Randomized Algorithms 2023A – Lecture 11
Coresets for Clustering*

Robert Krauthgamer

1 Coresets for Clustering

Let $D(\cdot,\cdot)$ denote the Euclidean distance in $\mathbb{R}^d$.

Geometric Clustering: In the *k-median problem* the input is a set of n data points $X = \{x_1,\dots,x_n\} \subset \mathbb{R}^d$, and the goal is to find a set of k centers $C = \{c_1,\dots,c_k\} \subset \mathbb{R}^d$ that minimizes the objective function

$$f(X,C) := \sum_{x\in X} D(x,C) = \sum_{i\in[n]} \min_{j\in[k]} \|x_i - c_j\|_2.$$

Note that the centers are not required be from X (the version with this requirement is called discrete centers).

The *k-means problem* is similar but using squared distances.

Notation: We shall omit the subscript from all norms, as we always use ℓ_2 norms.

Observe that points need not be distinct, i.e., we consider multisets, which is equivalent to giving every point an integer weight, and admits a succinct representation. We thus would like to reduce the number of *distinct* points, denoted throughout by $|X|$.

Strong Coreset: Let $\epsilon \in (0,1/2)$ be an accuracy parameter. We say that $S \subset \mathbb{R}^d$ is a strong ε-coreset of X (for objective f, which in our case is k-median) if

$$\forall C = \{c_1,\dots,c_k\} \subset \mathbb{R}^d, \qquad f(X,C) \in (1\pm\varepsilon) f(S,C).$$

Note: A weak coreset is similar, except the above requirement is only for the optimal centers for the coreset, i.e., C' that minimizes $f(S,C')$.

Goal: We want to construct small coresets. If done without computing an optimal solution C^*, then it would be useful for computing a near-optimal solution, because it suffices to solve k-median

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

on the smaller instance S. If the construction requires computing C^*, it could still be useful when sending (communicating) or storing the data.

We focus henceforth on existence (of coresets of a certain size), the algorithmic implementation and applications are usually straightforward.

2 Coresets via Geometric Decomposition

Idea: Discretize the space to create a small set $\hat{S}$, and "snap" every point in X to its nearest neighbor in S. Throughout, the (closed) ball of radius $r > 0$ about $c \in \mathbb{R}^d$ is defined as

$$B(c, r) = \{z \in \mathbb{R}^d : \|z - c\| \leq r\}.$$

Lemma 1 (ε-Ball Cover): For every $\varepsilon \in (0, 1)$, the unit ball $B = B(\vec{0}, 1)$ in $\mathbb{R}^d$ can be covered by $(3/\varepsilon)^d$ balls of radius ε.

The conclusion is that every point in the unit ball can be "approximated" by one of those $(3/\varepsilon)^d$ centers, with additive error ε. This argument immediately extends to any ball of radius $r > 0$, except that the additive error is now εr.

Exer: Prove this lemma.

Hint: Construct the covering iteratively, and use the volume estimate $\mathrm{vol}(B(c, r)) = r^d \cdot \mathrm{vol}(B(\vec{0}, 1))$.

Theorem 2: Every set X of n points in $\mathbb{R}^d$ admits an ε-coreset S of cardinality $|S| = O(k(9/\varepsilon)^d \log n)$.

Proof: Was seen in class.

Exer: Modify the above proof to be algorithmic, by using an $O(1)$-approximation to the minimum cost (meaning a set C' such that $f(X, C') \leq O(1) \cdot f(X, C^*)$), which can be computed in polynomial time.

Exer: Extend this argument to k-means using the following generalized triangle inequality: For every $a, b, c \in \mathbb{R}^d$ and $\varepsilon \in (0, 1)$,

$$\big| \|a - c\|^2 - \|b - c\|^2 \big| \leq \tfrac{12}{\varepsilon} \|a - b\|^2 + 2\varepsilon \|a - c\|^2.$$

2.1 Strong Coresets via Importance Sampling

Definition: The *sensitivity* of a point $x \in X$ is

$$s(x) := \sup_{c \in \mathbb{R}^d} \frac{\|x - c\|}{\sum_{z \in X} \|z - c\|},$$

and the *total sensitivity* of X is $S(X) = \sum_{x \in X} s(x)$.

Observe that for a given $c \in \mathbb{R}^d$ (i.e., without the supremum) the above ratio is the "desired" sampling probability in Importance Sampling.

Importance Sampling approach: Suppose we sample one point, where each $x \in X$ is picked with probability $q(x) := \frac{s(x)}{S(X)}$. We then give the sampled x new weight $\frac{1}{q(x)}$. Of course, we should average a few repetitions to reduce variance.

Lemma 3: $S(X) \leq 6$.

Lemma 4: Let Y be a multiset of $m \geq 24/\varepsilon^2$ points, each sampled iid from X according to $q(\cdot)$. Then

$$\forall c \in \mathbb{R}^d, \qquad \Pr\Big[\tfrac{1}{m} \sum_{y \in Y} \frac{\|y - c\|}{q(y)} \in (1 \pm \varepsilon) \sum_{x \in X} \|x - c\|\Big] \geq 3/4.$$

This does not give a strong coreset, but it is an important step in that direction.

Proof of Lemma 3: Was seen in class by bounding each $s(x) \leq \frac{4}{n} + \frac{\|x - c^*\|}{\mathrm{OPT}/2}$.

Proof of Lemma 4: Was seen in class by applying the Importance Sampling Theorem seen in the previous class for each $y \in Y$.

Amplifying the probability: We would like to improve the success probability in Lemma 9 to $1 - \delta$. Using Chebyshev's inequality, this would require increasing m by a factor of $\frac{1}{\delta}$.

Using Chernoff-Hoeffding concentration bounds would be better and require increasing m only by a factor of $O(\log \frac{1}{\delta})$. But for this, we need that no one sample $y \in Y$ ever contributes too much, which indeed holds in our setting.

Lemma 5: $\hat{Z} \leq S(X) \cdot \mathbb{E}[\hat{Z}]$ with probability 1.

Exer: Prove this lemma.

Lemma 6: The success probability in Lemma 4 can be improved $1 - \delta$ by using $m \geq L\varepsilon^{-2} \log \frac{1}{\delta}$ for a suitable constant $L > 0$.

Exer: Prove this lemma using concentration bounds.

Strong Coreset: To obtain a strong coreset, we must consider any $c \in \mathbb{R}^d$. If there were only a few potential centers, then we could apply Lemma 11 to each of them together with a union bound.

The idea is then to discretize the space of potential centers using the ε-ball cover lemma, and show that it suffices to consider only these centers. Then it would suffice to apply Lemma 4 and a union bound.

Theorem 7: Let Y be a multiset of $m \geq L' d\varepsilon^{-2} \log \frac{1}{\varepsilon}$ points from X, each sampled iid according to distribution $q(.)$ and reweighted by $w(x) = \frac{1}{mq(x)}$, for a suitable constant $L' > 0$. Then with high probability, Y is a strong coreset for the 1-median of X.

Due to time constraints, we saw in class only an outline of the proof, which is based on the lemma below, and on discretizing the possible centers using a ball cover (namely, a cover of the ball $B^* = B(c^*, \frac{1}{\varepsilon}\frac{\mathrm{OPT}}{n})$ by balls of radius $\varepsilon\frac{\mathrm{OPT}}{n}$).

One potential obstacle is the total weight of Y. It need not be n, but with high probability should be close.

Lemma 8: Under the conditions of Lemma 6, i.e., $m \geq L\varepsilon^{-2} \log \frac{1}{\delta}$,

$$\Pr[w(Y) \in (1 \pm \varepsilon)n] \geq 1 - \delta.$$

Exer: Prove this lemma using concentration bounds.

Hint: Write $w(Y) = \frac{1}{m} \sum_{y \in Y} \frac{1}{q(y)}$, show a bound $\frac{1}{q(x)} \leq O(n)$ (with probability 1), and then use concentration bound.