Lecture 22 Core sets

A coreset: "Turning big data into tiny data"

Idea: dramatically reduce the number of points while the cost of <u>any</u> solution is well approximated

I. A coreset for (Euclidean) k-means clustering

Let $D = D_{\ell_2^2}$. Let $D(A,C) := \sum_{x \in A} \min_{c \in C} \|x-c\|^2$.

Let $w: A \to \mathbb{R}^+$ be a weight function.

Let $D(A,w,C) := \sum_{x \in A} w(x) \cdot \min_{c \in C} \|x-c\|^2$.

<u>Def</u>: Let $A \subseteq \mathbb{R}^d$ be a set of n points, $\varepsilon \in (0,1)$, and $k \in \mathbb{N}^{\geq 1}$. A set $S \subseteq \mathbb{R}^d$ together with a weight function $w: S \to \mathbb{R}^+$ is called a <u>(k,ε)-coreset</u> if for all $C \subseteq \mathbb{R}^d$ with $|C| \leq k$,

$$|D(A,C) - D(S,w,C)| \leq \varepsilon \cdot D(A,C).$$

<u>Basic idea</u>: Given A, first compute a discretization S of the space, snap each point to its nearest neighbor in S, and then use the (weighted) set S to approximate the cost function.

II Some tools:

Def: An <u>ε-ball-cover</u> of unit sphere (in Euclidean spaces) is a set of points B, such that for any point p in the unit sphere, the distance between p to B is at most ε.

Lemma1: Let U be the unit sphere in d-dimensional Euclidean space. Then for every $0<\varepsilon<1$ there exists an ε-ball-cover B of size $(1+\frac{2}{\varepsilon})^d$, i.e. for every point $p \in U$, $\min_{b \in B} \|p-b\| \le \varepsilon$.

Remark: ① The above ball-cover is existential, There are algorithms for constructing one using $\varepsilon^{-O(d)}$ points.

② If not unit sphere, say U is a ball of radius r, then <u>ε-ball-cover</u> is the set B with $\min_{b \in B} \|p-b\| \le \varepsilon \cdot r$.

Thm1: There exists a 6.357-approximation algorithm for k-means problem. [Ahmadian et. al. 2017]

Remark: Any constant approximation will work for constructing a coreset.

Lemma 2 (Generalized Triangle Inequality) Let a, b, c be points in Euclidean space. Then for any $\varepsilon \in (0,1)$, we have

$$\left| \|a-c\|^2 - \|b-c\|^2 \right| \le \frac{12}{\varepsilon} \|a-b\|^2 + 2\varepsilon \cdot \|a-c\|^2$$

$$\left| D(a,C) - D(b,C) \right| \le \frac{12}{\varepsilon} D(a,b) + 2\varepsilon \cdot D(a,C)$$

III. A (non-efficient) alg for constructing a (k,ε)-coreset for k-means clustering

Input A, k

1. first compute a 10-approximation C' to the k-means problem, which can be done in polynomial time. (Eg. By Thm 1)
   Let $C_1, \dots, C_k$ be the corresponding k clusters with centers $c_1, \dots, c_k$.
2. for each $j = 1, \dots, k$,
   Let $F = C_j$
   Let B^i be the ball with radius $r_i = \frac{2^i}{n} \sum_{x \in F} \|x - c_j\|^2$ centered at c_j and
   let S^i be the $\frac{\varepsilon}{192}$-ball-cover of B^i, for $i = 0, \dots, \log(n)$

let $S_j = \bigcup_{i=0}^{\log 10n} S^i$

3. For each $x \in A$, let $B(x) \in \bigcup_{j=1}^{k} S_j$ be the nearest point in the union of all ball-covers.
   Let $S = \bigcup_{x \in A} B(x)$.
   For each $y \in S$, let $w(y)$ be the number of $x \in A$ such that $B(x) = y$.
4. Output (S, w)

IV Analysis:

Lemma 3: Let F be a set of n points. Let B^i be the ball with radius $r_i := \sqrt{\frac{2^i}{n} \sum_{x \in F} \|x\|^2}$ centered at the origin and let S^i be the $\frac{\varepsilon}{3}$-ball-cover of B^i. Denote by $S = \bigcup_{i=0}^{\log 10n} S^i$. Then $\sum_{x \in A} \min_{s \in S} \|x - s\|^2 \le \varepsilon^2 \sum_{x \in F} \|x\|^2$.

Proof: $F_{close} = \{ y \in F \mid \|y\|^2 \le \frac{1}{n} \sum_{x \in F} \|x\|^2 \}$

$F_{far} = F \setminus F_{close}$.

Since $|F_{close}| \le n$, we have

$$\sum_{x \in F_{close}} \min_{s \in S^0} \|x - s\|^2 \le |F_{close}| \cdot \frac{1}{n} \sum_{x \in F} \|x\|^2 \cdot \frac{\varepsilon^2}{9}$$

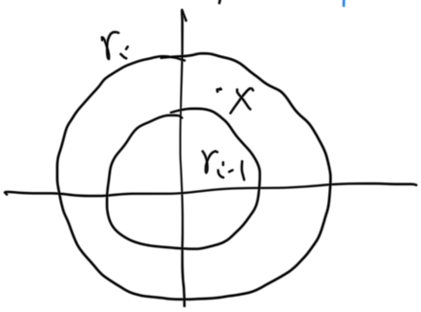
$$\le \frac{\varepsilon^2}{9} \sum_{x \in F} \|x\|^2.$$

For points in F_{far}, consider any point x in $B^i \setminus B^{i-1}$ for $i \in \{1, \dots, \log 10n\}$. We have

$$\min_{s \in S^i} \|x - s\|^2 \le \frac{\varepsilon^2}{9} \cdot r_i^2 \le \frac{4\varepsilon^2}{9} \cdot r_{i-1}^2 \le \frac{4\varepsilon^2}{9} \cdot \|x\|^2$$

Summing up over all points, we have $\sum_{x \in F} \min_{s \in S} \|x - s\|^2$

$$\le \frac{\varepsilon^2}{9} \cdot \sum_{x \in F} \|x\|^2 + \frac{4\varepsilon^2}{9} \sum_{x \in F} \|x\|^2 < \varepsilon^2 \sum_{x \in F} \|x\|^2$$



the best poly(k, ε^{-1}).

> **Thm 1.** The alg described above finds a coreset for k-means consisting of $O(k \varepsilon^{-d} \log n)$ points, where d is the (constant) dimension.

Proof: For each $F \in \{C_1, \dots, C_k\}$, we have a total of $\log 10n$ balls with varying radii. For each such ball of radius r, we compute an $\frac{\varepsilon}{48}$-ball-cover.

For any point $x \in A$, let $B(x)$ be the nearest point in the union of all ball-covers.

By Lemma 3, we have $\sum_{x \in A} \|x - B(x)\|^2 = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - B(x)\|^2$

$$\le \left(\frac{\varepsilon}{16}\right)^2 \cdot \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$$

$$\le \left(\frac{\varepsilon}{16}\right)^2 \cdot 10 \cdot D(A, C^*)$$

i.e. OPT

Recall the definition of S, w from the algorithm.

Now consider an arbitrary set of centers C. We have

$$\left| D(A, C) - D(S, w, C) \right|$$

$$= \left| \sum_{x \in A} \min_{c \in C} \|x - c\|^2 - \sum_{x \in A} \min_{c \in C} \|B(x) - c\|^2 \right|$$

$$\le \sum_{x \in A} \left| \min_{c \in C} \|x - c\|^2 - \min_{c \in C} \|B(x) - c\|^2 \right|$$

$$\le \frac{12}{\varepsilon} \sum_{x \in A} \|x - B(x)\|^2 + 2\varepsilon \sum_{x \in A} \min_{c \in C} \|x - c\|^2$$

$$\le \frac{12}{\varepsilon} \left(\frac{\varepsilon}{16}\right)^2 \cdot 10 \cdot D(A \cdot C^*) + 2\varepsilon \cdot \sum_{x \in A} \min_{c \in C} \|x - c\|^2$$

$$< 2 \cdot \varepsilon \cdot D(A, C^*) + 2\varepsilon \cdot \sum_{x \in A} \min_{c \in C} \|x - c\|^2$$

$$\le 4\varepsilon \cdot \sum_{x \in A} \min_{c \in C} \|x - c\|^2$$

Rescaling ε by a factor of $1/4$ finishes the proof. $\square$

Space: $O(k \cdot \log n)$ ball-covers, each with $O(\varepsilon^{-d})$ points.

V: Constructing coresets in the streaming (i.e. points arrive sequencially; limited storage)

Merge & Reduce technique.

① Composability

$\boxed{\text{Lemma}}$: Let $A_1, A_2 \subseteq \mathbb{R}^d$ be two disjoint point sets. Assume S_1 with $w_1 : S_1 \to \mathbb{R}$ and S_2 with $w_2 : S_2 \to \mathbb{R}$ are (k, ε)-coresets for A_1, A_2, respectively. Then $S_1 \cup S_2$ with $w_1 + w_2 : S_1 \cup S_2 \to \mathbb{R}$ is a (k, ε)-coreset for $A_1 \cup A_2$.

② Idea:

Assume n points in the sequence are partitioned into $\frac{n}{s}$ batches, each of size s.

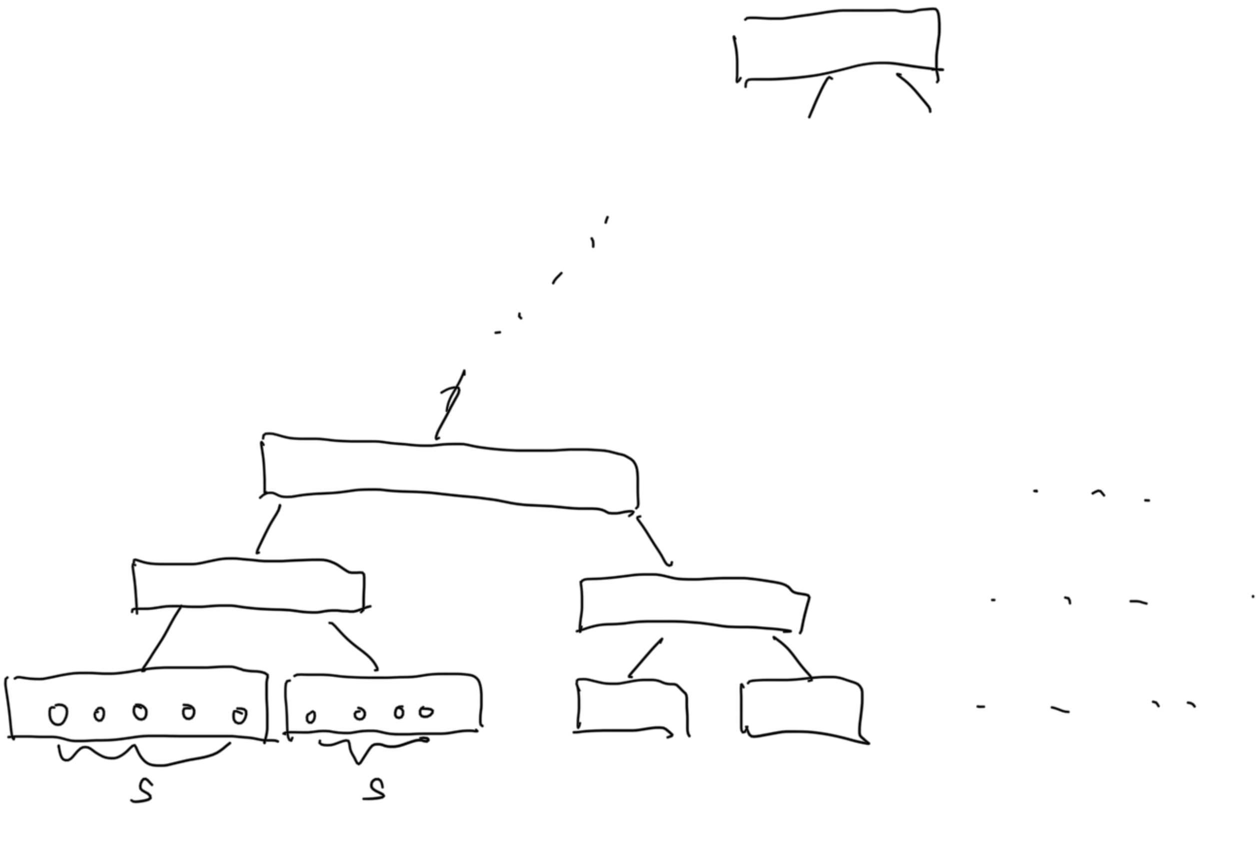
These batches form the leaves of a binary tree of height $h \leq \log(n/s)$

For each batch, we compute a coreset

Whenever we have computed a coreset for two children of a node, we aggregate them by recomputing a coreset of the union of the children and storing it in the parent node.

The children can be then deleted at this point.

We only need store ≤ 2 coresets at each level.



Note: the final output (i.e. coreset at the root node) has approximation $(1+\varepsilon')^{\log n}$, where ε' is the parameter for each leaf. So we set $\varepsilon' = \frac{\varepsilon}{2(\log n+1)}$, then

$$(1+\varepsilon')^{\log n} < 1+\varepsilon.$$

3 The algorithm

Initialize (k, ε, s)

1. Initialize an empty coreset B and initialize an empty dynamic array A
2. Set $\varepsilon' = \varepsilon / (2(\log n + 1))$, $i_{max} = 0$
3. Store k, s.

Update $(x \in \mathbb{R}^d)$

1. Store x in B
2. If $|B| = s$ do
3.   Set $i = 0$
4.   While $A[i] \neq \phi$ do
5.   Compute a (k, ε')-coreset for $(A[i] \cup B, k)$, store it in B and empty $A[i]$
6.   Set $A[i] = \phi$, $i = i+1$, update $i_{max} = \max\{i, i_{max}\}$
7.   End While
8.   Set $A[i] = B$ and empty B
9. EndIf

After any update, we can call the following query( ) to get the coreset for the points read so far.

query( )

1. Set $T = B$
2. For $i = 0, \dots, i_{max}$, do
3. Set $T = T \cup A[i]$
4. Compute a (k, ε')-coreset S for (T, k)
5. Return S

Thm: The above alg:

Space $O(k \varepsilon^{-d} (\log n)^{d+2})$

(k, ε)-coreset of size $O(k (\log n)^{d+1} \cdot \varepsilon^{-d})$.