Complementi di Algoritmi e Strutture Dati

k-Means++

Docente: *Nicolò Cesa-Bianchi* versione 14 ottobre 2025

These lecture notes are based on a set of slides written by Marco Bressan in 2023.

Recall the k-means problem: given a set $\mathcal{X} \subset \mathbb{R}^d$ of size n and $1 < k < n$, find

$$\mathcal{C}^* \in \operatorname*{argmin}_{\boldsymbol{c}_1,\dots,\boldsymbol{c}_k \in \mathbb{R}^d} \Phi(\boldsymbol{c}_1,\dots,\boldsymbol{c}_k)$$

where, for any $\mathcal{C} \subset \mathbb{R}^d$,

$$\Phi(\mathcal{C}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \phi(\mathcal{C},\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{c}_i\in\mathcal{C}} \|\boldsymbol{x}-\boldsymbol{c}_i\|^2$$

Let $\mathrm{OPT} = \Phi(\mathcal{C}^*)$ and, for any $\mathcal{C} \subset \mathbb{R}^d$ and $A \subseteq \mathcal{X}$, let

$$\phi(\mathcal{C},A) = \sum_{\boldsymbol{x}\in A} \phi(\mathcal{C},\boldsymbol{x})$$

We identify a clustering $\mathcal{C}$ via its centers $\{\boldsymbol{c}_1,\dots,\boldsymbol{c}_k\}$ or with its clusters $\{C_1,\dots,C_k\}$. Note that, for any clustering $\mathcal{C}$ output by Lloyd's algorithm, including the optimal clustering $\mathcal{C}^*$,

$$\phi(\mathcal{C},C) = \sum_{\boldsymbol{x}\in C} \|\boldsymbol{x}-\boldsymbol{\mu}_C\|^2 \qquad \text{for all } C\in\mathcal{C}, \text{ where } \boldsymbol{\mu}_C \text{ is the centroid of } C. \tag{1}$$

We proved that Lloyd's algorithm does not have any approximation guarantee because, while outliers can contribute a lot to the overall cost, they are are not favored in the initial random draw of centers.

Algoritmo 1 `k-means++`

Input: Finite set of points $\mathcal{X} \subset \mathbb{R}^d$, integer $1 < k < |\mathcal{X}|$.
1: Draw a center $\boldsymbol{c}_1$ u.a.r. from $\mathcal{X}$ and let $\mathcal{C}_1 = \{\boldsymbol{c}_1\}$
2: **for** $i = 2,\dots,k$ **do**
3: draw $\boldsymbol{c}_i$ from $\mathcal{X}$ according to the distribution $\mathbb{P}(\boldsymbol{c}_i = \boldsymbol{x} \mid \mathcal{C}_{i-1}) = \dfrac{\phi(\mathcal{C}_{i-1},\boldsymbol{x})}{\Phi(\mathcal{C}_{i-1})}$
4: $\mathcal{C}_i = \mathcal{C}_{i-1} \cup \{\boldsymbol{c}_i\}$
5: **end for**
Output: The output of Lloyd's algorithm run with initial centers $\boldsymbol{c}_1,\dots,\boldsymbol{c}_k$

We prove a simplified version of the following theorem.

Teorema 1 *The clustering $\mathcal{C}$ found by `k-means++` satisfies* $\mathbb{E}\big[\Phi(\mathcal{C})\big] \le 8(\ln k + 2)\mathrm{OPT}$.

Note that the currently best known approximation algorithms for k-means is based on a linear programming rounding approach and produces a clustering with a cost $c \times \mathrm{OPT}$ where $c \in [6,7]$.

Consider any optimal clustering $\mathcal{C}^* = (A_1,\dots,A_k)$ and for let $\mathcal{C}_i$ be the clustering of `k-means++` after drawing the first i centers in Line 3.

Lemma 2 *For any $A \in \mathcal{C}^*$ and for any $i \in [k]$,*

$$\mathbb{E}\Big[\phi(\mathcal{C}_i, A) \,\Big|\, \boldsymbol{c}_i \in A, \mathcal{C}_{i-1}\Big] \le 8\,\phi(\mathcal{C}^*, A)$$

Dimostrazione. Consider first $i = 1$. Then $\mathcal{C}_{i-1} = \mathcal{C}_0 = \emptyset$ and c_i is drawn according to the uniform distribution over $\mathcal{X}$, and we can write

$$\begin{aligned}
\mathbb{E}\big[\phi(\mathcal{C}_1, A) \mid \boldsymbol{c}_1 \in A\big] &= \frac{1}{|A|}\sum_{\boldsymbol{a}\in A}\left(\sum_{\boldsymbol{x}\in A}\|\boldsymbol{x}-\boldsymbol{a}\|^2\right) && (\mathcal{C}_1 = \{\boldsymbol{c}_1\})\\
&\le \frac{1}{|A|}\sum_{\boldsymbol{a}\in A}\left(\sum_{\boldsymbol{x}\in A}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 + |A|\,\|\boldsymbol{a}-\boldsymbol{\mu}\|^2\right) && (\boldsymbol{\mu} \text{ is the centroid of } A)\\
&= \sum_{\boldsymbol{x}\in A}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 + \sum_{\boldsymbol{a}\in A}\|\boldsymbol{a}-\boldsymbol{\mu}\|^2\\
&= 2\sum_{\boldsymbol{x}\in A}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 = 2\,\phi(\mathcal{C}^*, A) && \text{(because of (1).)}
\end{aligned}$$

In particular, note that

$$\frac{1}{|A|}\sum_{\boldsymbol{a}\in A}\sum_{\boldsymbol{x}\in A}\|\boldsymbol{x}-\boldsymbol{a}\|^2 \le 2\,\phi(\mathcal{C}^*, A) \tag{2}$$

Now assume $i > 1$. Then

$$\mathbb{P}\big(\boldsymbol{c}_i = \boldsymbol{a} \mid \boldsymbol{a} \in A, \mathcal{C}_{i-1}\big) = \frac{\phi(\mathcal{C}_{i-1}, \boldsymbol{a})}{\sum_{\boldsymbol{x}\in A}\phi(\mathcal{C}_{i-1}, \boldsymbol{x})}$$

For any $\boldsymbol{x}, \boldsymbol{a} \in A$, let $\boldsymbol{c}$ be the center of $\mathcal{C}_{i-1}$ closest to $\boldsymbol{x}$. Then

$$\begin{aligned}
\min_{j=1,\dots,i-1}\|\boldsymbol{a}-\boldsymbol{c}_j\| &\le \|\boldsymbol{a}-\boldsymbol{c}\|\\
&\le \|\boldsymbol{x}-\boldsymbol{c}\| + \|\boldsymbol{a}-\boldsymbol{x}\| && \text{(by the triangular inequality.)}
\end{aligned}$$

Using $(a+b)^2 \le 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$ and $\|\boldsymbol{x}-\boldsymbol{c}\|^2 = \phi(\mathcal{C}_{i-1}, \boldsymbol{x})$ we get

$$\phi(\mathcal{C}_{i-1}, \boldsymbol{a}) \le 2\Big(\phi(\mathcal{C}_{i-1}, \boldsymbol{x}) + \|\boldsymbol{a}-\boldsymbol{x}\|^2\Big)$$

By averaging the above inequality over all $\boldsymbol{x} \in A$, we get

$$\phi(\mathcal{C}_{i-1}, \boldsymbol{a}) \le \frac{2}{|A|}\sum_{\boldsymbol{x}\in A}\Big(\phi(\mathcal{C}_{i-1}, \boldsymbol{x}) + \|\boldsymbol{a}-\boldsymbol{x}\|^2\Big)$$

Note also that, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\phi(\mathcal{C}_i, \boldsymbol{x}) = \min\Big\{\phi(\mathcal{C}_{i-1}, \boldsymbol{x}), \|\boldsymbol{x}-\boldsymbol{c}_i\|^2\Big\}$$

Therefore, for $\mathcal{C}_i = \mathcal{C}_{i-1} \cup \{\boldsymbol{c}_i\}$,

$$
\begin{aligned}
\mathbb{E}\big[\phi(\mathcal{C}_i, A) \mid \boldsymbol{c}_i \in A, \mathcal{C}_{i-1}\big] &= \sum_{\boldsymbol{a}\in A} \frac{\phi(\mathcal{C}_{i-1}, \boldsymbol{a})}{\sum_{\boldsymbol{x}\in A} \phi(\mathcal{C}_{i-1}, \boldsymbol{x})} \phi(\mathcal{C}_i, A) \\
&\leq \frac{2}{|A|} \sum_{\boldsymbol{a}\in A} \sum_{\boldsymbol{x}\in A} \frac{\phi(\mathcal{C}_{i-1}, \boldsymbol{x}) + \|\boldsymbol{a} - \boldsymbol{x}\|^2}{\sum_{\boldsymbol{x}'\in A} \phi(\mathcal{C}_{i-1}, \boldsymbol{x}')} \sum_{\boldsymbol{a}'\in A} \min\left\{\phi(\mathcal{C}_{i-1}, \boldsymbol{a}'), \left\|\boldsymbol{a}' - \boldsymbol{a}\right\|^2\right\} \\
&= \frac{2}{|A|} \sum_{\boldsymbol{a}\in A} \frac{\sum_{\boldsymbol{x}\in A} \phi(\mathcal{C}_{i-1}, \boldsymbol{x})}{\sum_{\boldsymbol{x}'\in A} \phi(\mathcal{C}_{i-1}, \boldsymbol{x}')} \sum_{\boldsymbol{a}'\in A} \min\left\{\phi(\mathcal{C}_{i-1}, \boldsymbol{a}'), \left\|\boldsymbol{a}' - \boldsymbol{a}\right\|^2\right\} \\
&\quad + \frac{2}{|A|} \sum_{\boldsymbol{a}\in A} \sum_{\boldsymbol{x}\in A} \frac{\|\boldsymbol{a} - \boldsymbol{x}\|^2}{\sum_{\boldsymbol{x}'\in A} \phi(\mathcal{C}_{i-1}, \boldsymbol{x}')} \sum_{\boldsymbol{a}'\in A} \min\left\{\phi(\mathcal{C}_{i-1}, \boldsymbol{a}'), \left\|\boldsymbol{a}' - \boldsymbol{a}\right\|^2\right\} \\
&\leq \frac{2}{|A|} \sum_{\boldsymbol{a}\in A} \sum_{\boldsymbol{a}'\in A} \left\|\boldsymbol{a}' - \boldsymbol{a}\right\|^2 + \frac{2}{|A|} \sum_{\boldsymbol{a}\in A} \sum_{\boldsymbol{x}\in A} \|\boldsymbol{a} - \boldsymbol{x}\|^2 \\
&= \frac{4}{|A|} \sum_{\boldsymbol{a}\in A} \sum_{\boldsymbol{x}\in A} \|\boldsymbol{x} - \boldsymbol{a}\|^2 \\
&\leq 8\, \phi(\mathcal{C}^*, A) \qquad \text{(because of (2).)}
\end{aligned}
$$

concluding the proof. □

A cluster $A \in \mathcal{C}^*$ is uncovered in $\mathcal{C}_i$ if $A \cap \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_i\} = \emptyset$. Lemma 2 shows that we pay $\mathcal{O}(\mathrm{OPT})$ for every optimal cluster that we cover. This justifies the following simplifying assumptions, stating that the cost of each optimal cluster is set to 1, and we pay 1 for each optimal cluster that is covered and L for each optimal cluster that remains uncovered.

Assunzione 3 *For all $A \in \mathcal{C}^*$:*

1. $\phi(\mathcal{C}^*, A) = 1$
2. *for all $i \in [k]$, if A is covered in $\mathcal{C}_i$, then then $\phi(\mathcal{C}_i, A) = 1$; otherwise, $\phi(\mathcal{C}_i, A) = L$.*

Lemma 4 *Under the above assumptions, $\mathbb{E}\big[\Phi(\mathcal{C})\big] \leq \big(2 + \ln k\big)\mathrm{OPT}$.*

Dimostrazione. Let $\mathcal{C}_i = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_i)$. Conventionally, $\mathcal{C}_0 = \emptyset$ and $\Phi(\mathcal{C}_0) = kL$ (as if there were a default faraway center). Now, observing that $\mathcal{C} = \mathcal{C}_k$,

$$
\Phi(\mathcal{C}_k) = \Phi(\mathcal{C}_0) + \sum_{i=0}^{k-1} \left(\Phi(\mathcal{C}_{i+1}) - \Phi(\mathcal{C}_i)\right)
$$

Taking expectations,

$$\begin{aligned}
\mathbb{E}\big[\Phi(\mathcal{C}_k)\big] &= \Phi(\mathcal{C}_0) + \sum_{i=0}^{k-1} \big(\mathbb{E}\big[\Phi(\mathcal{C}_{i+1})\big] - \mathbb{E}\big[\Phi(\mathcal{C}_i)\big]\big) \\
&= kL + \sum_{i=0}^{k-1} \big(\mathbb{E}\big[\Phi(\mathcal{C}_{i+1})\big] - \mathbb{E}\big[\Phi(\mathcal{C}_i)\big]\big) \\
&= k + \sum_{i=0}^{k-1} \big((L-1) + \mathbb{E}\big[\Phi(\mathcal{C}_{i+1})\big] - \mathbb{E}\big[\Phi(\mathcal{C}_i)\big]\big)
\end{aligned}$$

Let N_i the number of uncovered clusters in $\mathcal{C}_i$. Because of our assumptions, $\Phi(\mathcal{C}_i) = N_i L + (k - N_i)$.

For any uncovered A, the probability that at round $i+1$ we choose a center from A is

$$\mathbb{P}(\boldsymbol{c}_{i+1} \in A \mid \mathcal{C}_i) = \frac{\phi(\mathcal{C}_i, A)}{\Phi(\mathcal{C}_i)} = \frac{L}{N_i L + (k - N_i)}$$

So the probability p_{i+1} that we choose a center from some uncovered cluster is:

$$\mathbb{P}\Big(\exists A \in \mathcal{C}^* \,:\, \boldsymbol{c}_{i+1} \in A \,\wedge\, A \cap \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_i\} = \emptyset \mid \mathcal{C}_i\Big) = \frac{N_i L}{N_i L + (k - N_i)} \geq \frac{(k-i)L}{(k-i)L + i}$$

where in the last inequality we used $N_i \geq k - i$.

Now, if $\boldsymbol{c}_{i+1}$ does not cover any A that was uncovered in $\mathcal{C}_i$ (which happens with probability $1 - p_{i+1}$), then $\Phi(\mathcal{C}_{i+1}) \leq \Phi(\mathcal{C}_i)$. On the other hand, if $\boldsymbol{c}_{i+1}$ covers some A that was uncovered in $\mathcal{C}_i$ (which happens with probability p_{i+1}), then $\Phi(\mathcal{C}_{i+1}) = \Phi(\mathcal{C}_i) - L + 1 = \Phi(\mathcal{C}_i) - (L-1)$. Therefore

$$\begin{aligned}
(L-1) + \mathbb{E}\big[\Phi(\mathcal{C}_{i+1}) \mid \mathcal{C}_i\big] - \mathbb{E}\big[\Phi(\mathcal{C}_i) \mid \mathcal{C}_i\big] &\leq (L-1) + 0 \times (1 - p_{i+1}) - (L-1)p_{i+1} \\
&\leq (L-1) - (L-1)\frac{(k-i)L}{(k-i)L + i} \\
&= (L-1)\left(\frac{i}{(k-i)L + i}\right) \\
&< L\frac{i}{(k-i)L + i} \\
&< L\frac{k}{(k-i)L} = \frac{k}{k-i}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\big[\Phi(\mathcal{C}_k)\big] &= k + \sum_{i=0}^{k-1} \big((L-1) + \mathbb{E}\big[\Phi(\mathcal{C}_{i+1})\big] - \mathbb{E}\big[\Phi(\mathcal{C}_i)\big]\big) \\
&= k + \sum_{i=0}^{k-1} \mathbb{E}\Big[(L-1) + \mathbb{E}\big[\Phi(\mathcal{C}_{i+1}) \mid \mathcal{C}_i\big] - \mathbb{E}\big[\Phi(\mathcal{C}_i) \mid \mathcal{C}_i\big]\Big] \\
&\leq k + \sum_{i=0}^{k-1} \frac{k}{k-i} \\
&= k + k\sum_{i=1}^{k} \frac{1}{i} \leq k(2 + \ln k)
\end{aligned}$$

where we used the bound on the harmonic sum $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{k} \leq 1+\ln k$. The proof is concluded by noticing that, under our assumptions, $\mathrm{OPT} = \Phi(\mathcal{C}^*) = k$. $\square$