Coresets for the DP-Means Clustering Problem

Olivier Bachem
ETH Zurich

Mario Lučić
ETH Zurich

Andreas Krause
ETH Zurich

Abstract

We present coresets, a technique for approximately solving clustering problems on a small subset of the data, for the DP-Means clustering problem. DP-Means is a variant of K-Means where the number of clusters is not fixed but inferred from the data. We first show the existence of coresets of size $O(d^d k^* \log n/\epsilon^d)$ where k^* is the optimal number of centers and then propose a practical coreset construction algorithm that can be used to solve large instances of DP-Means clustering. We empirically demonstrate that coresets significantly outperform naive uniform subsampling and observe substantial speedups in runtime while achieving a low approximation error compared to solving the full instance.

1 Introduction

Scalable training of Bayesian nonparametric models is a notoriously difficult problem. One approach is to apply the technique of *small variance asymptotics* to the Gaussian Dirichlet Process mixture model [1]. This leads to a discrete optimization problem similar to K-Means clustering where the number of clusters is not fixed but inferred from the data. *DP-Means clustering* allows a solution to have an arbitrary number of clusters but imposes a penalty proportional to the number of clusters leading to a tradeoff between number of clusters used and quantization error achieved. As is the case with K-Means clustering, it is challenging to solve the DP-Means clustering problem for datasets where the number of samples is prohibitively large. The "naive" approach to this problem is to solve the clustering problem on a random subset of the data with the hope that the solution on this subset is close to the solution on the full dataset.

The concept of *coresets* originating from computational geometry offers a better solution. Coresets are weighted subsets of the data such that the quality of any clustering can be approximated on the coreset instead of the full dataset. This allows for fast approximate inference for large datasets by solving the clustering problem on the small coreset. Coreset constructions have been developed for a variety of unsupervised learning problems including K-Median/K-Means [2, 3, 4, 5, 6, 7, 8], LSA/PCA [8], K-Lines [9] and K-Segments [10].

Our contribution is the development of coresets for the DP-Means clustering problem. We first show the existence of coresets for DP-Means and then propose a practical coreset construction algorithm. In experiments, we finally verify its superior performance compared to naive subsampling and its speedup compared to solving DP-Means on the full dataset.

2 Coresets for the DP-Means clustering problem

2.1 DP-Means clustering problem

In *DP-Means clustering* an arbitrary number of cluster centers can be used; however, for each cluster center a penalty of λ is added to the objective function.

For $\lambda > 0$, a weighted set $\mathcal{P}$ of n points in $\mathbb{R}^d$ and a non-empty set of cluster centers $Q \subset \mathbb{R}^d$, the *DP-Means objective function*, also called the *DP-Means cost function*, is defined as

$$\text{cost}_{DP}(\mathcal{P}, Q) = \sum_{p \in \mathcal{P}} w_p \min_{q \in Q} \text{dist}(p, q)^2 + |Q|\lambda$$

The *DP-Means clustering problem* is to find a finite, non-empty set of cluster centers $Q \subset \mathbb{R}^d$ minimizing the DP-Means objective function.

2.2 Coreset definition

Similar to existing definitions for K-Means (e.g. [2]), we propose the following coreset definition for DP-Means:

Definition 2.1. Let $\epsilon > 0$ and $\mathcal{P}$ be a point set in $\mathbb{R}^d$. The weighted set $\mathcal{C}$ is an $(\epsilon, \bar{k})$-coreset for the DP-Means clustering of $\mathcal{P}$ if for any query, i.e. non-empty set Q, of at most $\bar{k}$ centers in $\mathbb{R}^d$

$$|\mathrm{cost}_{DP}(\mathcal{P}, Q) - \mathrm{cost}_{DP}(\mathcal{C}, Q)| \leq \epsilon \, \mathrm{cost}_{DP}(\mathcal{P}, Q)$$

If this property holds with $\bar{k} = \infty$, the weighted set $\mathcal{C}$ is called an ϵ-coreset.

2.3 Theoretical existence result using exponential grids

Our first result is the existence of ϵ-coresets for the DP-Means clustering problem that are sublinear in n if the optimal number of centers k^* is sublinear in n.

Theorem 2.2. Let $\epsilon > 0$ and let $\mathcal{P}$ be a set of n points in $\mathbb{R}^d$. Then there exists an ϵ-coreset for the DP-Means clustering of $\mathcal{P}$ with size $O\left(\frac{d^d k^* \log n}{\epsilon^d}\right)$ where k^* is the optimal number of centers.

Proof sketch. This result is obtained by applying the exponential grid approach of [2] to the DP-Means cost function. Assume that the optimal solution to the DP-Means clustering problem is known. One could then build an exponential grid around each of the cluster centers and project all data points in a grid cell to an arbitrary representative. It can then be shown that the number of grid cells is $O\left(\frac{d^d k^* \log n}{\epsilon^d}\right)$ and that the sum of the cost differences induced by the projection is bounded by $\epsilon \, \mathrm{cost}_{DP}(\mathcal{P}, Q)$ implying the required result.

2.4 Practical coresets using importance sampling

Our second contribution is a practical coreset construction for the DP-Means clustering problem that can be used to solve large problem instances. There are two differences to the theoretical setting in the last section: Firstly, we use a coreset construction technique based on importance sampling [7] that allows one to generate coresets of a specified size. This technique has been successfully applied in [11], both theoretically and empirically. Secondly, the existence result builds upon knowing the optimal solution which, in practice, is not the case. A common way of addressing this issue is finding a (rough) approximation of the optimal solution.

For a DP-Means problem instance defined by a set of points $\mathcal{P}$ in $\mathbb{R}^d$ and a hyperparameter $\lambda > 0$, we propose the following coreset construction:

Step 1 To find a (rough) approximation of the optimal solution we propose the algorithm *DP-Means++* (Algorithm 1) which is inspired by the seeding step of K-Means++ [12]. The difference is that the number of centers sampled using D^2-sampling is not fixed but inferred from the data using a stopping condition. Intuitively, this stopping condition manages the tradeoff between quantization error and penalization term which is the essential challenge of DP-Means clustering. Furthermore, Algorithm 1 not only returns a set A consisting of k' cluster centers but one obtains an upper bound $\bar{k} = k' \left(16(\log_2 k' + 2) + 1\right)$ on the optimal number of cluster centers k^*.

Step 2 We sample a $(\epsilon, \bar{k})$-coreset using the importance sampling scheme proposed in Algorithm 2. As the DP-Means cost function is a sum of (independent) cost contributions of all points, any importance sampling scheme produces an unbiased estimator. A key step to obtaining a coreset is bounding the variance of the sampling scheme [6]. This is achieved by sampling a point $p \in \mathcal{P}$ with probability proportional to its sensitivity $s(p)$ which is an upper bound to the maximum ratio between the individual cost contribution of p and the average cost contributions of all points [6]. We have derived the necessary bounds on the sensitivity and the required coreset size for the DP-Means clustering problem.

(Step 3) To approximately solve the full problem, any DP-Means solver can finally be applied to the $(\epsilon, \bar{k})$-coreset under the assumption that the solver respects the known upper bound $\bar{k}$, i.e. it is ensured that it only evaluates the DP-Means cost function for clusterings with less than $\bar{k}$ cluster centers. Both a brute-force approach based on solving K-Means for different values of k and the DP-Means algorithm [1] satisfy this requirement.

Algorithm 1 DP-Means++

Require: Set of data points $\mathcal{P}$, parameter λ
Uniformly sample $a \in \mathcal{P}$ and set $A = \{a\}$
while $\sum_{p \in \mathcal{P}} \operatorname{dist}(p, A)^2 > 16\lambda|A|(\log_2 |A| + 2)$ **do**
Sample point $a \in \mathcal{P}$ with probability $m(a) = \frac{\operatorname{dist}(a,C)^2}{\sum_{p' \in \mathcal{P}} \operatorname{dist}(p',C)^2}$ and add it to A
return approximate solution A of cardinality k'

Algorithm 2 Importance sampling scheme

Require: Set of data points $\mathcal{P}$, approximate solution A of cardinality k'
$\alpha \leftarrow 16(\log_2 k' + 2) + 2$
for $a \in A$ **do** $P_a \leftarrow$ points $p \in \mathcal{P}$ whose closest center in A is a
for $a \in A$ and $p \in P_a$ **do** $s(p) \leftarrow \frac{2\alpha \operatorname{dist}(p,A)^2}{\operatorname{cost}_{DP}(\mathcal{P},A)/|\mathcal{P}|} + \frac{4\alpha \sum_{p' \in P_a} \operatorname{dist}(p',A)^2}{|P_a| \operatorname{cost}_{DP}(\mathcal{P},A)/|\mathcal{P}|} + \frac{4|\mathcal{P}|}{|P_a|} + 1$
for $p \in \mathcal{P}$ **do** $q(p) \leftarrow \frac{s(p)}{\sum_{p' \in \mathcal{P}} s(p')}$
$m \leftarrow O\left(\frac{dk'^3 \log k'}{\epsilon^2}\right)$
$\mathcal{C} \leftarrow$ sample m weighted points from $\mathcal{P}$ where each point p has weight $\frac{1}{m \cdot q(p)}$ and is sampled with probability $q(p)$
return coreset $\mathcal{C}$

The main theoretical contribution of this section is that our method constructs valid $(\epsilon, \bar{k})$-coresets (Theorem 2.3). This implies that, given an optimal solver for the DP-Means problem, an arbitrarily small approximation error can be obtained by solving on the coreset (Corollary 2.4).

Theorem 2.3. Let $\epsilon > 0$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the results of Algorithm 1. Then with constant probability the weighted set $\mathcal{C}$ is an $(\epsilon, \hat{k})$-coreset with $\hat{k} = k' \left(16(\log_2 k' + 2) + 1\right)$ where k' is the number of centers returned by Algorithm 1.

Proof sketch. The proof builds upon Theorem 4.1 and Theorem 4.4 in [7]. Firstly, we bound the sensitivities $s(p)$ for the DP-Means cost function using the (rough) approximation of the optimal solution obtained in DP-Means++. This allows us to derive the sampling probabilities $q(p)$. Secondly, we show that the total sensitivity is upper bounded by $O(k')$ and that the dimension of the function space induced by the DP-Means cost function is upper bounded by $d(\bar{k} + 1)$ where $\bar{k}$ is the maximal number of cluster centers.

Corollary 2.4. Let $\epsilon > 0$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the results of Algorithm 1. For any optimal solver Q mapping a set of data points to a set of cluster centers, we have $\operatorname{cost}_{DP}(\mathcal{P}, Q(\mathcal{C})) \leq \frac{1+\epsilon}{1-\epsilon} \operatorname{cost}_{DP}(\mathcal{P}, Q(\mathcal{P}))$

The number of points m to be sampled in the second step depends on the size k' of the DP-Means++ solution. In Theorem 2.5 we bound the coreset size using the optimal number of cluster centers k^*. While there is an exponential dependency on d, the coreset size is sublinear in the number of data points n and only exhibits quadratic dependence on $1/\epsilon$. In practice, we are further able to observe k' allowing for data-dependent coreset sizes as in Algorithm 2.

Theorem 2.5. Let $\epsilon > 0$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the results of Algorithm 1. Then with constant probability the weighted set $\mathcal{C}$ has size at most $\tilde{O}\left(d^{3d^2+4} k^{*3} (\log n)^{3d/2+3} / \epsilon^2\right)$ where k^* is the optimal number of centers of the DP-Means clustering problem and the $\tilde{O}(\cdot)$ notation neglects any $\log k^*$ and $\log \log n$ terms.

Proof sketch. The proof relies on bounding k' based on k^*. We first show that the quantization error of the optimal K-Means clustering decays exponentially as the number of clusters is increased. This is achieved using an exponential grid argument similar to [2] albeit at the cost of introducing an exponential dependency on the dimension d. Then, using properties of the quantization error of the optimal K-Means solution for $k = k^*$ and $k = k'$ as well as results on D^2-sampling [13], we show that with constant probability k' is of $O(d^{d \cdot d} k^* [\log n]^{d/2+1})$.

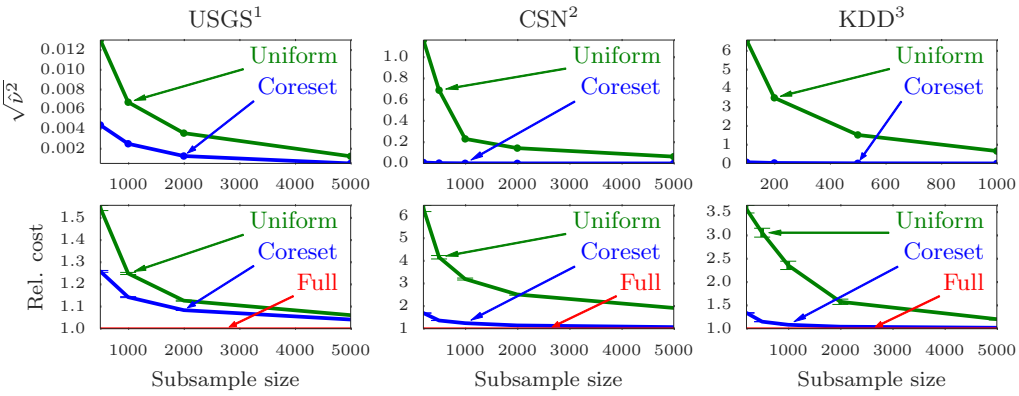

Figure 1: Experimental results for different datasets. The top row shows estimated variance for random query evaluations. The bottom row displays the relative cost of the solution obtained.

3 Experimental results

In the first set of experiments we compare the variance of the DP-Means cost estimate when approximated using coresets and uniform subsamples to see whether coresets exhibit a variance reducing property. We construct a random set of clusters Q by sampling uniformly from the original point set and then consider the relative error $\nu = (\text{cost}_{DP}(\mathcal{C}, Q) - \text{cost}_{DP}(\mathcal{P}, Q))/\text{cost}_{DP}(\mathcal{P}, Q)$ for the weighted subset $\mathcal{C}$. By repeating this 460 times we are able obtain an unbiased estimator $\hat{\nu}^2$ of $\mathbb{E}\left[\nu^2\right]$, both for coresets and uniform subsampling. The top row of Figures 1 shows the estimated variance for several datasets and subsample sizes. As expected, the variance for both coresets and the uniform subsampling decreases with increasing subsample sizes and, more importantly, coresets exhibit substantially lower variance than uniform subsampling.

Table 1: Runtime and performance comparison for KDD with subsample size $s = 5000$

	Sampling time	**Solving time**	**Total time**	**Speedup**	**Cost** (10^9)	**Rel. cost**
Uniform	0.0 s	9.9 s	9.9 s	43.8x	304.3	123.6%
Coreset	0.3 s	10.9 s	11.2 s	38.8x	251.9	102.3%
Full	-	434.3 s	434.3 s	1.0x	246.2	100.0%

In a second set of experiments we obtain a solution Q by solving DP-Means on coresets, uniform subsamples and the full dataset[4]. In the bottom row of Figure 1 we display the average DP-Means cost relative to the cost of the full solution. The average cost decreases with increasing subsample size and the coresets significantly outperform uniform subsampling. The results for the the KDD datasets and subsample size 5000 (see Table 1) illustrate the practical relevance of these results. Instead of using the full data, coresets with size 3.43% of the full data allow us to achieve a speedup of 38.8 times with only 2.3% additional error. Similarly naive uniform subsampling leads to an additional error of 23.6% with a runtime comparable to that of coresets. In fact, the runtime of the sampling step for coresets, i.e. DP-Means++, is negligible taking only 0.3 seconds or 2.5% of the total runtime of 11.2 seconds[5].

[1] **USGS** [14]: location of 59209 earthquakes between 1972 and 2010 mapped to 3D space using WGS 84
[2] **CSN** [15]: >7GB of cellphone accelerometer data processed into 80000 observations and 17 features
[3] **KDD** [16]: 145751 samples with 74 features measuring the match between a protein and a native sequence
[4] We use K-Means++ to solve K-Means for different values of k.
[5] All calculations were run on an Intel Xeon machine with 24 2.9GHz processors and 256GB RAM.

References

[1] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 513–520, 2012.

[2] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-Sixth Annual ACM symposium on Theory of Computing*, pages 291–300. ACM, 2004.

[3] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. In *Proceedings of the Twenty-First Annual Symposium on Computational Geometry*, pages 126–134. ACM, 2005.

[4] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry*, pages 11–18. ACM, 2007.

[5] Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[6] Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 598–607. Society for Industrial and Applied Mathematics, 2010.

[7] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the Forty-Third Annual ACM symposium on Theory of Computing*, pages 569–578. ACM, 2011.

[8] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.

[9] Dan Feldman, Amos Fiat, and Micha Sharir. Coresets for weighted facilities and their applications. In *47th Annual IEEE Symposium on Foundations of Computer Science*, pages 315–324. IEEE, 2006.

[10] Cynthia Sung, Dan Feldman, and Daniela Rus. Trajectory clustering for motion prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1547–1552. IEEE, 2012.

[11] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In *Advances in Neural Information Processing Systems*, pages 2142–2150, 2011.

[12] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[13] Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 15–28. Springer, 2009.

[14] United States Geological Survey. Global earthquakes (1.1.1972-19.3.2010). Retrieved from the mldata.org repository `https://mldata.org/repository/data/viewslug/global-earthquakes/`, 2010.

[15] Matthew Faulkner, Michael Olson, Rishi Chandy, Jonathan Krause, K Mani Chandy, and Andreas Krause. The next big one: Detecting earthquakes and other rare events from community-based sensors. In *10th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 13–24. IEEE, 2011.

[16] KDD Cup 2004. Protein Homology Dataset. Available at `http://osmot.cs.cornell.edu/kddcup/datasets.html`, 2004.