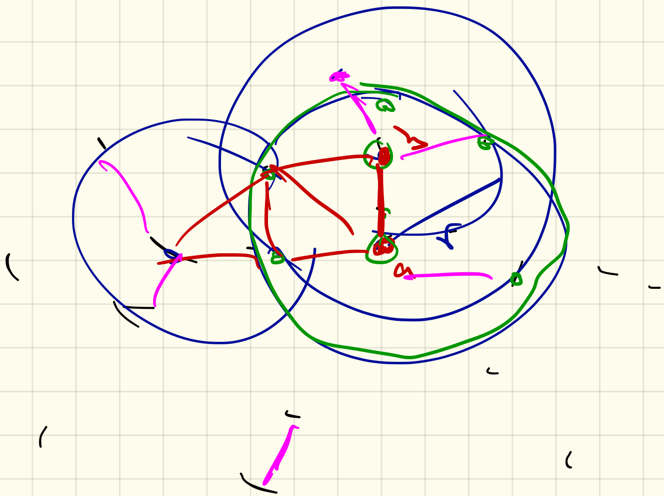
- <u>Computational Complexity.</u>
  - $O(n^2)$ pairwise dist.
  - $O(n)$ nodes in hierarchy.

  $O(n^3) \longrightarrow O(n^2 \log n)$ (w/ PQ)

  ? $O(n \log n)$ or $O(nk)$

DB Scan $\qquad \mathcal{O}(n \log n)$?

$X,\ d,\quad$ param: r (radius) $\qquad T$ (threshold density)



each $B_r(x)$

$|B_r(x) \cap X| = 8$

if $\geq T$

$x \to$ "core point"

link core points $d(x, x') < r$

$\hookrightarrow$ cluster.

K-Center Cluster $\qquad O(nk)$

Gonzalez Algo.

In: X, d, k $\qquad$ any d metric

Out: $c_1, c_2, \dots, c_k \leftarrow$ centers of clusters,

$c_1 \leftarrow$ arbitrarily(X)

for $j = 2$ to k

$$c_j = \underset{x \in X}{\arg\max} \left(\cancel{\arg} \min_{c \in \{c_1, \dots, c_{j-1}\}} \| x - c \| \right)$$