Introduction to Coresets

Isak Falk[1]

May 27, 2020

[1] ucabitf@ucl.ac.uk

Outline

Informal Statement

Coreset is a modern data summarization that approximates the original data in some provable sense with respect to a (usually infinite) set of questions, queries or models and an objective loss/cost function. – (Introduction to Coresets: Accurate Coresets [5])

Formal Statement

Ingredients

- $\mathcal{X}$ is called the *query set*
- $P' = (P, w)$ is a weighted set called the *input set*
  - The weighing function $w : P \to \mathbb{R}$ assigns a weight to each $p \in P$
- $f : P \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is called the *cost function*
- $\ell : \cup_{n=1}^{\infty} \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is called the *loss*

Query space

The tuple $(P, w, \mathcal{X}, f, \ell)$ is called a *query space*.

Accurate Coresets

For a query space $(P, w, \mathcal{X}, f, \ell)$, the *fitting error* of any weighted set $C' = (C, u) = (\{c_1, \ldots, c_m\}, u)$ and $x \in \mathcal{X}$ is

$$f_\ell(C', x) = \ell(u(c_1)f(c_1, x), \cdots, u(c_m)f(c_m, x)).$$

C' is called an *accurate coreset* for the query space $(P, w, \mathcal{X}, f, \ell)$ if for every $x \in \mathcal{X}$

$$f_\ell(C', x) = f_\ell(P', x)$$

Practical Examples

Accurate coresets can be found for many settings including least squares

Practical Examples

Name	Input Weighted Set (P, w) of size $\lvert P\rvert = n$	Query Set $\mathcal{X}$	cost function $f : P \times \mathcal{X}$	loss for $f(p, x)$ over $p \in P$	Coreset C	Coreset Weights	Const. time	Query time	Section
1-Center	$P \subseteq \ell \subseteq \mathbb{R}^d$ $w \equiv 1$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = \lVert p - x\rVert$	$\lVert\cdot\rVert_\infty$	$C \subseteq P$ $\lvert C\rvert = 2$	$u \equiv 1$	$O(n)$	$O(d)$	3.1
Monotonic function	$P \subseteq \mathbb{R}$ $w \equiv 1$	$\mathcal{X} = \{g \mid g$ is monotonic decreasing/increasing or increasing and then decreasing function$\}$	$f(p, g) = g(p)$	$\lVert\cdot\rVert_\infty$	$C \subseteq P$ $\lvert C\rvert = 2$	$u \equiv 1$	$O(n)$	$O(1)$	3.2
Vectors sum (1)	$P \subseteq \mathbb{R}^d$ $w : P \to \mathbb{R}$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = p - x$	Σ	$C \subseteq \mathbb{R}^d$ $\lvert C\rvert = 1$	$u \equiv \sum_{p\in P} w(p)$	$O(n)$	$O(d)$	3.3
Vectors sum (2)	$P \subseteq \mathbb{R}^d$ $w : P \to \mathbb{R}$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = p - x$	Σ	$C \subseteq P$ $\lvert C\rvert \le d + 1$	$u : C \to \mathbb{R}$ $\sum_{p\in C} u(p) = \sum_{p\in P} w(p)$	$O(nd^2)$	$O(d^2)$	3.3.1
Vectors sum (3)	$P \subseteq \mathbb{R}^d$ $w : P \to [0, \infty)$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = p - x$	Σ	$C \subseteq P$ $\lvert C\rvert \le d + 2$	$u : C \to \left[0, \sum_{p\in P} w(p)\right]$ $\sum_{p\in C} u(p) = \sum_{p\in P} w(p)$	$O(\min\{n^2d^2, nd + d^4 logn\})$	$O(d^2)$	3.3.2
1-Mean (1)	$P \subseteq \mathbb{R}^d$ $w : P \to \mathbb{R}$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = w(p) \lVert p - x\rVert^2$	$\lVert\cdot\rVert_1$	$C \subseteq \mathbb{R}^d \times \mathbb{Z} \times \mathbb{R}$ $\lvert C\rvert = 3$ Different loss	*unweighted*	$O(nd)$	$O(d)$	3.4
1-Mean (2)	$P \subseteq \mathbb{R}^d$ $w : P \to \mathbb{R}$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = w(p) \lVert p - x\rVert^2$	$\lVert\cdot\rVert_1$	$C \subseteq P$ $\lvert C\rvert \le d + 2$	$u : C \to \mathbb{R}$ $\sum_{p\in C} u(p) = \sum_{p\in P} w(p)$ $\sum_{p\in C} u(p) \lVert p\rVert^2 = \sum_{p\in P} w(p) \lVert p\rVert^2$	$O(nd^2)$	$O(d^2)$	3.4.1
1-Mean (3)	$P \subseteq \mathbb{R}^d$ $w : P \to [0, \infty)$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = w(p) \lVert p - x\rVert^2$	$\lVert\cdot\rVert_1$	$C \subseteq P$ $\lvert C\rvert \le d + 3$	$u : C \to \left[0, \sum_{p\in P} w(p)\right]$ $\sum_{p\in C} u(p) = \sum_{p\in P} w(p)$ $\sum_{p\in C} u(p) \lVert p\rVert^2 = \sum_{p\in P} w(p) \lVert p\rVert^2$	$O(\min\{n^2d^2, nd + d^4 logn\})$	$O(d^2)$	3.4.2
1-Segment	$P = \{(t_i \mid p_i)\}_{i=1}^n \subseteq \mathbb{R}^{d+1}$ $w : P \to [0, \infty)$	$\mathcal{X} = \{g \mid g : \mathbb{R} \to \mathbb{R}^d\}$	$f((t, p), g) = \lVert p - g(t)\rVert^2$	$\lVert\cdot\rVert_1$	$C \subseteq \mathbb{R}^{d+1}$ $\lvert C\rvert = d + 2$	$u \equiv 1$	$O(nd^2)$	$O(d^2)$	3.5
Matrix 2-norm (1)	$P \subseteq \mathbb{R}^d$ $w : P \to [0, \infty)$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = (p^T x)^2$	$\lVert\cdot\rVert_1$	$C \subseteq \mathbb{R}^d$ $\lvert C\rvert = d$	$u \equiv 1$	$O(nd^2)$	$O(d^2)$	3.6
Matrix 2-norm (2)	$P \subseteq \mathbb{R}^d$ $w : P \to [0, \infty)$	$\mathcal{X} = \mathbb{R}^d$	$f(p, x) = (p^T x)^2$	$\lVert\cdot\rVert_1$	$C \subseteq P$ $\lvert C\rvert \le d^2 + 1$	$u : C \to \left[0, \sum_{p\in P} w(p)\right]$ $\sum_{p\in C} u(p) = \sum_{p\in P} w(p)$	$O(\min\{n^2d^4, nd^2 + d^8 logn\})$	$O(d^3)$	3.6.1
Least Mean Squares	$P = \{(a_i^T \mid b_i)\}_{i=1}^n \subseteq \mathbb{R}^{d+1}$ $w : P \to [0, \infty)$	$\mathcal{X} = \mathbb{R}^d$	$f((a^T \mid b), x) = (a^T x - b)^2$	$\lVert\cdot\rVert_1$	$C \subseteq P$ $\lvert C\rvert \le (d + 1)^2 + 1$	$u : C \to \left[0, \sum_{p\in P} w(p)\right]$ $\sum_{p\in C} u(p) = \sum_{p\in P} w(p)$	$O(\min\{n^2d^4, nd^2 + d^8 logn\})$	$O(d^3)$	3.7

Figure: Table of settings and Accurate Coresets

Example 1: 1-Center

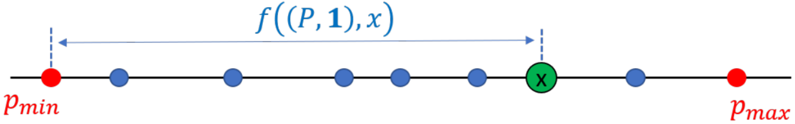
Problem

Given n points on the real line can we preprocess the points such that for any point $x \in \mathbb{R}$ we can find the distance from x to the point farthest away in P in $O(1)$ time?

Query Space

- $P = \{p_1, \ldots, p_n\} \in \mathbb{R}, w(p) = 1$
- $\mathcal{X} = \mathbb{R}$
- $f(p, x) = |p - x|$
- $\ell(\cdot) = \|\cdot\|_\infty$

Example 1: 1-Center



Solution

From the image

$$
\begin{aligned}
f_\ell(P', x) &= \|(|p_1 - x|, \ldots, |p_n - x|)\|_\infty \\
&= \max_{p \in P} |p - x| \\
&= \max_{p \in \{p_{min}, p_{max}\}} |p - x| \\
&= f_\ell(C', x)
\end{aligned}
$$

where $C = \{p_{min}, p_{max}\}, u(p) = 1$ is an accurate coreset for the query space.

Example 1: Generalisations

The previous accurate coreset can be generalised to when

- Unweighted P' is a line in $\mathbb{R}^d$
  - But not when P' is weighted
- Replacing $\mathcal{X}$ with all $g : \mathbb{R} \to [0, \infty)$ that are non-negative decreasing or increasing and increasing monotonic function after some point.

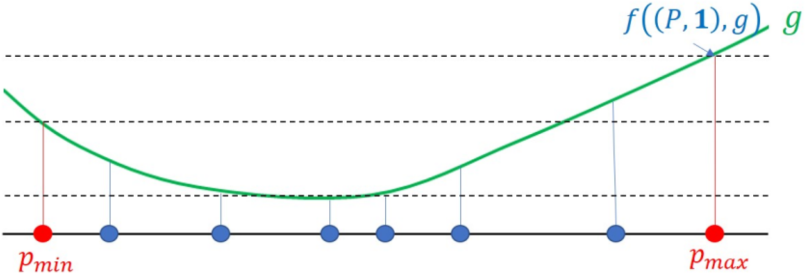


Figure: Non-negative decreasing g then increasing function

Example 2: Least Mean Squares Solvers

Problem

- $P = \{p_1, \dots, p_n\}$
  - $p_i = (a_i, b_i)^\mathsf{T}$, $a_i \in \mathbb{R}^d, b_i \in \mathbb{R}$ and $w : P \to [0, \infty)$.
  - Denote $w_i = w(p_i)$
- $\mathcal{X} = \mathbb{R}^d$
- $f((a, b)^\mathsf{T}) = (a^\mathsf{T} x - b)^2$
- $\ell(\cdot) = \|\cdot\|_1$

This leads to

$$f_\ell((P, w), x) = \sum_{i=1}^{n} w_i (a_i^\mathsf{T} x - b_i)^2$$

which is the weighted *least squares* objective function in statistics / ml.

Example 2: Least Mean Squares Solvers

Fact: Subset coreset of bounded weights

We will not prove, but use coresets for the matrix 2-norm. Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, where the i'th row of $\boldsymbol{A}$ is $\sqrt{w_i} p_i$. Then there exists a matrix $\boldsymbol{Z} \in \mathbb{R}^{d^2+1 \times d}$ such that he set of rows is a weighted subset of the rows of $\boldsymbol{A}$, for any $x \in \mathbb{R}^d$

$$\|\boldsymbol{A}x\|_2 = \|\boldsymbol{Z}x\|_2$$

and this corresponds to the query space $(P, w, \mathcal{X} = \mathbb{R}^d, f = \langle \cdot, \cdot \rangle^2, \|\cdot\|_1)$.

Example 2: Least Mean Squares Solvers

- Consider the original Least Squares query problem, but with $f(p, x) = (p^{\mathsf{T}} x)^2$.
- From previous slide there exist a coreset (C, u) of size $m = (d+1)^2 + 1$ for this query space where $C \subseteq P$ ($d+1$ since $p = (a, b)^{\mathsf{T}}$).

Example 2: Least Mean Squares Solvers

Call $C = \{(\hat{a}_1, \hat{b}_1)^\mathsf{T}, \ldots, (\hat{a}_m, \hat{b}_m)^\mathsf{T}\} = \{q_1, \ldots, q_m\}$ and $u_i = u(q_i)$. Since C is an accurate coreset, for any $x' \in \mathbb{R}^{d+1}$

$$\sum_{i=1}^{n} w_i (p_i^\mathsf{T} x')^2 = \sum_{j=1}^{m} u_i (q_i^\mathsf{T} x')^2$$

Example 2: Least Mean Squares Solvers

Choosing $x' = (x, -1)^\mathsf{T}$ we see that

$$\sum_{i=1}^{n} w_i (a_i^\mathsf{T} x - b)^2 = \sum_{i=1}^{n} w_i (p_i^\mathsf{T} x')^2$$
$$= \sum_{j=1}^{m} u_i (q_i^\mathsf{T} x')^2$$

which means that (C, u) *is an accurate coreset for the original least squares query space.*

Recent Work

Coresets (potentially approximate) have recently been applied in machine learning and statistics

- Logistic Regression [6]
- K-means [1]
- Kernel Density Estimation [7]
- Bayesian Inference [4, 3, 2]

References I

- Olivier Bachem, Mario Lucic, and Andreas Krause. "Scalable k-means clustering via lightweight coresets". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 1119–1127.
- Trevor Campbell and Tamara Broderick. "Automated Scalable Bayesian Inference Via Hilbert Coresets". In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 551–588.
- Trevor Campbell and Tamara Broderick. "Bayesian Coreset Construction Via Greedy Iterative Geodesic Ascent". In: *arXiv preprint arXiv:1802.01737* (2018).
- Jonathan Huggins, Trevor Campbell, and Tamara Broderick. "Coresets for scalable Bayesian logistic regression". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4080–4088.

References II

Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. "Introduction To Coresets: Accurate Coresets". In: *arXiv preprint arXiv:1910.08707* (2019).

Alexander Munteanu et al. "On coresets for logistic regression". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6561–6570.

Jeff M Phillips and Wai Ming Tai. "Near-Optimal Coresets of Kernel Density Estimates". In: *Discrete & Computational Geometry* (2019), pp. 1–21.