Clustering

Note Title 2/8/2016

Ill-defined

Hard clustering

Given set $X \subset M$, metric $d: M \times M \to \mathbb{R}^+$

cluster $S_i \subset X$

partition clustering $\mathcal{C}(X) = \{S_1, S_2, \ldots, S_k\}$

(1) each $S_i \subset X$

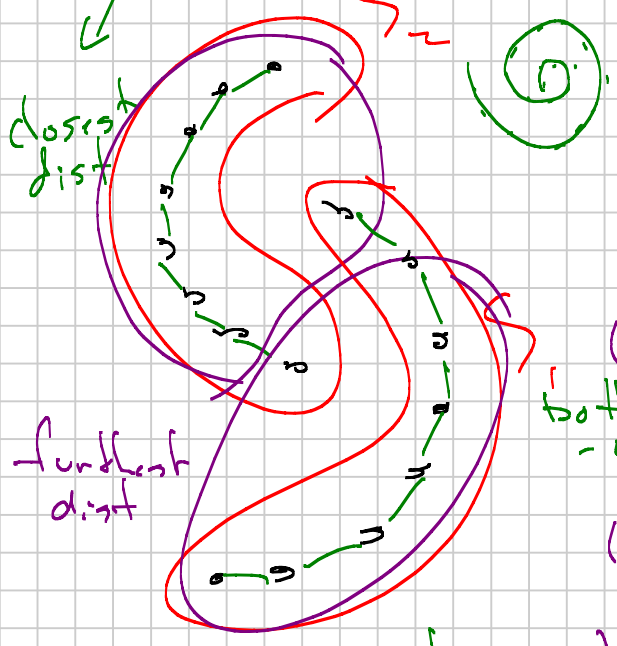
(2) each pair $S_i \cap S_j = \emptyset$ (often "soft")

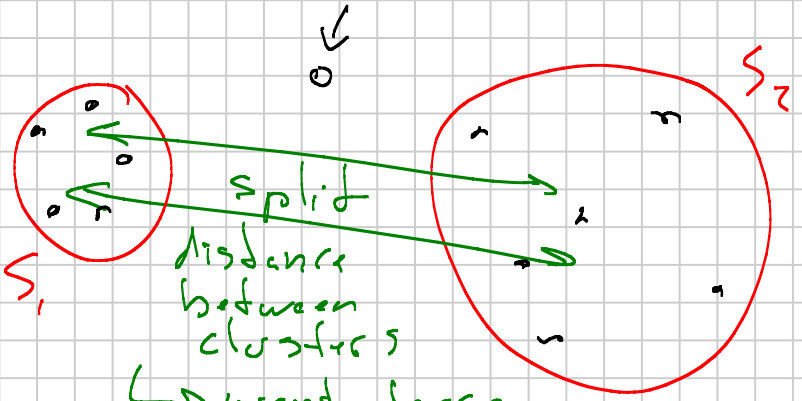
(3) $\bigcup_{i=1}^{k} S_i = X$.

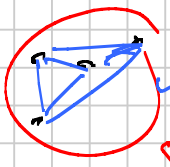
Objective !

manifold learning

"swiss roll"







width = distances in cluster

↳ want small

(1) Hierarchical

bottom-up Agglomerative clustering

top-down

(2) Assignment-based clustering (k-means)

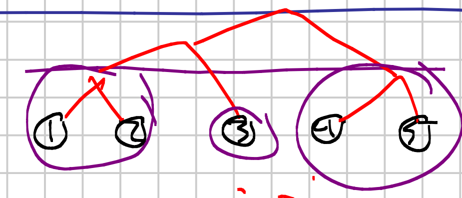
(3) Spectral (Graph

HAC

1. Each $x_i \in X$ is a separate cluster S_i
2. while Two clusters are close enough — $O(n)$
3. Find closest two clusters S_i, S_j — tie! — $O(n^2)$, $O(n \log n)$, $O(n^2 \log n)$
4. Merge S_i, S_j into → single cluster S_i

$+\ O(n^3)$

"Close" between clusters



- Create center of each cluster, dist between centers
  - + average (Euclidean mean)
  - + mode
  - + pick $c \in S_i$ minimize $\sum_{s \in S_i} d(c, s_i)^2$
  - + arbitrary point

- Distance between closest points

$$d(S_1, S_2) = \min_{s_1 \in S_1,\ s_2 \in S_2} d(s_1, s_2)$$

- Distance between farthest points

$$d(S_1, S_2) = \max_{s_1 \in S_1,\ s_2 \in S_2} d(s_1, s_2)$$

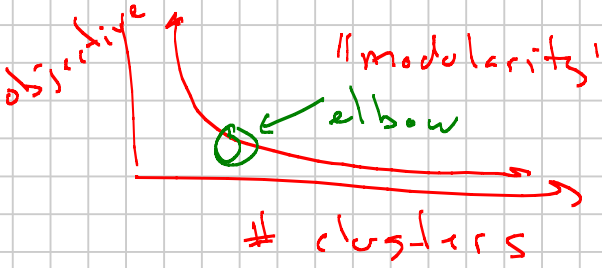
- Density of $S_1 \cup S_2$

width
- + radius of minimum enclosing ball
- + average distance (to center / among pts) after joining

What is close "enough"

- Reach k clusters
- radius of all clusters reach maximum
  ↳ merging would get "too big" (width)
- Split get "too small" clusters get too close
- density

Hidden parameters



Efficiency Runtime as function of $n = |X|$

- How many merges? $O(n)$ merges