Clustering

Sunday, September 19, 2021 5:51 PM

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

Usually a well-defined objective.

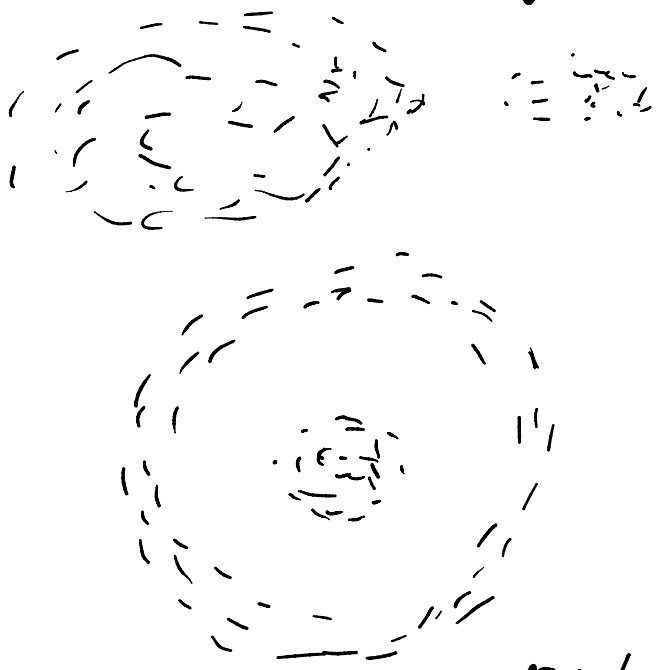
E.g. k-means, k-median, k-center, diameter.

— 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 $c_1 \dots c_k$ found by GREEDY, max distance to nearest center $\leq 2r$.

Proof,

$\Rightarrow \exists\ k+1$ pts $c_1 \dots c_{k+1}$

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

$\Rightarrow$ No two of $\{c_1 \dots c_{k+1}\}$ can belong to same cluster of radius r.

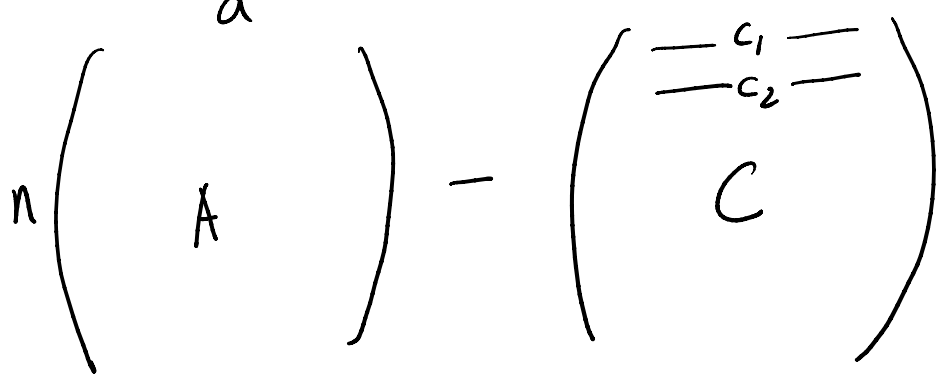
The Spectral Approach

– Project to span of top k singular vectors of $A \in R^{n \times d}$

– Cluster in R^k.

between

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



centers.

$\sigma^2(C) = \frac{\|A - C\|_2}{n}$: average variance of 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 $6\frac{k}{\varepsilon}\sigma(C)$.

$\|D - A\|$.

$$A_k = \operatorname*{argmin}_{D:\, rk(D) \le k} \|D - A\|_2$$

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

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

since $A_k - C$ has rank $\le 2k$.

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

(**) $$\le 2\|A - C\|_2$$

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

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

$\le 8k\sigma(C)^2 \cdot n$

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

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

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

i, j different clusters, $\notin B$

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

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

$$P_1(\text{we pick point} \notin B) \ge \left(1 - \frac{\varepsilon^2}{k}\right) \cdot \left(1 - \frac{(\varepsilon - \varepsilon^2)}{k}\right)^{k-1}$$

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

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

Then $\sigma(C) \le c_1 \sigma$ (distance to own center in one direction)

Separation needed: $\frac{15k}{\varepsilon}\sigma(C) = O\left(\frac{k}{\varepsilon} \cdot \sigma\right)$.

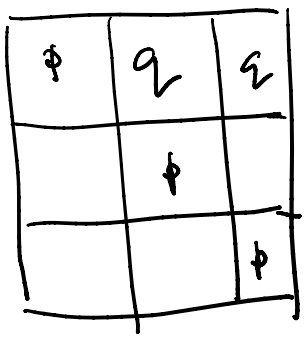
between centers

(a bit worse than what we got)

Example 2 Stochastic Block Models.

or Planted Partitions.

or Planted Partitions.



$i,j \in$ same block $\hookleftarrow$

$$\mathbb{P}((i,j) \in E) = \begin{cases} p \\ q & \text{different blocks.} \end{cases}$$

A: adjacency matrix of G.

$$\mathbb{E}(A) = C = \begin{pmatrix} pp\cdots p & q\cdots q \\ \vdots & \\ \vdots & \end{pmatrix}$$

Problem: Recover "planted" partition.

Apply spectral algorithm.

$$\|\mu_i - \mu_j\|^2 = \quad \cdot (p-q)^2 \cdot n$$

different clusters.

What about $\|A - C\|_2$?

$A - C$ is a random matrix with $\mathbb{E}(A-C)=0$ and independent entries.

Thm. R random with independent entries $\in [-1,1]$

$\mathbb{E}(R_{ij})=0$ $\mathbb{E}(R_{ij}^2) \le \sigma^2$.

$\|R\| \le (2+o(1))\,\sigma\sqrt{n}$.

$\mathbb{E}(R_{ij}) = 0 \quad \mathbb{E}(R_{ij}^2) \le \sigma^2$.

Then with prob. $1 - o(1)$, $\|R\|_2 \le (2 + o(1))\sigma\sqrt{n}$.

So, $\sigma(C)^2 = \dfrac{\|A - C\|_2^2}{n} = O(p)$

So by the spectral clustering theorem, it suffices to have

$$\|r_i - r_j\|^2 \gtrsim (p-q)^2 \cdot n > c\frac{k^2}{\varepsilon^2} \cdot p > \left(\frac{15k}{\varepsilon}\right)^2 \cdot \sigma(C)^2$$

or $|p - q| > c' \dfrac{k}{\varepsilon}\sqrt{\dfrac{p}{n}}$

if we set $p = \dfrac{a}{n}$, $q = \dfrac{b}{n}$

then this says. $|a - b| = \Omega(k) \cdot \sqrt{a}$ suffices.

This is information theoretically tight up to constant factors.

In fact for $k = 2$, $(a-b)^2 > 2a$ is necessary and sufficient!

... of random matrix bound?

Q. How to prove the random matrix bound?