Notes on k-Means

Karl Stratos

1 Problem

Let $\Omega \subseteq \mathbb{R}^d$ be a closed convex set. Given points $\mathcal{X} \subset \Omega$ and an integer k, we aim to compute k "centers" $\mathcal{M} \subset \Omega$ that minimize

$$\sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}} D(x, \mu) \tag{1}$$

where $D : \Omega \times \Omega \to [0, \infty)$ is a (not necessarily symmetric) distortion function. Appendix A gives examples of D. Dasgupta (2008) shows that minimizing (1) over $\mathcal{M}$ is NP-hard for $D(x, y) := ||x - y||_2^2$ and $k = 2$.

2 Algorithm

A heuristic for minimizing (1) can be derived as follows. Define a k-partition $\{C^\mu\}_{\mu \in \mathcal{M}}$ of $\mathcal{X}$ associated with centers $\mathcal{M}$ where

$$C^\mu := \left\{ x \in \mathcal{X} : \ \mu = \arg\min_{\mu' \in \mathcal{M}} D(x, \mu') \right\} \tag{2}$$

We assume that a tie $D(x, \mu) = D(x, \mu')$ is broken arbitrarily.

Proposition 2.1. *For each center $\mu \in \mathcal{M}$, let*

$$\nu^\mu := \arg\min_{\nu \in \Omega} \sum_{x \in C^\mu} D(x, \nu) \tag{3}$$

and let $\mathcal{N} = \{\nu^\mu : \ \mu \in \mathcal{M}\}$. Then

$$\sum_{x \in \mathcal{X}} \min_{\nu \in \mathcal{N}} D(x, \nu) \leq \sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}} D(x, \mu)$$

A proof is given in Appendix B. Thus given some initial k centers, we can repeatedly compute a k-partition of $\mathcal{X}$ based on (2) and new centers based on (3) to monotonically improve (1) until local convergence.

The resulting algorithm, shown in Figure 1, is often called **k-means** because for a wide class of distortion functions called Bregman divergences, ν^μ in (3) for cluster C^μ is simply the mean of C^μ (see Proposition A.1). Thus we calculate "k means" of $\{C^\mu\}_{\mu \in \mathcal{M}}$ in each iteration.

The runtime of the algorithm is $O(T |\mathcal{X}| kd)$, but note that we can easily parallelize cluster assignment and center computation to reduce the runtime to $O(T |\mathcal{X}| kd/\tau)$ where τ is the number of threads. In practice, we also need to handle an issue with empty clusters (see Appendix C).

k-means
Input: points $\mathcal{X} \subset \Omega$, initial k centers $\mathcal{M}^{(0)} \subset \Omega$, distortion $D : \Omega \times \Omega \to [0, \infty)$, number of iterations T

1. For $t = 1 \dots T$,
   (a) Calculate a k-partition $\{C^\mu\}_{\mu \in \mathcal{M}^{(t-1)}}$ of $\mathcal{X}$ where

$$C^\mu \leftarrow \left\{ x \in \mathcal{X} : \ \mu = \operatorname*{arg\,min}_{\mu' \in \mathcal{M}^{(t-1)}} D\left(x, \mu'\right) \right\}$$

   (b) Calculate new k centers $\mathcal{M}^{(t)} = \left\{ \nu^\mu : \ \mu \in \mathcal{M}^{(t-1)} \right\}$ where

$$\nu^\mu \leftarrow \operatorname*{arg\,min}_{\nu \in \Omega} \sum_{x \in \mathcal{C}^\mu} D\left(x, \nu\right)$$

Output: $\mathcal{M}^{(T)} \subset \Omega$ such that

$$\sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}^{(T)}} D\left(x, \mu\right) \leq \sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}^{(0)}} D\left(x, \mu\right)$$

Figure 1: The k-means clustering algorithm.

3 Guarantees

In this section, for clarity we only consider the squared Euclidean distance $D(x, y) := ||x - y||_2^2$ (with domain $\Omega = \mathbb{R}^d$) and use the following notation with respect to fixed $\mathcal{X} \subset \mathbb{R}^d$ and k. Denote the cost of proposed k centers $\mathcal{M} \subset \mathbb{R}^d$ by

$$\operatorname{cost}\left(\mathcal{M}\right) := \sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}} \ ||x - \mu||_2^2$$

and denote the optimal k centers by $\mathcal{M}^* := \operatorname{arg\,min}_{\mathcal{M} \subset \mathbb{R}^d : \ |\mathcal{M}| = k} \ \operatorname{cost}\left(M\right)$. The bad news is that k-means has no guarantee on the optimality of its output.

Proposition 3.1. *Let B be any constant. Then we can construct $\mathcal{X}$ and $\mathcal{M}^{(0)}$ such that no matter how large T is,*

$$\operatorname{cost}\left(M^{(T)}\right) \geq B \operatorname{cost}\left(M^*\right)$$

*where $\mathcal{M}^{(T)}$ is the output of k-**means**$(\mathcal{X}, \mathcal{M}^{(0)}, ||x - y||_2^2, T)$.*

A construction proving Proposition 3.1 is well-known and thus omitted. The good news is that it is possible to combat degenerate cases by randomizing the choice of initial centers. Arthur and Vassilvitskii (2007) propose a good randomized strategy called "k-means++" which is given in Figure 2. They show that k-means++ produces centers that are at most a factor of $\log k$ worse than the optimal centers in expectation!

Theorem 3.1 (Arthur and Vassilvitskii, 2007)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be any points. If $\mathcal{M}^+$ is the output of k-**means++**$(\mathcal{X})$, then*

$$\boldsymbol{E}\left[\operatorname{cost}\left(\mathcal{M}^+\right)\right] \leq O(\log k) \operatorname{cost}\left(\mathcal{M}^*\right)$$

where the expectation is with respect to the randomness of k-means++.

k-means++
Input: points $\mathcal{X} \subset \mathbb{R}^d$, number of centers k

1. Draw $\mu_1 \sim \mathcal{X}$ uniformly at random, and let $\mathcal{M}^+ \leftarrow \{\mu_1\}$.
2. For $i = 2 \dots k$,
   (a) Draw $\mu_i \sim \mathcal{X}$ with probability

$$\frac{\min_{\mu \in \mathcal{M}^+} ||x - \mu||_2^2}{\sum_{x' \in \mathcal{X}} \min_{\mu \in \mathcal{M}^+} ||x' - \mu||_2^2} \qquad \forall x \in \mathcal{X}$$

   and let $\mathcal{M}^+ \leftarrow \mathcal{M}^+ \cup \{\mu_i\}$.

Output: $\mathcal{M}^+ \subset \mathbb{R}^d$ such that

$$\boldsymbol{E}\left[\sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}^+} ||x - \mu||_2^2\right] \leq O(\log k) \min_{\substack{\mathcal{M} \subset \mathbb{R}^d: \\ |\mathcal{M}| = k}} \left(\sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}} ||x - \mu||_2^2\right)$$

Figure 2: The k-means++ algorithm.

A key part of the proof is that when a center is randomly selected *from the points $\mathcal{X}$ themselves*, it is worse than the optimal center only by a constant factor. A useful tool for showing this is the bias-variance decomposition of the expected squared error: for any constant $x \in \mathbb{R}^d$ and random variable $Z \in \mathbb{R}^d$,

$$\underbrace{\boldsymbol{E}\left[||x - Z||_2^2\right]}_{\text{squared error of } x} = \underbrace{||x - \boldsymbol{E}[Z]||_2^2}_{\text{bias of } x} + \underbrace{\boldsymbol{E}\left[||Z - \boldsymbol{E}[Z]||_2^2\right]}_{\text{variance of } Z}$$

The result is easy to show for a single cluster.

Lemma 3.2. *Let $C \subset \mathbb{R}^d$ be a nonempty set. If Z is drawn from C uniformly at random, then*

$$\boldsymbol{E}\left[\sum_{x \in C} ||x - Z||_2^2\right] = 2 \min_{z \in C} \sum_{x \in C} ||x - z||_2^2$$

Proof. The minimizer is given by the mean $z^* = (1/|C|) \sum_{x \in C} x = \boldsymbol{E}[Z]$, and

$$\begin{aligned}
\boldsymbol{E}\left[\sum_{x \in C} ||x - Z||_2^2\right] &= \sum_{x \in C} ||x - \boldsymbol{E}[Z]||_2^2 + |C| \, \boldsymbol{E}\left[||Z - \boldsymbol{E}[Z]||_2^2\right] \\
&= \sum_{x \in C} ||x - \boldsymbol{E}[Z]||_2^2 + \sum_{x \in C} ||x - \boldsymbol{E}[Z]||_2^2 \\
&= 2 \sum_{x \in C} ||x - z^*||_2^2
\end{aligned}$$

□

Lemma 3.2 applies immediately to the first center μ_1 selected by k-means++. Let $z_1 \in \mathcal{M}^*$ denote the mean of the cluster that μ_1 belongs to. Then since μ_1 is a uniformly random draw from that cluster, μ_1 is worse than z_1 only by a factor of 2 in expectation.

Here is a sketch of the proof. We can decompose the expected cost $\boldsymbol{E}\left[\text{cost}\left(\mathcal{M}^{+}\right)\right]$ of the centers selected by k-means++ into a sum of k components corresponding to the $t = 1 \ldots k$ iterations of the algorithm. At t-th component, we have a term that is a constant multiple of the optimal value associated with $\mathcal{M}^{*}$ (e.g., as in Lemma 3.2), plus a term that accounts for the suboptimality of the $1 \ldots t-1$ previous centers. This expression ends up taking the following form:

$$\begin{aligned}\boldsymbol{E}\left[\text{cost}\left(\mathcal{M}^{+}\right)\right] &\leq 8\,\text{cost}\left(\mathcal{M}^{*}\right)\left(1+1+\frac{1}{2}+\ldots+\frac{1}{k}\right)\\ &\leq 8\,\text{cost}\left(\mathcal{M}^{*}\right)\left(2+\log k\right)\end{aligned}$$

where we used the upper bound $1+\log k$ on the harmonic sum $1+(1/2)+\cdots+(1/k)$.

References

Arthur, D. and Vassilvitskii, S. (2007). 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.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, **6**(Oct), 1705–1749.

Dasgupta, S. (2008). *The hardness of k-means clustering*. Department of Computer Science and Engineering, University of California, San Diego.

A Choices of Distortion

We describe some well-known choices of the distortion function D underlying k-means.

A.1 Bregman Divergence

Given a closed convex set Ω, let $F : \Omega \to \mathbb{R}$ be a smooth and strictly convex function. The **Bregman divergence** $D_F : \Omega \times \Omega \to [0, \infty)$ associated with F is defined as

$$D_F(x, y) := F(x) - F(y) - \langle \nabla F(y), x - y \rangle \qquad \forall x, y \in \Omega$$

That is, it is the error of the first-order Taylor approximation of $F(x)$ at y. While it is not a metric (e.g., it does not satisfy the triangle inequality or symmetry), it has certain desirable properties including:

- $D_F(x, y) \geq 0$ for all $x, y \in \Omega$, with equality if and only if $x = y$. This follows because F is strictly convex.
- $D_F(x, y)$ is strictly convex in $x \in \Omega$ for any fixed $y \in \Omega$.

But the most useful property for k-means is that the solution of (3) is given by the mean for any choice of F.

Proposition A.1 (Banerjee *et al.*, 2005)**.** *Let $C \subset \Omega$ be a nonempty set with the mean $\mu^C := (1/|C|) \sum_{x \in C} x$. If $D_F : \Omega \times \Omega \to [0, \infty)$ is a Bregman divergence, then*

$$\mu^C = \arg\min_{\mu \in \Omega} \sum_{x \in C} D_F(x, \mu)$$

Proof. We have $\mu^C \in \Omega$ since Ω is closed and convex. Pick any $\mu \in \Omega$ and note that

$$\begin{aligned}
&\sum_{x \in C} D_F(x, \mu) - \sum_{x \in C} D_F\left(x, \mu^C\right) \\
&= \sum_{x \in C} F\left(\mu^C\right) - F(\mu) - \langle \nabla F(\mu), x - \mu \rangle + \langle \nabla F\left(\mu^C\right), x - \mu^C \rangle \\
&= |C| F\left(\mu^C\right) - |C| F(\mu) - |C| \langle \nabla F(\mu), \mu^C - \mu \rangle \\
&= |C| D_F\left(\mu^C, \mu\right) \geq 0
\end{aligned}$$

with equality if and only if $\mu = \mu^C$. □

Here are some choices of Ω and $F : \Omega \to \mathbb{R}$ that induce popular Bregman divergences.

Example A.1 (Squared Euclidean distance)**.** *Let $\Omega = \mathbb{R}^d$ and $F(x) := ||x||_2^2$. Then for all $x, y \in \Omega$,*

$$\begin{aligned}
D_F(x, y) &= ||x||_2^2 - ||y||_2^2 - 2\langle y, x - y \rangle \\
&= ||x||_2^2 + ||y||_2^2 - 2\langle x, y \rangle = ||x - y||_2^2
\end{aligned}$$

Example A.2 (KL divergence)**.** *Let $\Omega = \Delta^{d-1}$ and $F(p) := \sum_{i=1}^d p_i \log p_i$ (i.e., the negative entropy of a random variable X with $p_i = P(X = i)$). Then for all $p, q \in \Omega$,*

$$\begin{aligned}
D_F(p, q) &= \sum_{i=1}^d p_i \log p_i - \sum_{i=1}^d q_i \log q_i - \sum_{i=1}^d (1 + \log q_i)(p_i - q_i) \\
&= \sum_{i=1}^d p_i (\log p_i - \log q_i) - \left(\sum_{i=1}^d p_i - \sum_{i=1}^d q_i \right) = D_{KL}(p||q)
\end{aligned}$$

A.2 Non-Bregman Distortion

We can consider distortion functions that are not Bregman divergences such as the Manhattan and Euclidean distances on $\mathbb{R}^d$. The **Manhattan distance** is the difference in l_1 norm and has a closed-form solution for (3).

Proposition A.2. *Let $C \subset \mathbb{R}^d$ be a nonempty set. Let $\delta^C \in \mathbb{R}^d$ denote a vector such that δ_i^C is the median of $\{x_i : x \in C\}$. Then*

$$\delta^C = \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \sum_{x \in C} ||x - y||_1$$

We omit the proof, but the intuition is that the stationary condition of the objective

$$\sum_{x \in C:\ y_i \geq x_i} 1 = \sum_{x \in C:\ y_i < x_i} 1 \qquad \forall i = 1 \ldots d$$

is satisfied by taking the median. The **Euclidean distance** is the (non-squared) difference in l_2 norm. The minimizer of distortion for a nonempty $C \subset \mathbb{R}^d$,

$$\gamma^C := \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \sum_{x \in C} ||x - y||_2$$

is called the **geometric median** of C. There is no closed-form solution for γ^C, but an iterative algorithm such as Weiszfeld's algorithm can be used to optimize the objective. Since the objective is strictly convex, there is no issue of local optimum.

B Proof of Proposition 2.1

Proposition For each center $\mu \in \mathcal{M}$, let

$$C^\mu := \left\{ x \in \mathcal{X} :\ \mu = \operatorname*{arg\,min}_{\mu' \in \mathcal{M}} D(x, \mu') \right\}$$
$$\nu^\mu := \operatorname*{arg\,min}_{\nu \in \Omega} \sum_{x \in \mathcal{C}^\mu} D(x, \nu)$$

and let $\mathcal{N} = \{\nu^\mu :\ \mu \in \mathcal{M}\}$. Then

$$\sum_{x \in \mathcal{X}} \min_{\nu \in \mathcal{N}} D(x, \nu) \leq \sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}} D(x, \mu)$$

Proof. In the following, we write

$$C^\nu := \left\{ x \in \mathcal{X} :\ \nu = \operatorname*{arg\,min}_{\nu' \in \mathcal{N}} D(x, \nu') \right\}$$

for each $\nu \in \mathcal{N}$. Then

$$\begin{aligned}
\sum_{x \in \mathcal{X}} \min_{\mu \in \mathcal{M}} D(x, \mu) &= \sum_{\mu \in \mathcal{M}} \sum_{x \in C^\mu} D(x, \mu) \\
&\geq \sum_{\mu \in \mathcal{M}} \sum_{x \in C^\mu} D(x, \nu^\mu) && \text{(by definition)} \\
&\geq \sum_{\mu \in \mathcal{M}} \sum_{x \in C^{\nu^\mu}} D(x, \nu^\mu) = \sum_{x \in \mathcal{X}} \min_{\nu \in \mathcal{N}} D(x, \nu)
\end{aligned}$$

□

C Empty Clusters

In order to compute (3), we need C^μ to be nonempty. For instance, under a Bregman divergence D we must compute

$$\nu^\mu = \frac{1}{|C^\mu|} \sum_{x \in C^\mu} x$$

where an empty C^μ causes division by zero! Unfortunately, empty clusters can be created during the algorithm, especially if initial centers are bad: see `http://www.ceng.metu.edu.tr/~tcan/ceng465_f1314/Schedule/KMeansEmpty.html`. Some ways to handle this problem in practice are:

- When a center with an empty cluster is created, replace it with a random point.
- Restart the algorithm with a different choice of centers.