Hierarchical Agglomerative ~~Clustering~~

Density-based Clustering

Feb 17, 2025

Jeff M. Phillips

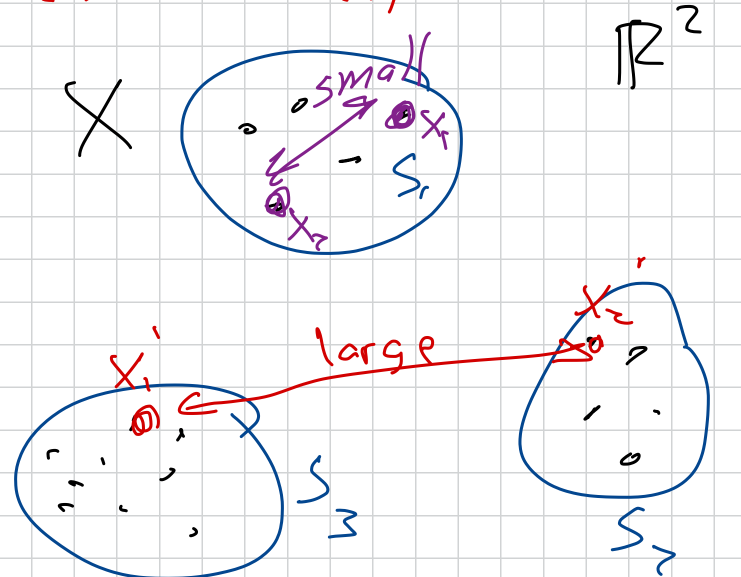
What is Clustering

Input

Data set $X = \{x_1, x_2, \dots x_n\} \in \mathcal{X}$

Distances $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$

often $\mathcal{X} \subset \mathbb{R}^d$ $\quad D(x_1, x_2) = \|x_i - x_2\|$

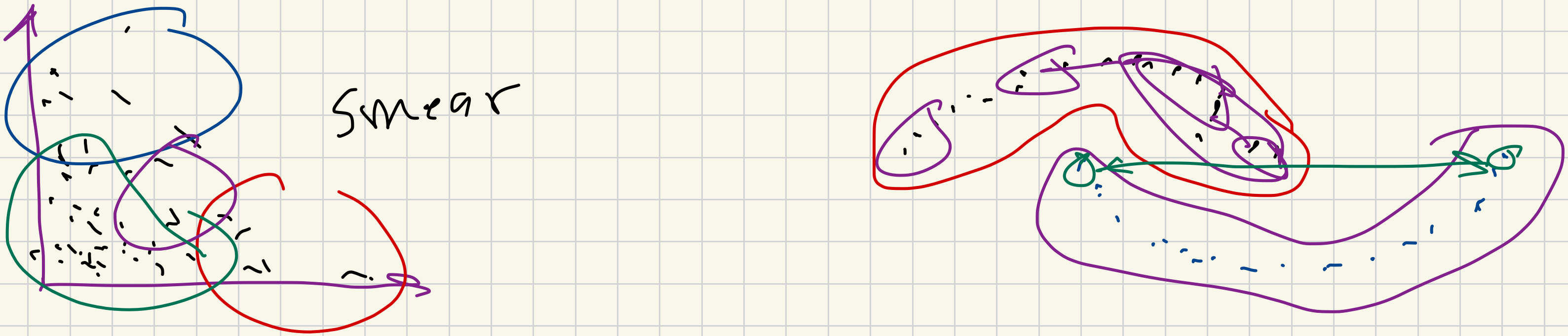


Goal Partition X into

k sets $S_1, S_2, \dots S_k$

$S_i \cap S_j = \emptyset$, $\bigcup_{j=1}^{k} S_j = X$

- points in cluster $x_1, x_2 \in S_j$ close $D(x_1, x_2)$ small
- points in diff clusters $x_1 \in S_1$, $x_2 \in S_2$ $D(x_1, x_2)$ large



- When data is easily or naturally clusterable, then most clustering algorithms work quickly and well

- When data is **not** easily or naturally clusterable, then no algorithm will find good clusters

Hierarchical Agglomerative Clustering (HAC)

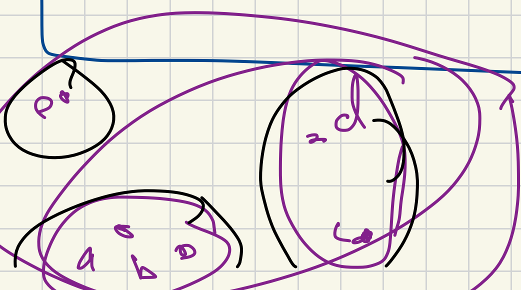
1. Each $x \in X$ is its own cluster

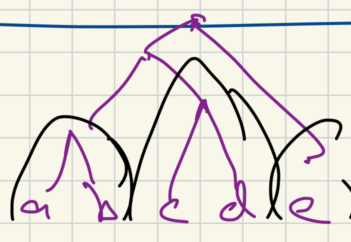
   $S_1 = \{x_1\}, \quad S_2 = \{x_2\}, \ldots$

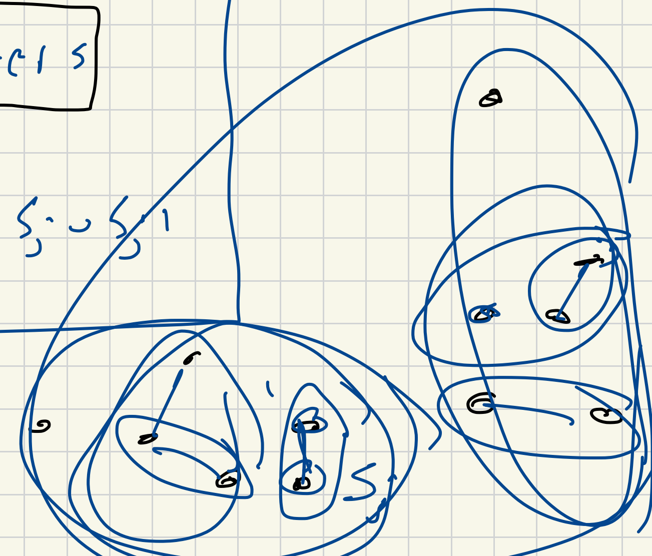
2. while (more than 1 cluster)

   a. Find two closest clusters $S_j, S_{j'}$.

   b. Merge $S_j, S_{j'}$ into $S_j = S_j \cup S_{j'}$







Find two closest clusters

distance $D(S_1, S_2)$ $\quad D: \mathcal{X} \times \mathcal{X}$

- Distribution Dist $W_{s,r}$, D_K

- $D(c_1, c_2)$

  c_j "center" S_j

  ↳ mean(S_j)

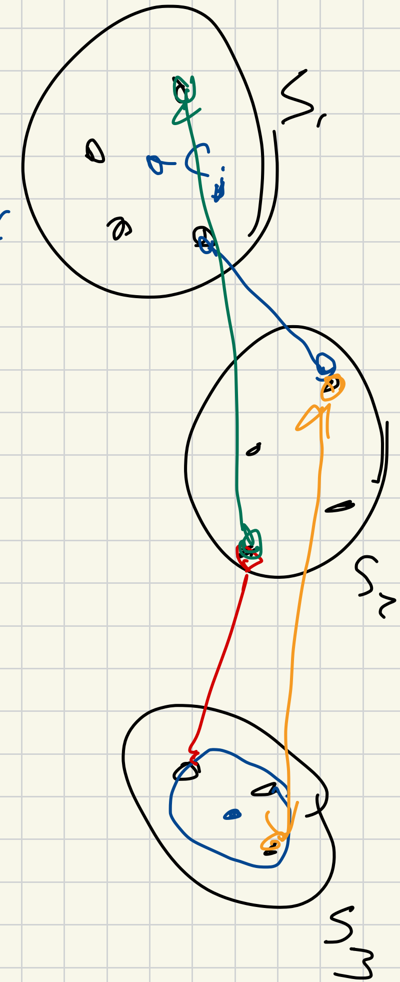
  ↳ center of MEB

Single-Link

- $D(S_1, S_2) = \min_{x_1 \in S_1, x_2 \in S_2} D(x_1, x_2)$

Complete-Link

- $D(S_1, S_2) = \max_{x_1 \in S_1, x_2 \in S_2} D(x_1, x_2)$



Runtime of HAC Single-Link

- How many merges? w $|X| = n$
  - ↳ $n-1 = O(n)$ merges.

Find closest 2 clusters.

Merge $S_j, S_{j'} \Rightarrow S_j''$ update the distance

Update $O(n)$ distances $S_i, S_{i+1} \dots$

maintain in PQ $O(\log n)$ time

each $O(1)$ time

↳ $O(n^2 \log n)$ time. slow

Density-based Clustering

DBScan ≈ chopped off Single-Link HAC

- radius r = clusters to be separated if $D_{SL}(s_1, s_2) > r$.
- threshold T = min # pts in ball radius r

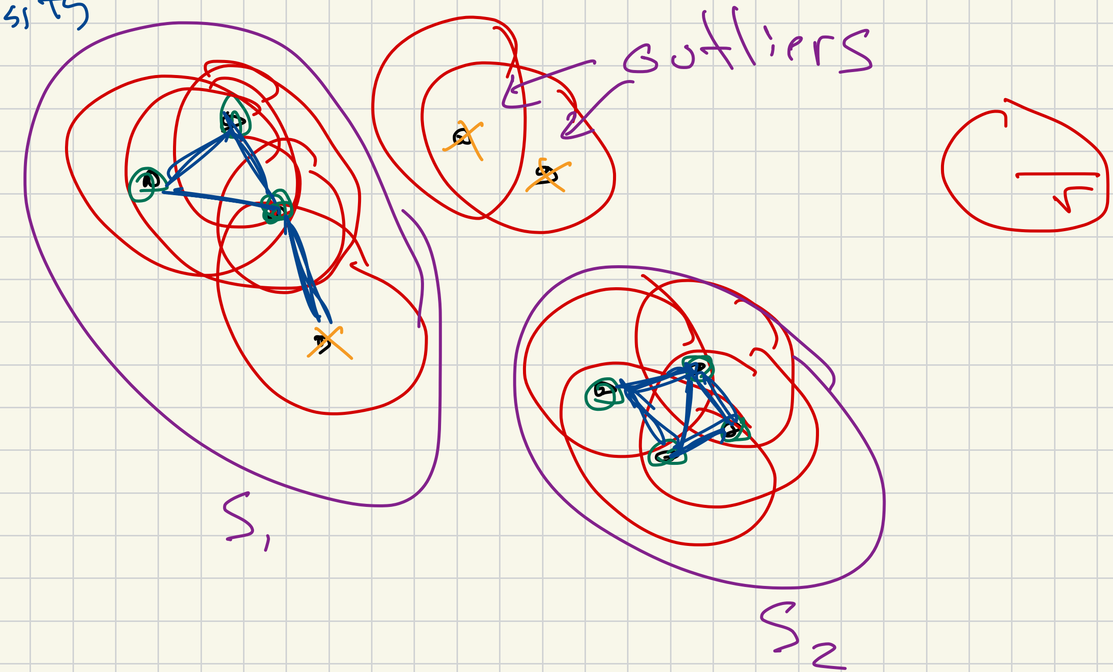
$T = 10$

1. For all $x \in X$, find # pts in ball $B_r(x)$ $O(n \log n)$
   if $> T \Rightarrow x$ is core
2. Build graph on core pts to all neighbors (radius r) $O(n^2)$
3. Clusters are connected components of Graph $O(n\alpha(n))$

threshold density

$T = 3$

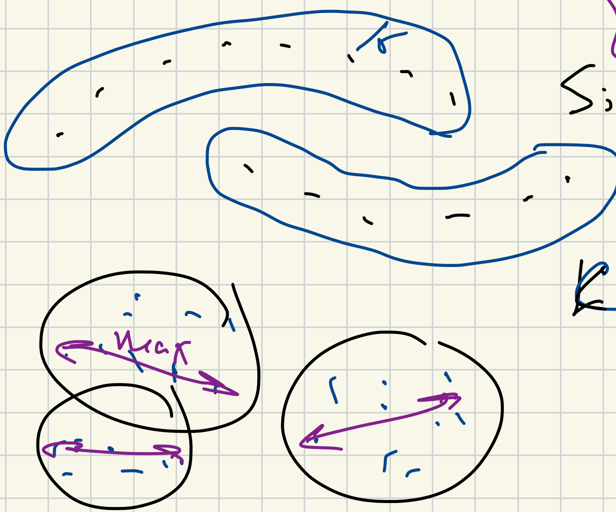
Edges



DBScan

How many Clusters?

- DBScan : depended on r, T (r: radius, T: threshold)
- HAC : Single, complete Link
  threshold



Single = min distance between points in diff clusters

Complete Link = max dist between pts in cluster

Plug in values, look at clusters!

- look at sample data

sanity check