CMPSCI590D: Algorithms for Data Science

**Analysis of K-means++**

Instructor: Barna Saha

# 1 K-means Problem

For the $k$-means problem, we are given an integer $k$ and a set of $n$ data points $\mathcal{X} \in \mathbb{R}^d$. The goal is to select $k$ centers $\mathcal{C}$ to minimize the following objective.

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} ||x - c||^2$$

That is we want to select $k$ centers to minimize the sum of squared distances of all points to their respective nearest center.

$k$-means problem is NP-Hard. Lloyd's algorithm is a popular heuristic that is employed to solve $k$-means problem. However, depending on the distribution of the data points, it is possible that Lloyd's algorithm converges to a local optimum that is far from the global optimum. $k$-means++ is a way to ensure that Llyod's algorithm does not converge to a local optimum which is far off. In other words, even with $k$-means++, it is possible that the algorithm converges to a local optimum, but that local optimum is close to the global optimum.

## 1.1 Details of $k$-means++ Algorithm

Let $D(x)$ denote the distance of $x$ from the closest center.

1. Take one center $c_1$, chosen uniformly at random from $\mathcal{X}$.
2. Take a new center $c_i$, choosing $x \in \mathcal{X}$ with probability $\frac{D(x)}{\sum_{x \in \mathcal{X}} D(x)^2}$
3. Repeat step 2 until we have taken $k$ centers altogether.
4. Now proceed as with the standard $k$-means algorithm.

## 1.2 Analysis

Here is a standard result from Linear Algebra.

**Lemma 1.** *Let $S$ be a set of points with center of mass $c(S)$, and let $z$ be an arbitrary point. Then*

$$\sum_{x \in S} ||x - z||^2 - \sum_{x \in S} ||x - c(S)||^2 = |S||c(S) - z||^2$$

**Theorem 2.** *If $C$ is constructed with k-means++, then the corresponding objective function $\phi$ satisfies*

$$E[\phi] \leq 8(\ln k + 2)\phi_{OPT}$$

.

The above claim holds after the first iteration. Later iterations can only improve the bound as $\phi$ decreases.

We will only show a partial result. We will show if cluster centers are chosen from each cluster then $k$-means++ is 8-competitive.

**Theorem 3.** *Let $\mathcal{C}$ contains the current centers that have already been chosen. Let $A$ be a cluster in the optimal solution not represented in $\mathcal{C}$.*

*If we add a random center to $\mathcal{C}$ from $A$, chosen with $D^2$ weighing then*

$$E[\phi(A)] \leq 8\phi_{OPT}(A)$$

.

*Proof.*

$$E[\phi(A)] = \sum_{a_0 \in A} \frac{D(a_0)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min\left(D(a)^2, ||a - a_0||^2\right)$$

Let $z$ be the closest center to some $a \in A$. Using triangle inequality, we have

$$D(a_0) \leq ||a_0 - z||_2 \leq D(a) + ||a - a_0||_2$$

Therefore,

$$D(a_0)^2 \leq (D(a) + ||a - a_0||_2)^2 \leq 2D(a)^2 + 2||a - a_0||^2$$

Summing over all $a \in A$, we get

$$|A|D(a_0)^2 \leq 2\sum_{a \in A} D(a)^2 + 2\sum_{a \in A} ||a - a_0||^2$$
$$\text{or}\,, D(a_0)^2 \leq \frac{2}{|A|}\sum_{a \in A} D(a)^2 + \frac{2}{|A|}\sum_{a \in A} ||a - a_0||^2$$

Using the above in the expression for $E[\phi(A)]$ we get

$$\begin{aligned}
E[\phi(A)] &= \sum_{a_0 \in A} \frac{D(a_0)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min\left(D(a)^2, ||a - a_0||^2\right) \\
&\leq \sum_{a_0 \in A} \frac{\frac{2}{|A|} \sum_{a \in A} D(a)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min\left(D(a)^2, ||a - a_0||^2\right) + \sum_{a_0 \in A} \frac{\frac{2}{|A|} \sum_{a \in A} ||a - a_0||^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} \min\left(D(a)^2, ||a - a_0||^2\right) \\
&\leq \sum_{a_0 \in A} \frac{\frac{2}{|A|} \sum_{a \in A} D(a)^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} ||a - a_0||^2 + \sum_{a_0 \in A} \frac{\frac{2}{|A|} \sum_{a \in A} ||a - a_0||^2}{\sum_{a \in A} D(a)^2} \sum_{a \in A} D(a)^2 \\
&= \frac{4}{|A|} \sum_{a_0 \in A} \sum_{a \in A} ||a - a_0||^2
\end{aligned}$$

Now use Lemma 1.

$$\begin{aligned}
E[\phi(A)] &= \frac{4}{|A|} \sum_{a_0 \in A} \sum_{a \in A} ||a - a_0||^2 \\
&= \frac{4}{|A|} \sum_{a_0 \in A} \sum_{a \in A} ||a - c(A)||^2 + |A|(||a_0 - c(A)||^2) \\
&= 8 \sum_{a \in A} ||a - c(A)||^2 = 8\phi_{OPT}(A)
\end{aligned}$$

This completes the proof.

□