Clustering

Sunday, September 19, 2021 5:51 PM

Goal:

Clustering refers to partitioning a set into "dissimilar" subsets of "similar" elements.

Usually a well-defined objective.

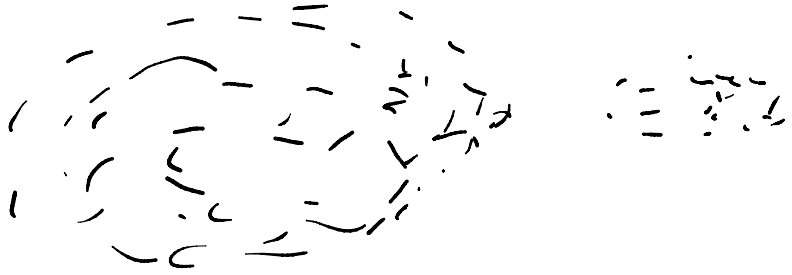
E.g. k-means, k-median, k-center, diameter.

- define

of with the goal of recovering some ground truth.

- all are NP-hard

No universal clustering criterion.



Depends on the context/application.

K-center: – start with any point in given set as first center. $C = \{c_1\}$

Repeat K-1 times [– add farthest point to C

Thm. Greedy algorithm is a factor 2 approximation.

Pf. Suppose OPT is r.

Claim: For the centers $c_1 \ldots c_K$ found by GREEDY, max distance to nearest center $\le 2r$.

If not,

$\Rightarrow \exists\ K+1$ pts $c_1 \ldots c_{K+1}$

s.t. $d(c_i, c_j) > 2r$.

$\Rightarrow$ No two of $\{c_1 \ldots c_{K+1}\}$ can belong

$\Rightarrow$ No loss of $\{c_1 \dots c_{k+1}\}$ to save cluster of radius r.

The Spectral Approach

- Project to span of top k singular vectors of $A \in \mathbb{R}^{n \times d}$
- Cluster in $\mathbb{R}^k$.

Idea: this should shrink distance between x and nearest center.

$$n\left(\begin{array}{c} \\ A \\ \\ \end{array}\right) - \left(\begin{array}{c} c_1 \\ c_2 \\ C \end{array}\right)$$

d

centers.

$\sigma^2(C) = \frac{\|A - C\|_2}{n}$: average variance of clusters.

clusters.

Thm If $\|C_i - C_j\| > 15\frac{k}{\varepsilon}\cdot\sigma(C)$ $\forall i \neq j$ and each cluster has $\leq \varepsilon n$, then Spectral clustering finds C' that differs from C in at most $\varepsilon^2 \cdot n$ points.

Algo. 1. Project to top k right Singular vectors of A.

Repeat k times. { 2. Take a random row, include all points within distance $6\frac{k}{\varepsilon}\sigma(C)$.

$$A_k = \underset{D:\, rk(D)\leq k}{\operatorname{argmin}} \|D - A\|_2$$

Lemma. for any C of rank k, $\|A_k - C\|_F^2 \leq 8k\|A - C\|_2^2$ any A.

Pf: (*) $\|A_k - C\|_F^2 \leq 2k\|A_k - C\|_2^2$

Since $A_k - C$ has rank $\leq 2k$.

Since $A_k - C$ has rank $\leq 2k$.

$$\|A_k - C\|_2 \leq \|A_k - A\|_2 + \|A - C\|_2$$

$$\leq 2\|A - C\|_2 \qquad (**)$$

$$(*) + (**) \Rightarrow \quad \|A_k - C\|_F^2 \leq 8k\,\|A - C\|_2^2 .$$

Pf (of Thm). Let v_i be i^{th} row of A_k.

Claim. Most v_i are within distance $\frac{3k}{\varepsilon}\sigma(C)$ of their center.

Let $B = \{i : \|v_i - c_i\| > \frac{3k}{\varepsilon}\sigma(C)\}$.

Then $\|A_k - C\|_F^2 \geq |B| \cdot \frac{9k^2}{\varepsilon^2} \cdot \sigma(C)^2$

$$\left(\leq 8k\,\sigma(C)^2 \cdot n \right.$$

$$\Rightarrow \quad |B| < \frac{\varepsilon^2}{k} \cdot n .$$

For $i, j \in$ same cluster and $\notin B$,

$\|v_i - v_j\| < (k \quad \sigma(C)$ ✓

$$\|v_i - v_j\| \le \frac{6K}{\varepsilon}\sigma(C).$$

i, j different cluster, $\notin B$

$$\|v_i - v_j\| > \frac{15K}{\varepsilon}\sigma(C) - \frac{6K}{\varepsilon}\sigma(C) = \frac{9K}{\varepsilon}\sigma(C).$$

Hence if we pick point not in B as the seed, all K times, all points not in B will be correctly classified.

$$P(\text{we pick point} \notin B) \ge \left(1 - \frac{\varepsilon^2}{K}\right)\cdot\left(1 - \frac{(\varepsilon-\varepsilon^2)}{K}\right)^{K-1}$$

$$\ge 1 - \frac{\varepsilon}{K}\cdot K = 1-\varepsilon.$$

Example 1 Mixture of K Gaussians, each with max covariance σ^2.

Then $\sigma(C) \le \sigma\sqrt{K}$ (distance to own center)

Separation needed: $\frac{15K}{\varepsilon}\sigma(C) \le 15\frac{K^{3/2}}{\varepsilon}\cdot\sigma$.
between centers

Example 2 Stochastic Block Models.

Example 2 Stochastic Block Models.
or Planted Partitions.