Lecture 8

TTIC 41000: Algorithms for Massive Data

Toyota Technological Institute at Chicago

Spring 2021

Instructor: Sepideh Mahabadi

This Lecture

- Core-sets for k-median

Coreset for 1-means

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

- A coreset from which we can estimate the 1-means cost, e.g., $Cost(P, q) = \sum_{p \in P} dist(p, q)^2$

- $\sum_{p \in P} ||p - q||^2 = \sum_{p \in P} \langle p - q, p - q \rangle = \sum_{p \in P} ||p||^2 + n||q||^2 - 2q \sum_{p \in P} p$
- So only keep the mean $\sum_{p \in P} p$
- Not a core-set exactly.

Coreset for k-center

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

- We showed a coreset from which we can estimate the 1-center cost, e.g., $Cost(P, q) = far(P, q) = \max_{p \in P} dist(p, q)$
- What about k-center cost?

Naïve Uniform Sampling

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

Importance Sampling

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

- $Pr \approx 1/n_i$ proportional to size of the cluster
- $weight \approx n_i$ proportional to the size

Importance Sampling

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

- $Pr \approx 1/n_i$ proportional to size of the cluster
- $weight \approx n_i$ proportional to the size
- Do we need the clusters?

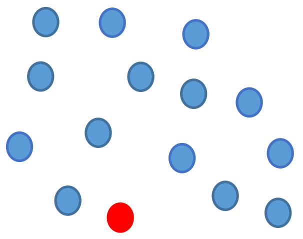
Importance Sampling

Given: a point set $P \subset \mathbb{R}^d$

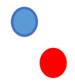
Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P,q) \approx Cost(C,q)$$

- $Pr \approx 1/n_i$ proportional to size of the cluster
- $weight \approx n_i$ proportional to the size
- Do we need the clusters?
  - Answer: some approximation suffices



Low Probability
Large weight



High Probability
Low weight

Importance Sampling

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P,q) \approx Cost(C,q)$$

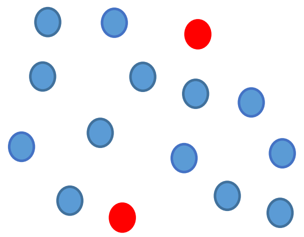
- $Pr \approx 1/n_i$ proportional to size of the cluster
- $weight \approx n_i$ proportional to the size
- Do we need the clusters?
  - Answer: some approximation suffices
  - Even bi-criteria approximation (pick more centers)

Low Probability
Large weight

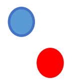
High Probability
Low weight

General approach

1. Find an approximate clustering (usually much easier)
2. Sample points based on their cluster size



Low Probability
Large weight



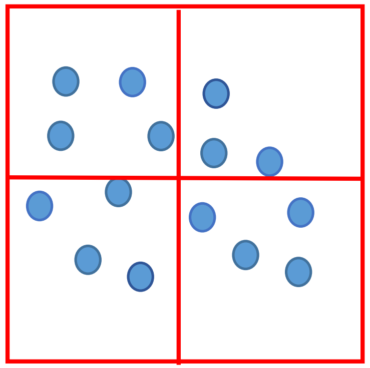
High Probability
Low weight

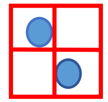
Even use previous approach

k-center

1. Find an approximate clustering (usually much easier)
2. Apply the grid on each cluster

- Exponential dependence on d





Coreset for 1-means

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

- A coreset from which we can estimate the 1-means cost, e.g., $\boldsymbol{Cost(P, q) = \sum_{p \in P} dist(p, q)^2}$

Coreset for k-median

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $Q \subseteq \left(\mathbb{R}^d\right)^k$

$$Cost(P, Q) \approx Cost(C, Q)$$

- A coreset from which we can estimate the k-median cost, e.g.,

$$Cost(P, Q) = \sum_{p \in P} dist(p, Q) = \sum_{p \in P} \min_{q \in Q} dist(p, q)$$

- Opening facilities, test several candidates.

Coreset for k-median

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $Q \subseteq \left(\mathbb{R}^d\right)^k$

- $\sum_{p \in P} dist(p, Q) \approx_{1+\epsilon} \sum_{c \in C} w(c) dist(c, Q)$

- A coreset from which we can estimate the k-median cost, e.g.,

$$Cost(P, Q) = \sum_{p \in P} dist(p, Q) = \sum_{p \in P} \min_{q \in Q} dist(p, q)$$

A general theorem

- Suppose the cost function satisfies
  - $Cost(P, Q) = \sum_{p \in P} w(p) dist(p, Q)$
- Sample C proportional to sensitivity$(p) = \max_{Q \in \mathcal{Q}} \frac{dist(p,Q)}{\sum_{p' \in P} dist(p',Q)}$
- Number of samples: $|C| \geq O\left(\frac{VC(\mathcal{Q})}{\epsilon^2} \cdot \sum_p \text{sensitivity}(p)\right)$
- Need to bound
  - $VC(\mathcal{Q})$: (roughly how many parameters one need to describe the query, e.g., kd)
  - Total sensitivity $\sum_p \text{sensitivity}(p)$ (for k-median can be bounded by k)
  - Gives coreset of size $O(\frac{k^2 d}{\epsilon^2})$

Applications

- k-means, k-median, k-center
- j-subspace: query q is a j-dimensional subspace.
- Projective clustering (j,k): query Q is a set of k j-dimensional subspaces.

K-median

- $\text{sensitivity}(p) = \frac{dist(p,Q^*)}{\sum_{p' \in P} dist(p',Q^*)} + \frac{1}{n_p}$
  - Q^* is the optimal k-means clustering (again we can use approximation)
  - n_p is the number of points in p's cluster
- Total sensitivity $= 1 + k$
- $|C| = \left(\frac{k^2 d}{\epsilon^2}\right)$
- Combining with PCA gives $\left(\frac{k^2 (\frac{k}{\epsilon})}{\epsilon^2}\right)$
  - Independent of n
  - Independent of d

Proof of the Theorem

Setup

Given: (P, w), $P \subseteq X$, $w: P \to [0,1)$, sum of weights are 1

- Core-sets for core-sets

Query space: $(P, w, \mathcal{Q}, f)$

- $f: P \times \mathcal{Q} \to [0, \infty)$
- $\bar{f}(P, w, Q) = \sum_{p \in P} w(p) \cdot f(p, q)$

Change multiplicative ϵ to additive error ϵ

- Goal: find $(C, u, \mathcal{Q}, f)$ such that for any $Q \in \mathcal{Q}$:
- $\left|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)\right| \le \epsilon$
- Why?
- Let $f(p, Q) \coloneqq \frac{dist(p,Q)}{Cost(P,Q)} = \frac{dist(p,Q)}{\sum_{p \in P} w(p) \cdot dist(p,Q)}$

Why

We want $|Cost(P,Q) - Cost(C,q)| \leq \epsilon \cdot Cost(P,Q)$

- $|Cost(P,Q) - Cost(C,q)| = \left|\sum_{p\in P} w(p) \cdot dist(p,Q) - \sum_{p\in C} u(p) \cdot dist(p,Q)\right| =$
- $Cost(P,Q) \cdot \left|\sum_{p\in P} w(p) \cdot f(p,Q) - \sum_{p\in C} u(p) \cdot f(p,Q)\right| =$
- $Cost(P,Q) \cdot \left|\bar{f}(P,w,Q) - \bar{f}(C,u,Q)\right|$

Change multiplicative ϵ to additive error ϵ

- Goal: find $(C, u, \mathcal{Q}, f)$ such that for any $Q \in \mathcal{Q}$:
- $\left|\bar{f}(P,w,Q) - \bar{f}(C,u,Q)\right| \leq \epsilon$
- Why?
- Let $f(p,Q) \coloneqq \frac{dist(p,Q)}{Cost(P,Q)} = \frac{dist(p,Q)}{\sum_{p\in P} w(p) \cdot dist(p,Q)}$

Intermediate goal

Given: (P, w), $P \subseteq X$, $w: P \to [0,1)$, sum of weights are 1

Query space: $(P, w, \mathcal{Q}, f)$

- $f: P \times \mathcal{Q} \to [0, \infty)$
- $\bar{f}(P, w, Q) = \sum_{p \in P} w(p) \cdot f(p, q)$

Let $f(p, Q) \coloneqq \frac{dist(p,Q)}{Cost(P,Q)} = \frac{dist(p,Q)}{\sum_{p \in P} w(p) \cdot dist(p,Q)}$

Intermediate Goal: find (C, u) such that for any $Q \in \mathcal{Q}$:

- $\left|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)\right| \leq \epsilon \cdot \max_{p \in P} f(p, q)$
- Why? (roughly, probability bounds work when parameters are between 0,1. otherwise a single input could have the maximum which is quite large. It is also hard to detect using uniform sampling. In other words the variance depends on the maximum).

Intermediate goal

Given: (P, w), $P \subseteq X$, $w: P \rightarrow [0,1)$, sum of weights are 1

Query space: $(P, w, \mathcal{Q}, f)$

- $f: P \times \mathcal{Q} \rightarrow [0, \infty)$
- $\bar{f}(P, w, Q) = \sum_{p \in P} w(p) \cdot f(p, q)$

Let $f(p, Q) \coloneqq \frac{dist(p,Q)}{Cost(P,Q)} = \frac{dist(p,Q)}{\sum_{p \in P} w(p) \cdot dist(p,Q)}$

Intermediate Goal: find (C, u) such that for any $Q \in \mathcal{Q}$:

- $\left|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)\right| \leq \epsilon \cdot \max_{p \in P} f(p, q)$
- Define $s(p) = \max_{Q \in \mathcal{Q}} f(p, Q)$
- Let $t = \sum_{p \in P} w(p) \cdot s(p)$
- Now let: $w'(p) \coloneqq w(p) \cdot \frac{s(p)}{t}$ and let $f'(p, Q) \coloneqq \frac{f(p,Q)}{s(p)}$ Thus $w(p) f(p, Q) = t \cdot w'(p) f'(p, Q)$

Core-set for the new weights

- Define $s(p) = \max_{Q \in \mathcal{Q}} f(p, Q)$
- Let $t = \sum_{p \in P} w(p) \cdot s(p)$
- Now let: $w'(p) \coloneqq w(p) \cdot \frac{s(p)}{t}$ and let $f'(p, Q) \coloneqq \frac{f(p,Q)}{s(p)}$ Thus $w(p) f(p, Q) = t \cdot w'(p) f'(p, Q)$

Suppose (C, u) is $\frac{\epsilon}{t}$ coreset for $(P, w', \mathcal{Q}, f')$, i.e., for any $Q \in \mathcal{Q}$:

- $\left|\overline{f'}(P, w', Q) - \overline{f'}(C, u, Q)\right| \le (\frac{\epsilon}{t}) \cdot \max_{p \in P} f'(p, q)$

Goal: for any $Q \in \mathcal{Q}$:

- $\left|\bar{f}(P, w, Q) - t \cdot \bar{f}(C, u, Q)\right| \le \epsilon$

Proof:

- $\bar{f}(P, w, Q) = t \cdot \overline{f'}(P, w', Q)$
- $\left|\bar{f}(P, w, Q) - t \cdot \bar{f}(C, u, Q)\right| = t \cdot \left|\overline{f'}(P, w', Q) - \overline{f'}(C, u, Q)\right| \le t \cdot \left(\frac{\epsilon}{t}\right) \cdot \max_{Q \in \mathcal{Q}} f'(p, q) \le \epsilon$

Goal: compute ϵ −approximation for f'

- For every positive $r > 0$ define $range(q, r) = \{p \in P | w(p) \cdot f(p, q) \leq r\}$
- Then the dimension of (P, w, Q, f) is the smallest d s.t. for any $S \subseteq P$,
- $|\{range(q, r) | q \in Q, r > 0\}| \leq 2^d$

- E.g. how many subsets can you cover with balls in $\mathbb{R}^d$? $n^{O(d)}$

- Coreset: Let C be a random sample of size $O((\frac{1}{\epsilon^2})(d + \log\frac{1}{\delta})$, then with probability $(1 - \delta)$, it is a ϵ-core-set

- $dist \to Cost \to f \to s(p) \to f', w' \to \epsilon - approx \to random\ sampling$

Bounding Sensitivity for k-median

- $s(p) = \max_{Q \in \mathcal{Q}} \frac{dist(p,Q)}{\sum_{p'} dist(p',Q)}$
- For a specific Q,
- $\frac{dist(p,Q)}{\sum_{p'} dist(p',Q)} \leq \frac{dist(p,q_i^*)}{\sum_{p'} dist(p',Q)} + \frac{dist(q_i^*,Q)}{\sum_{p'} dist(p',Q)} \leq \frac{dist(p,q_i^*)}{\sum_{p'} dist(p',q_i^*)} + \frac{dist(q_i^*,Q)}{\sum_{p'} dist(p',Q)}$
- $dist(q_i^*, Q) \leq dist(q_i^*, p') + dist(p', Q)$
- $|P_i| \cdot dist(q_i^*, Q) \leq \sum_{p' \in P_i} dist(q_i^*, p') + dist(p', Q) \leq 2 \sum_{p' \in P} dist(p', Q)$
- $s(p) \leq \frac{dist(p,q_i^*)}{\sum_{p'} dist(p',q_i^*)} + \frac{2}{|P_i|}$
- $\sum_{p \in P} s(p) = 1 + 2k$

K-median

- $\text{sensitivity}(p) = \frac{dist(p,Q^*)}{\sum_{p' \in P} dist(p',Q^*)} + \frac{1}{n_p}$
  - Q^* is the optimal k-means clustering (again we can use approximation)
  - n_p is the number of points in p's cluster
- Total sensitivity $= 1 + k$
- $|C| = \left(\frac{k^2 d}{\epsilon^2}\right)$
- Combining with PCA gives $\left(\frac{k^2 (\frac{k}{\epsilon})}{\epsilon^2}\right)$
  - Independent of n
  - Independent of d

Rough approximation

- Any bi-criteria approximation, e.g.,

Repeat for $\log n$ iterations:

1. Randomly sample k centers.
2. Remove half of the points that are closest to the centers