CMPSCI 711: More Advanced Algorithms

Section 3-1: Coresets and Clustering

Andrew McGregor

Last Compiled: April 29, 2012

Geometric Streams

- Consider a stream of points:

$$P = \langle p_1, \ldots, p_n \rangle$$

  where each $p_i \in \mathbb{R}^d$.
- What properties of P can we compute in sub-linear space?

Outline

Coresets

Clustering

Coresets

- *Goal:* Minimize a function $C_P : \mathbb{R}^d \to \mathbb{R}$ parameterized by $P \subset \mathbb{R}^d$.
- For example, finding the minimum enclosing ball corresponds to finding the ball's center $c = \text{argmin}\, C_p(x)$ and radius $C_P(c)$ where

$$C_P(x) = \max_{y \in P} \|x - y\|_2$$

- We'll assume C_P is monotone, i.e.,

$$\forall x \in \mathbb{R}^d, Q \subset P \ ; \quad C_Q(x) \leq C_P(x) \ .$$

- *Defn:* We say $Q \subseteq P$ is a α-coreset for P with respect to C if

$$\forall x \in \mathbb{R}^d, T \subset \mathbb{R}^d \ ; \quad C_{Q \cup T}(x) \leq C_{P \cup T}(x) \leq \alpha C_{Q \cup T}(x) \ .$$

- Hence, if we have a coreset Q for P then we can approximate the original problem up to a factor α.
- We'll first show that the existence of small coresets gives rise to small-space stream algorithms. We'll then show the small coresets exist of the minimum enclosing ball problem.

Properties of Coresets

- *Merge Property:* If Q is an α-core set for P and Q' is a β-coreset for P' then $Q \cup Q'$ is an $(\alpha\beta)$-coreset of $P \cup P'$.
- *Reduce Property:* If Q is an α-core set for P and R is a β-coreset for Q then R is an $(\alpha\beta)$-coreset of P.
- *Thm:* Suppose there exists an $(1+\delta)$-coreset of size $f(\delta)$ that is computable in linear space. Then there's a $O(f(\epsilon/\log n)\log n)$ space, $(1+O(\epsilon))$-approximation streaming algorithm.
- *Proof:* Via a recursive tree construction as in graph sparsification.

Minimum Enclosing Ball: Preliminaries

- For non-zero vectors $u, v \in \mathbb{R}^n$, define $\text{angle}(u, v) := \arccos \frac{u.v}{\|u\|_2 \|v\|_2}$
- For $\theta > 0$, we say $U = \{u_1, \ldots, u_t\} \subseteq \mathbb{R}^d \setminus \{0\}$ is a θ-grid if,

$$\forall x \in \mathbb{R}^d,\ \exists u \in U,\ \text{angle}(x, u) \leq \theta$$

- *Thm:* There exists a θ-grid U of size $O(1/\theta^{d-1})$ and we may assume that U consists of unit vectors.

Minimum Enclosing Ball: Coreset

- Given P, we'll construct a coreset $Q \subseteq P$ using a θ-grid U for some value of θ to be determined.
- For each $u \in U$, add the following points to Q:

$$\arg\max_{p \in P}(p.u) \quad \text{and} \quad \text{argmin}_{p \in P}(p.u)$$

- Need to show that for some $\alpha(\theta) \geq 1$, for any $T \subset \mathbb{R}^d$, $x \in \mathbb{R}^d$,

$$C_{Q \cup T}(x) \leq C_{P \cup T}(x) \leq \alpha(\theta) C_{Q \cup T}(x)$$

- Left inequality follows easily from the definition

$$C_Y(x) = \max_{y \in Y} \|x - y\|_2$$

- *Lemma:* Right inequality holds with $\alpha(\theta) = 1 + \theta^2$.
- Hence, setting $\theta = \sqrt{\epsilon}$ ensures Q is a $(1+\epsilon)$ coreset for P.

Proof of Lemma

- Consider arbitrary $T \subset \mathbb{R}^d$, $x \in \mathbb{R}^d$ and let z be farthest point from x in $P \cup T$.
- If $z \in T$: $C_{P\cup T}(x) = \|x - z\|_2 \leq C_{Q\cup T}(x)$
- If $z \in P$: There exists $u \in U$ such that angle$(u, z - x) \leq \theta$
  - Let y be point with $\|x - y\|_2 = \|x - z\|_2$ that maximizes $u.y$.
  - Let z' be the projection of z in the direction $y - x$.
  - By construction Q contains a point q with $u.z' \leq u.q$.
  - Hence,

$$C_{Q\cup T}(x) \geq C_Q(x) \geq \|x - z'\|_2 = \|x - z\|_2 \cos\theta = C_{P\cup T}(x)\cos\theta \ .$$

  - Result follows because $\frac{1}{\cos\theta} \leq 1 + \theta^2$ for small θ.

Outline

Coresets

Clustering

k-center

- Given a stream of distinct points $P = \{p_1, \ldots, p_n\}$, find the set of k points $Y \subset X$ that minimizes:

$$\max_i \min_{y \in Y} d(p_i, y)$$

  where d can be $\|\cdot\|_2$ or any metric. Let r be the optimum value.
- Can find 2 approx. in $O(k)$ space if you know r ahead of time.
  - Add a new point p to Y if $\min_{y \in Y} d(y, p) > 2r$.
  - Can never have more than k points in Y: Otherwise we'd have $k + 1$ points with all pairwise distances $> 2r$. Each optimal center covers at most one point in Y within radius r. Hence $|Y| \leq k$.
- Can find $(2 + \epsilon)$ approx. in $O(k\epsilon^{-1} \log(b/a))$ space if you know

$$a \leq r \leq b$$

- *Thm:* $(2 + \epsilon)$ approx. in $O(k\epsilon^{-1} \log \epsilon^{-1})$ space.

k-center: Sketch of Algorithm and Analysis

- Consider first $k+1$ points: this gives a lower bound a for r.
- Instantiate basic algorithm with guesses

$$\ell_1 = a,\ \ell_2 = (1+\epsilon)a,\ \ell_3 = (1+\epsilon)^2 a, \ldots\ \ell_{1+t} = O(\epsilon^{-1})a$$

- Say instantiation goes bad if it tries to open $(k+1)$-th center
- If instantiation for guess ℓ goes bad when processing $(j+1)$-th point
  - Let $q_1, \ldots, q_k$ be centers chosen so far.
  - Then $p_1, \ldots, p_j$ are all at most 2ℓ from some q_i.
  - Optimum for $\{q_1, \ldots, q_k, p_{j+1}, \ldots, p_n\}$ is at most $r + 2\ell$.
- Hence, for an instantiation with guess $2\ell/\epsilon$ only incurs a small error if we use $\{q_1, \ldots, q_k, p_{j+1}, \ldots, p_n\}$ rather than $\{p_1, \ldots, p_n\}$.