Bregman Power k-Means for Clustering Exponential Family Data

39 th International Conference on Machine Learning (ICML'22)

Adithya Vellal [1] Saptarshi Chakraborty [2] Jason Xu [1]

[1]Duke University [2]UC Berkeley

July 10, 2022

Partitional Clustering and k-means

- Data: n unlabeled observations. $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^p$.
- Goal: Find optimal partition $C = \{C_1, \ldots, C_k\}$ into k mutually exclusive and exhaustive groups.

Centroid-based clustering

Introduce cluster centroids, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\} \subset \mathbb{R}^p$.

k-means

Assign each observation to the cluster represented by the nearest center, minimizing within-cluster variance:

$$\min_C \sum_{j=1}^{k} \sum_{\boldsymbol{x}_i \in C_j} d(\boldsymbol{x}_i, \boldsymbol{\theta}_j)$$

Here $d(\cdot, \cdot)$ is a dissimilarity measure on $\mathbb{R}^p$.

Lloyd's algorithm

- Classically, $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

Greedy approach: seeks local minimizer of k-means objective, rewritten

$$\sum_{i=1}^{n} \min_{1 \leq j \leq k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2 := f_{-\infty}(\boldsymbol{\theta})$$

1. Update label assignments: $C_j^{(m)} = \{\boldsymbol{x}_i : \boldsymbol{\theta}_j^{(m)} \text{ is closest center}\}$
2. Recompute centers by averaging: $\boldsymbol{\theta}_j^{(m+1)} = \frac{1}{|C_j^{(m)}|} \sum_{\boldsymbol{x}_i \in C_j^{(m)}} \boldsymbol{x}_i$

Simple yet effective, remains most widely used clustering algorithm.

Bregman Hard Clustering

- Bregman divergence: $d_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.
- $\phi(\cdot)$ is convex, differentiable.
- Bregman hard clustering objective:

$$\min_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \min_{1 \le j \le k} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_j).$$

Mean as Minimizer

Let $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_{\ge 0}$ to be any continuous function with continuous first-order partial derivatives obeying $d(\boldsymbol{x}, \boldsymbol{x}) = 0$. Then the mean $\mathbb{E}[X]$ serves as the unique minimizer of $\mathbb{E}[d(X, \boldsymbol{y})]$ for $\boldsymbol{y} \in \mathbb{R}^p$ if and only if there exists some ϕ such that $d = d_\phi$.

Connection to Exponential Family

- The squared ℓ_2 distance is efficient if the clusters are normally distributed.
- Data generated from exponential family,

$$p(y|\theta, \tau) = C_1(y, \tau) \exp\left\{ \frac{y\theta - \phi^*(\theta)}{C_2(\tau)} \right\}.$$

- The negative log-likelihood of y can be written as its Bregman divergence to the mean:

$$-\ln p(y|\theta, \tau) = d_\phi\left(y, g^{-1}(\theta)\right) + C(y, \tau).$$

- In the context of clustering, they allow us to understand the analog of k-means minimizing the within-cluster variance in terms of the Bregman divergence based loss function.

Drawbacks of Lloyd's Algorithm

Too many local minimas!

- Sensitive to initialization, gets trapped in poor solutions, worsens in high dimensions.
- Objective is non-smooth, highly non-convex.
- Number of local minimas increase as the number of clusters (k) increases.

Power k-means with Bregman Divergence

The Proposed Objective function

$$f_s(\boldsymbol{\Theta}) = \sum_{i=1}^{n} M_s(d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_1), \ldots, d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_k)). \tag{1}$$

Here, $M_s(\boldsymbol{y}) = \left(\frac{1}{k} \sum_{i=1}^{k} y_i^s \right)^{1/s}$

Note that,

$$f_s(\boldsymbol{\Theta}) \downarrow f_{-\infty}(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Theta}} \sum_{i=1}^{n} \min_{1 \le j \le k} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_j)$$

- We implement a majorization-minimization (MM) to minimize the objective (1).
- The proposal runs with the same time complexity as Lloyd's k-means.
- As we take $s \to -\infty$, we get solutions to the Bregman hard clustering problem.

Theoretical Properties

- Model: $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} P$.

(Informal) Theorem 3.7

Assume that,

- $\|\boldsymbol{X}\|_2$ and $\phi(\boldsymbol{X})$ are sub-exponential.
- ϕ is τ_1 strongly convex and $\nabla\phi$ is τ_2-Lipschitz.

Then whenever $n \geq \log(2/\delta) \geq \frac{1}{2}$, with probability at least $1 - \delta - e^{-cn}$,

$$\text{Excess risk of } \widehat{\boldsymbol{\Theta}}_n \lesssim (\xi_P + \|\boldsymbol{\Theta}_*\|_F)\frac{k^{3/2-1/s}p}{\sqrt{n}} + k^{1-1/s}(1 + \xi_P + \|\boldsymbol{\Theta}_*\|_F)\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Significance of the Theoretical Analysis

- Unlike Paul et al. (2021, NeurIPS) we relax the bounded support assumption of P.
- The bound on the excess risk is in terms of the size of $\mathbf{\Theta}_*$.
- Matches with existing literature.
- We can recover strong consistency guarantees and $\sqrt{n}$-consistency of $\widehat{\mathbf{\Theta}}_n$.

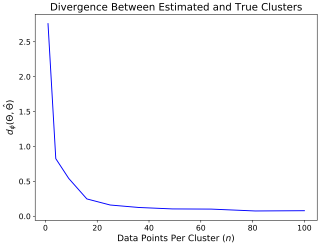

Figure: We see that the empirical convergence of Bregman power k-means to the true cluster centroids agrees with the $\mathcal{O}_P(n^{-1/2})$ convergence proposed in Theorem 3.8.

Experimental Results

	Lloyd's	Bregman Hard	Power	Bregman Power
Gaussian	0.828 (0.012)	0.837 (0.012)	0.927 (0.003)	0.927 (0.003)
Binomial	0.730 (0.014)	0.886 (0.011)	0.915 (0.004)	0.931 (0.003)
Poisson	0.723 (0.014)	0.882 (0.010)	0.888 (0.006)	0.916 (0.004)
Gamma	0.484 (0.009)	0.868 (0.005)	0.677 (0.008)	0.879 (0.004)

Table: Results for experiment 1. Mean and (standard deviation) ARI of Lloyd's algorithm, Bregman hard clustering, and their power means counterparts.

Experiment 2

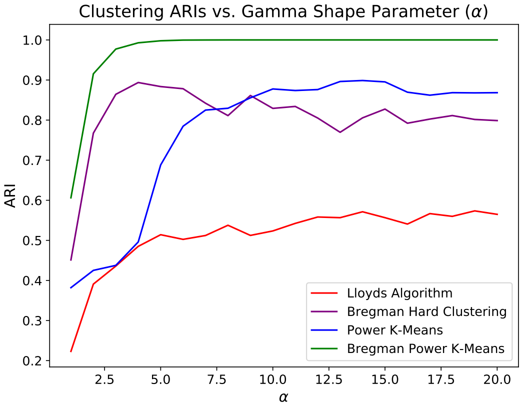

Figure: Performance as Gamma shape parameter varies.

Thank You!

https://arxiv.org/abs/2012.10929