DS-563 / CS-543 Lecture 15 / 2023-03-21

Today:

- coresets for k-median
- new theme: sampling from probability distributions
  - number of samples needed to learn a discrete distribution

---

Popular clustering objectives:

Goal: given a set S of points, select up to k points $Q = \{q_1, \dots, q_k\}$ to minimize

k-means: $\sum_{p \in S} \min_{q \in Q} \left(\text{dist}(p,q)\right)^2$

k-median: $\sum_{p \in S} \min_{q \in Q} \text{dist}(p,q)$

k-center: $\max_{p \in S} \min_{q \in Q} \text{dist}(p,q)$

Main difference: different sensitivity to outliers

Today: k-median

Notation: $opt_k(X)$ = cost of optimal k-median clustering for X

Assumption: we have an algorithm that computes α-approximation to k-median

$\quad$ || solution of cost $\leq \alpha \cdot opt_k$ (set of points)

$\alpha = O(1)$ possible in polynomial time

We write $Alg(X)$ to denote the output of this algorithm on set X

Our coreset construction for set S:

- $Q \leftarrow Alg(S)$
- for each $p \in S$: $q_p \leftarrow q \in Q$ closest to p
- return multiset composed of $|\{s \in S : q_s = q\}|$ copies of each $q \in Q$

Idea: round each point to the closest center

Denote the output of our coreset construction "coreset(S)"

Main claim:

$$S = S_1 \cup S_2 \cup \dots \cup S_t$$

$\mathrm{Alg}\left(\bigcup_{i=1}^{t} \mathrm{coreset}(S_i)\right)$ is a solution of cost $\alpha(\alpha+2)\,\mathrm{opt}_k(S)$

Proof:

Claim 1: $\underbrace{\sum_{i=1}^{t} \mathrm{opt}_k(S_i)}_{} \leq \mathrm{opt}_k(S)$

allows for assigning points to more than k centers, reusing centers corresponding to $\mathrm{opt}_k(S)$ allows for equality but may do better

Notation: For each $p \in S$, let q_p be the point in $\bigcup_{i=1}^{t} \text{coreset}(S_i)$ representing it

Claim 2: $\underbrace{\sum_{p \in S} \text{dist}(p, q_p)}_{\text{total dislocation of points in coresets}} \leq \alpha \, \text{opt}_k(S)$

$$\sum_{p \in S} \text{dist}(p, q_p) = \sum_{i=1}^{t} \sum_{p \in S_i} \text{dist}(p, q_p)$$

$$= \sum_{i=1}^{t} \begin{pmatrix} \text{cost of clustering} \\ \text{in } Alg(S_i) \end{pmatrix} \leq \sum_{i=1}^{t} \alpha \, \text{opt}_k(S_i)$$

$$\leq \alpha \, \text{opt}_k(S)$$

via Claim 1

Claim 3: $\text{opt}_k\left(\bigcup_{i=1}^{t} \text{coreset}(S_i)\right) \leq (\alpha+1) \, \text{opt}_k(S)$

- Q^* = optimal solution for S
- For each $p \in S$, let $q_p^* \in Q$ be the closest center for p
- $\begin{pmatrix} \text{cost of clustering } \bigcup_{i=1}^{t} \text{coreset}(S_i), \\ \text{using } Q^* \end{pmatrix}$

$$\leq \sum_{p \in S} \text{dist}(q_p, q_p^*)$$

$$\leq \underbrace{\sum_{p \in S} dist(q_p, p)}_{\substack{\leq \alpha\, opt_k(S) \\ \text{via Claim 2}}} + \underbrace{\sum_{p \in S} dist(p, q^*_p)}_{= opt_k(S)}$$

$$\leq (\alpha + 1)\, opt_k(S)$$

Claim 4: $\left(\text{cost of clustering } S \text{ with points in } Alg\left(\bigcup_{i=1}^{t} coreset(S_i)\right) \right)$ $\circledast =$

$$\leq \alpha(\alpha+2)\, opt_k(S)$$

[This will finish the proof of the main claim]

Notation: For each $p \in S$, let $\hat{q}_p$ be the closest center in $Alg\left(\bigcup_{i=1}^{t} coreset(S_i)\right)$ to q_p

$$\circledast \leq \sum_{p \in S} dist(p, \hat{q}_p)$$

$$\leq \sum_{p \in S} dist(p, q_p) + \sum_{p \in S} dist(q_p, \hat{q}_p)$$

$$\leq \alpha \, opt_k(S) + [\text{cost of clustering Alg}(\bigcup_{i=1}^{t} \text{coreset}(S_i))]$$

$$\leq \alpha \, opt_k(S) + \alpha \, opt_k\left(\bigcup_{i=1}^{t} \text{coreset}(S_i)\right)$$

$$\leq \alpha \, opt_k(S) + \alpha(\alpha+1) \, opt_k(S)$$

$$\leq \alpha(\alpha+2) \, opt_k(S) \quad \blacksquare$$

New Theme: Discrete Distributions

Access to unknown distribution(s) on a discrete set via:

- independent samples
- and/or probability mass function queries
- and/or ...

Goals:

- Learn the distribution
- check if the distribution has some specific property (is it uniform?)
- estimate a parameter of the distribution (what is its entropy? what is the support size?)

Total variation distance:

D_1 & D_2 ~ two distributions on $[n]$

p_i = probability a sample drawn from D_1 is i

q_i = probability a sample drawn from D_2 is i

For any subset $S \subseteq [n]$, define

$$p(S) = \sum_{x \in S} p_x$$

$$q(S) = \sum_{x \in S} q_x$$

Total variation distance between D_1 & D_2 is

$$d_{TV}(D_1, D_2) = \max_{S \subseteq [n]} |p(S) - q(S)|$$

$$= \frac{1}{2} \| (p_1, \ldots, p_n) - (q_1, \ldots, q_n) \|_1$$

Homework 2

Now: learning arbitrary discrete distribution

Model: distribution D on $[n]$

$\boxed{\text{Algorithm}} \leftarrow$ independent samples from D

$\downarrow$

output D'

Goal: with probability $9/10$,

$$d_{TV}(D, D') \leq \varepsilon \quad \leftarrow \text{input parameter}$$

How many samples are needed?

"Easy" $O\left(\frac{1}{\varepsilon^3} n \log n\right)$ bound:

- elements of probability $< \frac{\varepsilon}{100 n}$ are negligible (their total mass $\leq \frac{\varepsilon}{100}$)
- can estimate probabilities of heavier elements up to multiplicative factor of $(1 \pm \varepsilon/100)$

  via Chernoff + Union bound

A better $O(n/\varepsilon^2)$ bound

Algorithm:

sufficiently large constant

- collect $t = C \cdot n/\varepsilon^2$ independent samples
- output distribution D' s.t.

  probability of $i \in [n]$ is $\frac{\#\text{ samples equal to } i}{t}$

This is known as the empirical distribution of the samples

Why this works:

Notation: For any $S \subseteq [n]$,

$p(S) =$ total probability of elements of S in D

$q(S) =$ total probability of elements of S in D'

Need to show that with probability $9/10$, for all $S \subseteq [n]$, $|p(S) - q(S)| \leq \varepsilon$

Reminder: Hoeffding's Inequality (simplified)

$X_1, \ldots, X_t \in [0,1]$ independent random variables

$$\Pr\left(\left|\sum_{i=1}^{t} (X_i - \mathbb{E}[X_i])\right| \geq \Delta\right) \leq 2e^{-\frac{2\Delta^2}{t}}$$

Consider a specific $S \subseteq [n]$.

Define $X_i = \begin{cases} 1 & \text{if } i\text{-th sample in } S \\ 0 & \text{otherwise} \end{cases}$ for $i \in [t]$

For each $i \in [t]$, $\mathbb{E}[X_i] = p(S)$

And $\left(\sum_{i=1}^{t} X_i\right)/t = q(S)$

From Hoeffding's Inequality with $\Delta = \varepsilon t$:

$$\Pr\left(\left|\sum_{i=1}^{t}(X_i - \mathbb{E}[X_i])\right| \geq \varepsilon t\right) \leq 2e^{-2\varepsilon^2 t}$$

$$\Pr\left(\left|\sum_{i=1}^{t} X_i - t \cdot p(S)\right| \geq \varepsilon t\right) \leq 2e^{-2\varepsilon^2 t}$$

$$\Pr\left(\left|q(S) - p(S)\right| \geq \varepsilon\right) \leq 2e^{-2Cn}$$

By the union bound, this holds for all $S \subseteq [n]$, with probability $1 - 2^n \cdot 2e^{-2Cn} \geq 1 - 2^{n+1} \cdot 2^{-2Cn}$

$$\geq 1 - 2^{1-9n} \geq 1 - 2^{-8n} \geq \frac{9}{10}$$

$C \geq 5$

$\blacksquare$

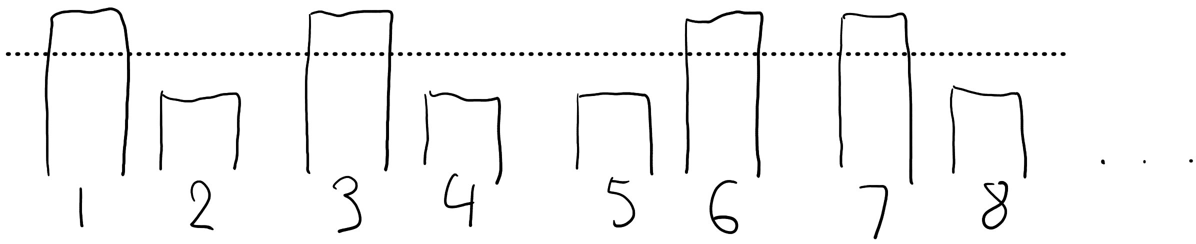
Note: $\Omega(n/\varepsilon^2)$ needed

Intuition:

$\Omega(1/\varepsilon^2)$ coin tosses needed to distinguish

coins	heads	$\frac{1}{2}-\varepsilon$	vs.	$\frac{1}{2}+\varepsilon$
	tails	$\frac{1}{2}+\varepsilon$		$\frac{1}{2}-\varepsilon$

Hard to learn distribution



Each pair $(2i-1, 2i)$ simulates a coin biased in a random direction

Probabilities either $\frac{1}{n}(1-c\cdot\varepsilon), \frac{1}{n}(1+c\cdot\varepsilon)$ or the other way around

some constant

The algorithm "needs to learn" the direction of the bias for most coins