Coresets

Lecture 25

Dec 1, 2022

Dealing with Big Data

Compute a smaller summary *quickly*, and use summary instead of original data

- Sampling
- Sketching
- Dimensionality reduction (JL, Subspacee embeddings)
- Streaming summaries
- $\cdots$

Dealing with Big Data

Compute a smaller summary *quickly*, and use summary instead of original data

- Sampling
- Sketching
- Dimensionality reduction (JL, Subspacee embeddings)
- Streaming summaries
- $\cdots$

Today: Coresets a technique from computational geometry

Coresets

$\mathcal{P}$: a collection of n points in $\mathbb{R}^d$

Want to compute some function $f(\mathcal{P})$

- k-cluster $\mathcal{P}$ according to some objective (k-means, k-median, k-center etc)
- find smallest radius ball that encloses $\mathcal{P}$

Coresets

$\mathcal{P}$: a collection of n points in $\mathbb{R}^d$

Want to compute some function $f(\mathcal{P})$

- k-cluster $\mathcal{P}$ according to some objective (k-means, k-median, k-center etc)
- find smallest radius ball that encloses $\mathcal{P}$

Coreset: $\mathcal{Q}$ s.t. $|\mathcal{Q}|$ small and $f(\mathcal{Q}) \simeq f(\mathcal{P})$

- Depends on f
- Ideally, $\mathcal{Q}$ should be computable quickly

Coresets

$\mathcal{P}$: a collection of n points in $\mathbb{R}^d$

Want to compute some function $f(\mathcal{P})$

- k-cluster $\mathcal{P}$ according to some objective (k-means, k-median, k-center etc)
- find smallest radius ball that encloses $\mathcal{P}$

Coreset: $\mathcal{Q}$ s.t. $|\mathcal{Q}|$ small and $f(\mathcal{Q}) \simeq f(\mathcal{P})$

- Depends on f
- Ideally, $\mathcal{Q}$ should be computable quickly

Originally $\mathcal{Q} \subset \mathcal{P}$ (or a weighted subset) and hence name coreset

Part I

Minimum Enclosing Ball

Minimum Enclosing Ball

Given n points $\mathcal{P} \in \mathbb{R}^d$ find smallest radius ball $B(x, r)$ that $\mathcal{P} \subseteq B(x, r)$

Minimum Enclosing Ball

Given n points $\mathcal{P} \in \mathbb{R}^d$ find smallest radius ball $B(x, r)$ that $\mathcal{P} \subseteq B(x, r)$

Exact computation is difficult especially when d is large. Can reduce to convex quadratic optimization leading to arbitrarily good approximation.

Minimum Enclosing Ball

Given n points $\mathcal{P} \in \mathbb{R}^d$ find smallest radius ball $B(x, r)$ that $\mathcal{P} \subseteq B(x, r)$

Exact computation is difficult especially when d is large. Can reduce to convex quadratic optimization leading to arbitrarily good approximation.

Theorem

For any $\mathcal{P} \in \mathbb{R}^d$ there is a set $\mathcal{Q} \subseteq \mathcal{P}$ such that $|\mathcal{Q}| \leq 2/\epsilon$ and MEB of $\mathcal{Q}$ is a $\frac{1}{1+\epsilon}$ approximation to MEB of $\mathcal{P}$.

$\mathcal{Q}$ is an ϵ-coreset for $\mathcal{P}$.

Minimum Enclosing Ball

Given n points $\mathcal{P} \in \mathbb{R}^d$ find smallest radius ball $B(x, r)$ that $\mathcal{P} \subseteq B(x, r)$

Exact computation is difficult especially when d is large. Can reduce to convex quadratic optimization leading to arbitrarily good approximation.

Theorem

For any $\mathcal{P} \in \mathbb{R}^d$ there is a set $\mathcal{Q} \subseteq \mathcal{P}$ such that $|\mathcal{Q}| \leq 2/\epsilon$ and MEB of $\mathcal{Q}$ is a $\frac{1}{1+\epsilon}$ approximation to MEB of $\mathcal{P}$.

$\mathcal{Q}$ is an ϵ-coreset for $\mathcal{P}$.

No dependence on n or d! Differs from sampling/sketching approaches

MEB Algorithm

MEB-Coreset:

$S_1 \leftarrow \{\texttt{arbitrary } p \in \mathcal{P}\}$
`for` $i = 2$ `to` T `do`
  $c_i \leftarrow \texttt{MEB center of } S_{i-1}$
  $p_i \leftarrow \arg\max_{p \in \mathcal{P}} d(p, c_i)$
  $S_i = S_{i-1} \cup \{p\}$
`end for`
`Output` S_T

MEB Algorithm

MEB-Coreset:

```
S_1 ← {arbitrary p ∈ P}
for i = 2 to T do
    c_i ← MEB center of S_{i-1}
    p_i ← arg max_{p∈P} d(p, c_i)
    S_i = S_{i-1} ∪ {p}
end for
Output S_T
```

Claim: If $T = 2/\epsilon$ then S_T is an ϵ-coreset for $\mathcal{P}$.

Analysis: basic lemma about MEB

Lemma

Suppose MEB of $\mathcal{P}$ *is defined by center* c *and radius* R*. Then for every closed half space* H *containing* c *there is a point* $p \in \mathcal{P} \cap H$ *such that* $d(p, c) = R$*.*

Analysis: basic lemma about MEB

Lemma

Suppose MEB of $\mathcal{P}$ is defined by center c and radius R. Then for every closed half space H containing c there is a point $p \in \mathcal{P} \cap H$ such that $d(p, c) = R$.

Proof by contradiction: if not true, for some $\delta > 0$, $d(p, c) \leq R - \delta$ for all $p \in \mathcal{P} \cap H$ (using closedness here). Consider ball of radius R around c. Shifting ball by $\delta/2$ orthogonal to H will create new ball with all points in $\mathcal{P}$ strictly contained inside it. Implies we can shrink ball contradicting the optimality of R.

Analysis of coreset algorithm

c_i MEB center of S_i and r_i radius for S_i.

Let R be optimum radius for $\mathcal{P}$. We have $r_i \leq R$ for all i since $S_i \subseteq \mathcal{P}$. Also $r_{i+1} \geq r_i$ for all i since $S_i \subseteq S_{i+1}$.

Analysis of coreset algorithm

c_i MEB center of S_i and r_i radius for S_i.

Let R be optimum radius for $\mathcal{P}$. We have $r_i \leq R$ for all i since $S_i \subseteq \mathcal{P}$. Also $r_{i+1} \geq r_i$ for all i since $S_i \subseteq S_{i+1}$.

Observation: Let $q \in \mathcal{P} \setminus S_i$ be farthest point from c_i. If $d(c_i, q) = r_i$ then $R \leq r_i$ which implies $r_i = R$.

Analysis of coreset algorithm

c_i MEB center of S_i and r_i radius for S_i.

Let R be optimum radius for $\mathcal{P}$. We have $r_i \leq R$ for all i since $S_i \subseteq \mathcal{P}$. Also $r_{i+1} \geq r_i$ for all i since $S_i \subseteq S_{i+1}$.

Observation: Let $q \in \mathcal{P} \setminus S_i$ be farthest point from c_i. If $d(c_i, q) = r_i$ then $R \leq r_i$ which implies $r_i = R$.

Hence interesting case is when $d(c_i, q) > r_i$. Which implies $r_{i+1} > r_i$. How much bigger does r_{i+1} get?

Define $\lambda_i = \frac{r_i}{R}$.

Analysis of coreset algorithm

Lemma

Either $\boldsymbol{r_i} = \boldsymbol{R}$ or $\boldsymbol{\lambda_{i+1}} \geq \frac{1+\lambda_i^2}{2}$.

Assuming lemma and solving recurrence, $\boldsymbol{\lambda_i} \geq (1 - \frac{1}{1+\frac{i}{2}})$. Thus, if $\boldsymbol{T} = 2/\epsilon$, $\boldsymbol{\lambda_T} \geq \frac{1}{1+\epsilon}$.

Proof of Lemma

Exists $q \in \mathcal{P} \setminus S_i$ such that $d(c_i, q) > R$. Let $\delta_i = d(c_{i+1}, c_i)$ be amount that center moves. $\delta_i > 0$ since $d(c_i, q) > R$.

Proof of Lemma

Exists $q \in \mathcal{P} \setminus S_i$ such that $d(c_i, q) > R$. Let $\delta_i = d(c_{i+1}, c_i)$ be amount that center moves. $\delta_i > 0$ since $d(c_i, q) > R$.

Two lower bounds on r_{i+1}

- By triangle inequality between c_i, c_{i+1}, q we have $d(c_i, c_{i+1}) + d(c_{i+1}, q) \geq d(c_i, q)$ which implies that $\delta_i + r_{i+1} \geq R$ and hence $r_{i+1} \geq R - \delta_i$.

Proof of Lemma

Exists $q \in \mathcal{P} \setminus S_i$ such that $d(c_i, q) > R$. Let $\delta_i = d(c_{i+1}, c_i)$ be amount that center moves. $\delta_i > 0$ since $d(c_i, q) > R$.

Two lower bounds on r_{i+1}

- By triangle inequality between c_i, c_{i+1}, q we have $d(c_i, c_{i+1}) + d(c_{i+1}, q) \geq d(c_i, q)$ which implies that $\delta_i + r_{i+1} \geq R$ and hence $r_{i+1} \geq R - \delta_i$.
- Consider closed half space H containing c_i orthogonal to line segement connecting c_i and c_{i+1} (and not containing c_{i+1}). By basic lemma there exists $p \in S_i$ such that $d(c_i, p) = r_i$. Implies $r_{i+1} \geq d(c_{i+1}, p) \geq \sqrt{r_i^2 + \delta_i^2}$.

Proof of Lemma

Exists $q \in \mathcal{P} \setminus S_i$ such that $d(c_i, q) > R$. Let $\delta_i = d(c_{i+1}, c_i)$ be amount that center moves. $\delta_i > 0$ since $d(c_i, q) > R$.

Two lower bounds on r_{i+1}

- By triangle inequality between c_i, c_{i+1}, q we have $d(c_i, c_{i+1}) + d(c_{i+1}, q) \geq d(c_i, q)$ which implies that $\delta_i + r_{i+1} \geq R$ and hence $r_{i+1} \geq R - \delta_i$.
- Consider closed half space H containing c_i orthogonal to line segement connecting c_i and c_{i+1} (and not containing c_{i+1}). By basic lemma there exists $p \in S_i$ such that $d(c_i, p) = r_i$. Implies $r_{i+1} \geq d(c_{i+1}, p) \geq \sqrt{r_i^2 + \delta_i^2}$.

Therefore $\lambda_{i+1} = \frac{r_{i+1}}{R} \geq \frac{1}{R} \max(R - \delta_i, \sqrt{r_i^2 + \delta_i^2})$.

Proof of Lemma

$$\lambda_{i+1} = \frac{r_{i+1}}{R} \geq \frac{1}{R} \max \left\{ R - \delta_i, \sqrt{r_i^2 + \delta_i^2} \right\}$$

Minimized when $R - \delta_i = \sqrt{r_i^2 + \delta_i^2} = \sqrt{\lambda_i^2 R^2 + \delta_i^2}$ which is when $\delta_i = \frac{(1-\lambda_i^2)R}{2}$.

Proof of Lemma

$$\lambda_{i+1} = \frac{r_{i+1}}{R} \geq \frac{1}{R} \max\left\{R - \delta_i, \sqrt{r_i^2 + \delta_i^2}\right\}$$

Minimized when $R - \delta_i = \sqrt{r_i^2 + \delta_i^2} = \sqrt{\lambda_i^2 R^2 + \delta_i^2}$ which is when $\delta_i = \frac{(1-\lambda_i^2)R}{2}$.

Thus

$$\lambda_{i+1} = \frac{r_{i+1}}{R} \geq \frac{R - \frac{(1-\lambda_i^2)R}{2}}{R} \geq \frac{1+\lambda_i^2}{2}$$

which finishes the proof.

Streaming Coresets

Suppose $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{\boldsymbol{n}}$ come in a stream. Can we compute a small coreset for $\mathcal{P}$?

Streaming Coresets

Suppose $p_1, p_2, \ldots, p_n$ come in a stream. Can we compute a small coreset for $\mathcal{P}$?

Can use Merge and Reduce approach for MEB to maintain an ϵ-coreset storing $O(\frac{\log^2 n}{\epsilon})$ points

Part II

Clustering

Clustering

Given n objects/items $\mathcal{P}$ and integer k find *partition* of $\mathcal{P}$ into k clusters $C_1, \ldots, C_k$ of similar items

Huge topic with many approaches based on domain/application

Center based metric-space clustering:

- $(\mathcal{P}, d)$ is metric space. $d(p, q)$ is distance between p and q
- find centers $S = \{c_1, c_2, \ldots, c_k\}$ such that $C_i = \{p \in \mathcal{P} : c_i \text{ is closest center to } p\}$.
- different objectives define different optimization problems: k-median, k-means, k-center etc
- choice of centers: $S \subset \mathcal{P}$ or S can be in ambient space if $\mathcal{P} \in \mathbb{R}^d$. Typically within factor of 2 in objective but clustering quality and algorithmic difficulty can be different.

k-median, k-means, k-center

Given $\mathcal{P}$ and k find k centers S such that

- k-median: minimize $\sum_{p \in \mathcal{P}} d(p, S)$
- k-means: minimize $\sum_{p \in \mathcal{P}} (d(p, S))^2$
- k-center: minimize $\max_{p \in \mathcal{P}} d(p, S)$
- spacial cases of ℓ_p clustering: minimze $\sum_{p \in \mathcal{P}} (d(p, S))^p$ for some $p \geq 1$.

Coresets for Clustering

Given $\mathcal{P}, k$ and ϵ find *weighted* point set $\mathcal{Q}$ such that clustering cost of $\mathcal{Q}$ is ϵ-approximation to that of $\mathcal{P}$.

Two techniques:

- In geometric settings of low dimension via gridding techniques [HarPeled-Mazumdar]
- Higher dimensions and metric spaces [Chen, Feldman-Langberg] and many others using importance sampling

Many results including very recent work: size of coreset, running time to build coreset, dependence on d vs k, etc etc

Coresets for Clustering

Given $\mathcal{P}, k$ and ϵ find *weighted* point set $\mathcal{Q}$ such that clustering cost of $\mathcal{Q}$ is ϵ-approximation to that of $\mathcal{P}$.

Some known results:

- $O(\text{poly}(k, \log n, 1/\epsilon)$ for a ϵ-approximate core set for k-median and k-means in general metric spaces [Chen'09]
- $O(kd/\epsilon^2)$ for points in $\mathbb{R}^d$ [Feldman-Langberg'11]
- $O(\text{poly}(k, 1/\epsilon))$ independent of dimension [Feldman-Schmidt-Sohler'13, Sohler-Woodruff'19]
- Dimension reduction to $O(k \log k/\epsilon^2)$ dimensions [Makarychev-Makarychev-Razenshteyn'19]

Importance Sampling for Coresets

High-level idea: Start with a crude approximation and use it for sampling [Chen]. Refined substantially later [Feldman-Langberg] and follow up work.

(α, β)-bicriteria-approximation for k-clustering:

- centers S such that $|S| \leq \alpha k$
- $\text{cost}(S, \mathcal{P}) \leq \beta \cdot \text{cost}(S^*, \mathcal{P})$ where S^* is an optimal center set

Here $\alpha, \beta \geq 1$. Both # of centers and cost approximate

Computing (α, β)-approximation fast is possible using various ideas.

Coresets for k-median

Suppose S is an (α, β)-bicriteria-approximation for k-median
$S = \{c_1, c_2, \ldots, c_h\}$ partitions $\mathcal{P}$ into $\mathcal{P}_1, \ldots, \mathcal{P}_h$

$\text{cost}(S, \mathcal{P}) = \sum_{i=1}^{h} \text{cost}(c_i, \mathcal{P}_i)$

Coresets for k-median

Suppose S is an (α, β)-bicriteria-approximation for k-median
$S = \{c_1, c_2, \ldots, c_h\}$ partitions $\mathcal{P}$ into $\mathcal{P}_1, \ldots, \mathcal{P}_h$

$\text{cost}(S, \mathcal{P}) = \sum_{i=1}^{h} \text{cost}(c_i, \mathcal{P}_i)$
Intuitively treat as h separate 1-median problems.

Coresets for k-median

Suppose S is an (α, β)-bicriteria-approximation for k-median
$S = \{c_1, c_2, \ldots, c_h\}$ partitions $\mathcal{P}$ into $\mathcal{P}_1, \ldots, \mathcal{P}_h$

$\text{cost}(S, \mathcal{P}) = \sum_{i=1}^{h} \text{cost}(c_i, \mathcal{P}_i)$
Intuitively treat as h separate 1-median problems.

Consider c_1 and $\mathcal{P}_1$. $\text{cost}(c_1, \mathcal{P}_1) = \sum_{p \in \mathcal{P}_1} d(p, c_1)$ Hence sample a point $p \in \mathcal{P}_i$ with probability $d(p, c_1)/\text{cost}(c_1, \mathcal{P}_1)$. Take several samples to control variance etc.

Coresets for k-median

Suppose S is an (α, β)-bicriteria-approximation for k-median
$S = \{c_1, c_2, \ldots, c_h\}$ partitions $\mathcal{P}$ into $\mathcal{P}_1, \ldots, \mathcal{P}_h$

$\text{cost}(S, \mathcal{P}) = \sum_{i=1}^{h} \text{cost}(c_i, \mathcal{P}_i)$
Intuitively treat as h separate 1-median problems.

Consider c_1 and $\mathcal{P}_1$. $\text{cost}(c_1, \mathcal{P}_1) = \sum_{p \in \mathcal{P}_1} d(p, c_1)$ Hence sample a point $p \in \mathcal{P}_i$ with probability $d(p, c_1)/\text{cost}(c_1, \mathcal{P}_1)$. Take several samples to control variance etc.

Actual scheme and analysis more tricky. Have to argue that sampling is good for potentially $\binom{n}{k}$ clusterings; coreset size becomes $\text{poly}(k, \log n)$. Geometry/VC-Dimension analysis to avoid dependence on n and reduce to d. Can change d to k via dimensionality reduction (not easy).