k-means++ under Approximation Stability

Manu Agarwal, Ragesh Jaiswal and Arindam Pal

Arindam Pal

arindamp@cse.iitd.ac.in

TCS Innovation Labs Kolkata

Department of Computer Science, IIT Delhi

May 20, 2013

TAMC 2013, University of Hong Kong

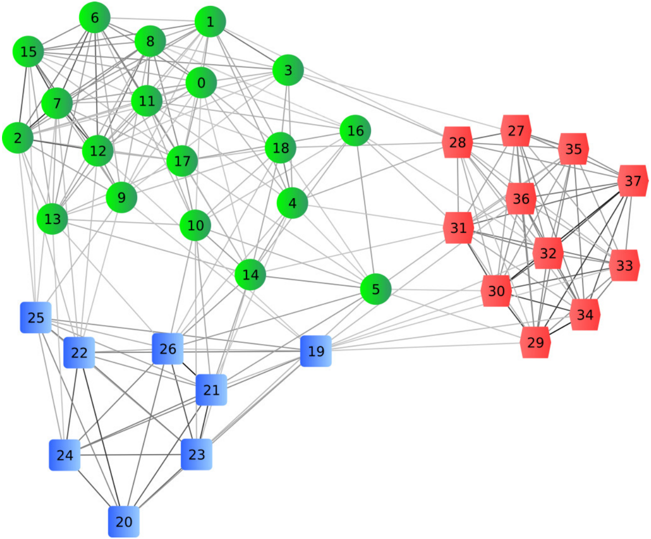
Agenda

- The clustering problem and k-means clustering
- Llyod's algorithm and k-means++
- Approximation stability and distance between clusterings
- Our contributions
- Analysis of k-means++
- Conclusion and future work

The clustering problem

- Given a set of data points, we need to group them together so that *similar* points are in the same group and *dissimilar* points are in different groups.
- Typically, these points live on a *metric space*.
- These groups are called *clusters*.
- There is an *objective function* which has to be optimized.

Example of a clustering



k-means clustering problem

- Suppose we have a k-clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$.
- The point c_i is the center of cluster C_i.
- A point x is assigned to cluster C_i if $d(x, c_i) \leq d(x, c_j)$ for any $j \neq i$.
- For the k-means clustering, we have to minimize the following objective function.

$$\Phi(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2.$$

Llyod's algorithm

1. Choose k initial centers $\mathcal{C} = \{c_1, \ldots, c_k\}$ arbitrarily.
2. For each $i \in \{1, \ldots, k\}$, set the cluster C_i to be the set of points in $\mathcal{X}$ that are closer to c_i than to c_j for any $j \neq i$.
3. For each $i \in \{1, \ldots, k\}$, set c_i to be the centroid of all points in C_i.

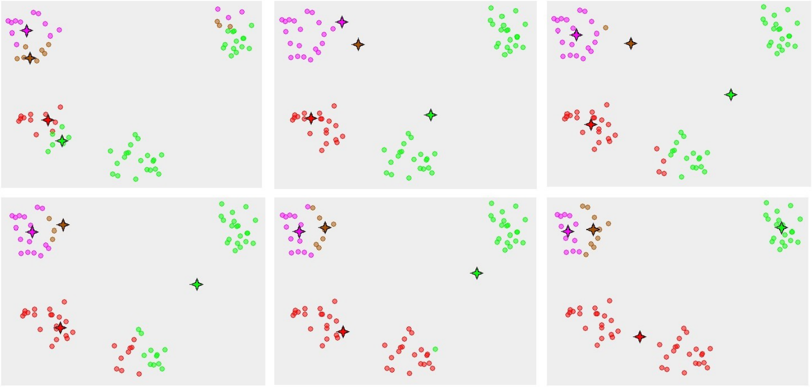
$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x.$$

4. Repeat Steps 2 and 3 until $\mathcal{C}$ does not change.

Problems with Llyod's algorithm

- Since it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum.
- The result depends on the initial clusters.
- There exist certain point sets (even on the plane), on which the algorithm takes exponential time ($2^{\Omega(n)}$) to converge.
- However, the smoothed running time of k-means is polynomial.
- k-means assumes that the clusters are spherical that are separable in a way so that the mean value converges towards the cluster center.

k-means convergence to a local minimum



k-means++: Initialization of cluster centers

1. Choose the first center c_1 uniformly at random from $\mathcal{X}$.
2. Choose the next center c_i with probability $\frac{D(c_i)^2}{\sum_{x \in S} D(x)^2}$.
3. Here $D(x)$ is the shortest distance from a point x to the closest center we have already chosen.
4. Repeat Step 2, until k centers are chosen.

Performance of k-means++

- k-means++ is $O(\log k)$-competitive in expectation.
- There are examples on which k-means++ is $\Omega(\log k)$-competitive in expectation.
- So, this is a tight analysis.
- Can k-means++ do better if the data has additional properties?

Distance between two clusterings

- Suppose we have two k-clusterings $\mathcal{C} = \{C_1, \ldots, C_k\}$ and $\mathcal{C}' = \{C'_1, \ldots, C'_k\}$ of a point set $\mathcal{X}$.
- Distance between $\mathcal{C}$ and $\mathcal{C}'$ is the fraction of points on which they disagree under the optimal matching of clusters in $\mathcal{C}$ to clusters in $\mathcal{C}'$.
- Formally,

$$dist(\mathcal{C}, \mathcal{C}') = \min_{\sigma \in \mathcal{S}_k} \frac{1}{n} \sum_{i=1}^{k} |C_i \setminus C'_{\sigma(i)}|,$$

  where $\mathcal{S}_k$ is the set of all permutations $\sigma : \{1, \ldots, k\} \mapsto \{1, \ldots, k\}$.
- Two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are ϵ-close if $dist(\mathcal{C}, \mathcal{C}') < \epsilon$.

Approximation stability

- Suppose we are given an objective function Φ such as k-means or k-median.
- The point set $\mathcal{X}$ satisfies (c, ϵ)-approximation stability if all clusterings $\mathcal{C}$ with $\Phi(C) \leq c \cdot \Phi_{OPT}$ are ϵ-close to the target clustering $\mathcal{C}_T$.
- At most ϵ fraction of points have to be reassigned in $\mathcal{C}$ to match $\mathcal{C}_T$.
- We can assume w.l.o.g that $\mathcal{C}_T$ is the optimal clustering.

Our results for large clusters

- Let $0 < \epsilon, \alpha \leq 1$. If a dataset satisfies $(1 + \alpha, \epsilon)$-approximation stability and each optimal cluster has size at least $\frac{60\epsilon n}{\alpha^2}$, then the k-means++ algorithm gives an 8-approximation to the k-means objective with probability $\Omega(\frac{1}{k})$.
- Let $0 < \epsilon \leq 1$ and $\alpha > 1$. If a dataset satisfies $(1 + \alpha, \epsilon)$-approximation stability and each optimal cluster has size at least $70\epsilon n$, then the k-means++ algorithm gives an 8-approximation to the k-means objective with probability $\Omega(\frac{1}{k})$.
- We also generalize these results for k-medians with respect to distance measures that satisfy approximate symmetry and approximate triangle inequality.

Lower bound example for small clusters

- We show that there exists a dataset $\mathcal{X} \in \mathbb{R}^d$ such that the following holds:
  - $\mathcal{X}$ satisfies the $(1+\alpha, \epsilon)$ approximation stability property.
  - k-means++ achieves an approximation factor of $\frac{1}{2}\log k$ with probability at most $e^{-\sqrt{k}-o(1)}$.

An important result [BBG09]

- Let $C_1^*, ..., C_k^*$ denote the optimal k clusters with respect to the k-means objective function and let $c_1^*, ..., c_k^*$ denote the centroids of these optimal clusters.
- For a point $x \in \mathcal{X}$, let $w(x)$ be its distance from the closest center and $w_2(x)$ be its distance from the second closest center.
- Suppose OPT is the cost of the optimal clustering.
- If the dataset satisfies $(1+\alpha, \epsilon)$-approximation-stability for the k-means objective, then
  1. If $\forall i, |C_i^*| \geq 2\epsilon n$, then less than ϵn points have $w_2^2(x) - w^2(x) \leq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n}$.
  2. For any $t > 0$, at most $t\epsilon n$ points have $w^2(x) \geq \frac{\mathrm{OPT}}{t\epsilon n}$.

Preliminaries

- Let $c_1, ..., c_i$ be the centers chosen by the first i iterations of k-means++.
- Suppose $j_1, ..., j_i$ are the indices of the optimal clusters to which these centers belong.
- Define $J_i = \{j_1, \ldots, j_i\}$ and $\bar{J}_i = \{1, ..., k\} \setminus J_i$.
- J_i is the set of indices of the clusters that are covered at the end of the i^{th} iteration.

- Let B_1 be the subset of points in $\bar{\mathcal{X}}_i$ such that for any point $x \in B_1$, $w_2^2(x) - w^2(x) \leq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n}$.
- Let B_2 denote the subset of points in $\bar{\mathcal{X}}_i$ such that for every point $x \in B_2$, $w^2(x) \geq \frac{\alpha^2 \cdot \mathrm{OPT}}{6\epsilon n}$.
- We know that $|B_1| \leq \epsilon n$ and $|B_2| \leq \frac{6\epsilon n}{\alpha^2}$.
- Let $B = B_1 \cup B_2$ and $\bar{B} = \bar{\mathcal{X}}_i \setminus B$.
- We know that $|B| \leq \frac{7\epsilon n}{\alpha^2}$.

A key lemma

Lemma

Let $\beta = \frac{1-\frac{\alpha}{2}}{6+\alpha}$. *For any* $x \in \bar{B}$ *we have,* $D^2(x, c_t) \geq \beta \cdot D^2(x, c^*_{j_t})$.

- Proof: Let j be the index of the optimal cluster to which x belongs.
- Note that $w^2(x) = D^2(x, c^*_j)$ and $w_2^2(x) \leq D^2(x, c^*_{j_t})$.
- For any $x \in \bar{B}$, we have:

$$w_2^2(x) - w^2(x) \geq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n} \geq \frac{6w^2(x)}{\alpha}$$

$$\Rightarrow \quad w_2^2(x) \geq \left(1 + \frac{6}{\alpha}\right) \cdot w^2(x) \qquad (1)$$

- Suppose that $D^2(x, c_t) < \beta \cdot D^2(x, c^*_{j_t})$.

- Then we get the following inequalities.

$$
\begin{aligned}
& 2 \cdot D^2(x, c_j^*) + 2 \cdot D^2(x, c_t) \geq D^2(c_t, c_j^*) \quad (\Delta \text{ inequality}) \\
\Rightarrow\ & 2 \cdot D^2(x, c_j^*) + 2 \cdot D^2(x, c_t) \geq D^2(c_t, c_{j_t}^*) \quad (D^2(c_t, c_j^*) \geq D^2(c_t, c_{j_t}^*)) \\
\Rightarrow\ & 2 \cdot D^2(x, c_j^*) + 2 \cdot D^2(x, c_t) \geq \frac{1}{2} \cdot D^2(x, c_{j_t}^*) - D^2(x, c_t) \\
\Rightarrow\ & 3 \cdot D^2(x, c_t) \geq \frac{1}{2} \cdot D^2(x, c_{j_t}^*) - 2 \cdot D^2(x, c_j^*) \\
\Rightarrow\ & 3\beta \cdot D^2(x, c_{j_t}^*) > \frac{1}{2} \cdot D^2(x, c_{j_t}^*) - 2 \cdot D^2(x, c_j^*) \\
& (\text{using assumption } D^2(x, c_t) < \beta \cdot D^2(x, c_{j_t}^*)) \\
\Rightarrow\ & D^2(x, c_j^*) > \frac{1 - 6\beta}{4} \cdot D^2(x, c_{j_t}^*) \\
\Rightarrow\ & w^2(x) > \frac{1}{1 + \frac{6}{\alpha}} \cdot w_2^2(x) \quad (D^2(x, c_{j_t}^*) \geq w_2^2(x) \text{ and } \beta = \frac{1 - \frac{\alpha}{2}}{6 + \alpha})
\end{aligned}
$$

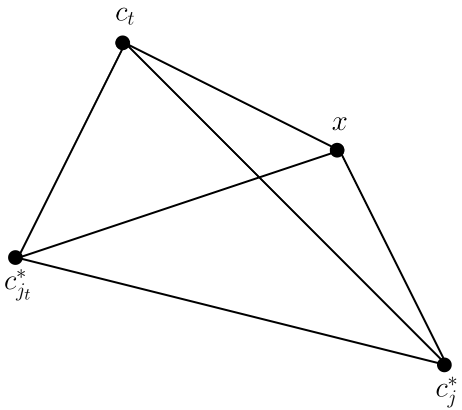


Figure: x belongs to the uncovered cluster j.

- This contradicts with Equation (1). Hence, for any $x \in \bar{B}$ and any $t \in \{1, ..., i\}$, we have $D^2(x, c_t) \geq \beta \cdot D^2(x, c^*_{j_t})$.

- Let $W_{min} = \min_{t \in [k]} \left(\sum_{x \in C_t^*, x \in \bar{B}} w_2^2(x) \right)$.
- Let C_i denote the set of centers $\{c_1, ..., c_i\}$ that are chosen in the first i iterations of k-means++.
- Let $\mathcal{X}_i = \cup_{t \in J_i} C_t^*$ and $\bar{\mathcal{X}_i} = \mathcal{X} \setminus \mathcal{X}_i$.
- $\mathcal{X}_i$ denotes the points that are covered by the algorithm after step i.
- For any subset of points $Y \subseteq \mathcal{X}$, $\phi_{C_i}(Y)$ is the cost of the points in Y with respect to the centers C_i, i.e., $\phi_{C_i}(Y) = \sum_{x \in Y} \min_{c \in C_i} D^2(x, c)$.
- We have $\phi_{\{c_1, \ldots, c_i\}}(\bar{\mathcal{X}_i}) \geq \beta \cdot (k - i) \cdot W_{min}$.

- Let E_i denote the event that the set J_i contains i distinct indices from $\{1, ..., k\}$.
- This means that the first i sampled centers cover i optimal clusters.
- The next Lemma is from [AV07] and shows that given that event E_i happens, the expected cost of points in $\mathcal{X}_i$ with respect to C_i is at most some constant times the optimal cost of $\mathcal{X}_i$ with respect to $\{c_1^*, ..., c_k^*\}$.
- $\forall i, \mathbf{E}[\phi_{\{c_1,...,c_i\}}(\mathcal{X}_i)|E_i] \leq 4 \cdot \phi_{\{c_1^*,...,c_k^*\}}(\mathcal{X}_i)$.

- From the last lemma, we get $\Pr\left[\phi_{\{c_1,\ldots,c_k\}}(\mathcal{X}) \leq 8 \cdot \phi_{\{c_1^*,\ldots,c_k^*\}}(\mathcal{X})\right] \geq \frac{1}{2}\Pr[E_k]$.
- We also show that $\Pr[E_{i+1}|E_i] \geq \frac{k-i}{k-i+1}$.
- This gives $\Pr[E_k] \geq \frac{1}{k}$.
- Hence, $\Pr\left[\phi_{\{c_1,\ldots,c_k\}}(\mathcal{X}) \leq 8 \cdot \phi_{\{c_1^*,\ldots,c_k^*\}}(\mathcal{X})\right] \geq \frac{1}{2k}$.
- Thus, the k-means++ algorithm gives an 8-approximation to the k-means objective with probability $\Omega(\frac{1}{k})$.

Conclusion and future work

- In this work, we showed that the k-means++ algorithm gives a constant factor approximation to the k-means and k-median objective with probability $\Omega(\frac{1}{k})$, provided all the clusters are large.
- We also showed that for small clusters, there is a dataset on which k-means++ can't achieve a constant factor approximation.
- Can we improve the upper and lower bounds in the analysis?