大数据算法-2025 春

Lecture 17: Beyond Worst Case: K-means

2025.5.29

Lecturer: 丁虎 *Scribe:* 王向禄

1 基本知识

Definition 1.1. 若$\Delta_k^2(X) \leq \varepsilon^2 \Delta_{k-1}^2(X)$, 则称$X$ 是ε-Separated. 其中，X 表示 $\mathbb{R}^d$ 中的 n 个输入点构成的集合，$\Delta_k^2(X)$ 表示对 X 进行 k-means 聚类时所得的最小代价（即最优解的 cost）[1]。

Lemma 1.2. 对于任意点集 $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$，定义其均值为：

$$\mu(X) = \frac{1}{n} \sum_{x \in X} x,$$

则有以下恒等式成立：

$$\forall x \in X, \quad \sum_{y \in X} \|x - y\|^2 = \Delta_1^2(X) + n\|x - \mu(X)\|^2,$$

进而有：

$$\sum_{x \in X} \sum_{y \in X} \|x - y\|^2 = n\Delta_1^2(X) + n \sum_{x \in X} \|x - \mu(X)\|^2.$$

注意到右侧第二项等于 $n\Delta_1^2(X)$，因此：

$$\sum_{x \in X} \sum_{y \in X} \|x - y\|^2 = n\Delta_1^2(X) + n\Delta_1^2(X) = 2n\Delta_1^2(X).$$

Lemma 1.3. 设点集 X 被划分为两个子集 X_1 和 X_2，记 $n_1 = |X_1|, n_2 = |X_2|, n = |X|$，则：

1. $\Delta_1^2(X) = \Delta_1^2(X_1) + \Delta_1^2(X_2) + \frac{n_1 n_2}{n} \|\mu(X_1) - \mu(X_2)\|^2$
2. $\|\mu(X_1) - \mu(X)\|^2 \leq \frac{\Delta_1^2(X)}{n} \cdot \frac{n_2}{n_1}$

2 The 2-Means Problem

设 $k=2$，μ_1, μ_2 为最佳的两个子簇中心点。

假设:

$$\Delta_2^2(X) \le \varepsilon^2 \Delta_1^2(X)$$

算法:

1. **采样（Sampling）:** 从集合 X 中随机选取一对点作为初始中心，选取点对 $x, y \in X$ 的概率与 $\|x-y\|^2$ 成正比 ($\frac{\|x-y\|^2}{\sum_{x,y\in X}\|x-y\|^2}$)。设 $\hat{\mu}_1, \hat{\mu}_2$ 为选出的两个中心点。
2. **"Ball-k-Means" 步骤:** 对于每个 $\hat{\mu}_i$，以其为中心、半径为 $r = \|\hat{\mu}_1 - \hat{\mu}_2\|/3$ 的球中，计算集合 X 在该球内部分的质心，记为 $\bar{\mu}_i$。返回 $\bar{\mu}_1, \bar{\mu}_2$ 作为最终中心。

运行时间（Running Time）: 整个算法的运行时间为 $O(nd)$。步骤 (2) 显然只需 $O(nd)$ 时间。我们将证明：采样步骤可以在 $O(nd)$ 时间内实现。

考虑以下的两步采样过程:

(a) 首先，从集合 X 中选择一个点 x，其被选中的概率为:

$$\frac{\sum_{y\in X}\|x-y\|^2}{\sum_{x,y\in X}\|x-y\|^2} = \frac{\Delta_1^2(X) + n\|x-\mu(X)\|^2}{2n\Delta_1^2(X)}$$

（由引理 1.2 得出）;

(b) 然后，从 X 中选择第二个中心 y，其被选中的概率为:

$$\frac{\|y-\hat{\mu}_1\|^2}{\Delta_1^2(X) + n\|\mu(X)-\hat{\mu}_1\|^2}$$

这个两步采样过程等价于步骤 (1) 中的采样过程，即以如下概率选择点对 $x_1, x_2 \in X$:

$$\frac{\|x_1-x_2\|^2}{\sum_{(x,y)\in X}\|x-y\|^2}$$

由于 $\Delta_1^2(X)$ 可预先计算，因此每一步都只需 $O(nd)$ 时间。

Lemma 2.1. $\max(r_1^2, r_2^2) \le \frac{\varepsilon^2}{1-\varepsilon^2}\|\mu_1-\mu_2\|^2 = O(\varepsilon^2)\|\mu_1-\mu_2\|^2$. 其中 $r_i^2 = \frac{\Delta_1^2(X_i)}{n_i}$

Proof. 根据引理 1.3 的第 (1) 点，有

$$\Delta_1^2(X) = \Delta_2^2(X) + \frac{n_1 n_2}{n}\|\mu_1 - \mu_2\|^2,$$

这等价于

$$\frac{n}{n_1 n_2} \cdot \Delta_2^2(X) = \frac{\|\mu_1 - \mu_2\|^2 \cdot \Delta_2^2(X)}{\Delta_1^2(X) - \Delta_2^2(X)}.$$

这意味着：

$$r_1^2 \cdot \frac{n}{n_2} + r_2^2 \cdot \frac{n}{n_1} \le \frac{\varepsilon^2}{1-\varepsilon^2}\|\mu_1 - \mu_2\|^2.$$

□

假设 $\rho = \frac{100\varepsilon^2}{1-\varepsilon^2}$，我们要求 $\rho < \frac{1}{4}$，因此 $\varepsilon^2 < \frac{1}{401}$。我们定义簇 X_i 的核心为：

$$X_i^{\text{cor}} = \left\{ x \in X_i : \|x - \mu_i\|^2 \le \frac{r_i^2}{\rho} \right\}.$$

由 Markov 不等式可知，$|X_i^{\text{cor}}| \ge (1-\rho)n_i$，对 $i = 1, 2$ 成立。

Lemma 2.2. $\Pr\left[\{\hat{c}_1, \hat{c}_2\} \cap X_1^{cor} \neq \varnothing \quad \text{且} \quad \{\hat{c}_1, \hat{c}_2\} \cap X_2^{cor} \neq \varnothing\right] \ge 1 - 4\rho.$

Proof. 为简化表达，我们假设所有点按 $\frac{1}{\|\mu_1 - \mu_2\|}$ 缩放（因此 $\|\mu_1 - \mu_2\| = 1$）。

由引理 1.3 的 (1) 部分可得：

$$\Delta_1^2(X) = \Delta_2^2(X) + \frac{n_1 n_2}{n}\|\mu_1 - \mu_2\|^2 \Rightarrow \Delta_1^2(X) \le \frac{n_1 n_2}{n(1-\varepsilon^2)} \quad \text{（因为}\Delta_2^2(X) < \varepsilon^2 \Delta_1^2(X)\text{）}.$$

令 μ_i' 为 X_i^{cor} 的质心。由引理 1.3 (2)（令 $S = X_i$，$S_1 = X_i^{\text{cor}}$）有：

$$\|\mu_i' - \mu_i\|^2 \le \frac{\rho}{1-\rho} r_i^2.$$

记事件发生的概率为 A/B，其中

$$A = \sum_{x \in X_1^{\text{cor}}} \sum_{y \in X_2^{\text{cor}}} \|x - y\|^2 = \sum_{x \in X_1^{\text{cor}}} \left(\Delta_1^2(X_2^{\text{cor}}) + |X_2^{\text{cor}}| \|x - \mu_2'\|^2 \right),$$

整理得：

$$A = |X_1^{\text{cor}}| \Delta_1^2(X_2^{\text{cor}}) + |X_2^{\text{cor}}| \Delta_1^2(X_1^{\text{cor}}) + |X_1^{\text{cor}}||X_2^{\text{cor}}| \|\mu_1' - \mu_2'\|^2 \ge (1-\rho)^2 n_1 n_2 \|\mu_1' - \mu_2'\|^2.$$

$$B = \sum_{(x,y) \subseteq X} \|x - y\|^2 = n\Delta_1^2(X) \le \frac{n_1 n_2}{1-\varepsilon^2}.$$

结合 Lemma 2.1 以及 $\|\mu_i' - \mu_i\|$ 的上界，可得：

$$\|\mu_1' - \mu_2'\| \geq \|\mu_1 - \mu_2\| - 2 \cdot \frac{\rho}{1-\rho} \max(r_1, r_2) \geq 1 - 2\varepsilon\sqrt{\frac{\rho}{(1-\rho)(1-\varepsilon^2)}} \geq 1 - \frac{\rho}{5\sqrt{1-\rho}}.$$

综上：

$$A \geq \left(1 - 2\rho - \frac{2\rho}{5\sqrt{1-\rho}}\right) n_1 n_2, \quad 且 \quad \frac{A}{B} \geq 1 - 4\rho.$$

□

Lemma 2.3. 对于任意的 i，我们有 $X_i^{cor} \subseteq B_i \subseteq X_i$ 因此 $\|\bar{\mu}_i - \mu_i\|^2 \leq \frac{\rho}{1-\rho} \cdot r_i^2$。其中 $B_i = \left\{x \in X : \|x - \hat{\mu}_i\| \leq \frac{\|\hat{\mu}_1 - \hat{\mu}_2\|}{3}\right\}$.

Proof. 令

$$\theta = \frac{\varepsilon}{\sqrt{\rho(1-\varepsilon^2)}} \leq \frac{1}{10}$$

根据引理 2.1，可得：

$$\|\hat{\mu}_i - \mu_i\| \leq \sqrt{\frac{\rho}{1-\rho}} \cdot r_i \leq \theta\|\mu_1 - \mu_2\|, \quad 对 i = 1, 2$$

因此：

$$\frac{4}{5} \leq \frac{\|\hat{\mu}_1 - \hat{\mu}_2\|}{\|\mu_1 - \mu_2\|} \leq \frac{6}{5}.$$

对任意 $x \in B_i$，有：

$$\|x - \mu_i\| \leq \|x - \hat{\mu}_i\| + \|\hat{\mu}_i - \mu_i\| \leq \frac{\|\mu_1 - \mu_2\|}{2}$$

所以 $x \in X_i$。对于任意 $x \in X_i^{\text{cor}}$，由于：

$$\|x - \hat{\mu}_i\| \leq 2\theta\|\mu_1 - \mu_2\| \leq \frac{\|\hat{\mu}_1 - \hat{\mu}_2\|}{3}$$

所以 $x \in B_i$。因此有 $X_i^{\text{cor}} \subseteq B_i \subseteq X_i$。

根据引理 1.3 的第 (2) 部分，取 $S = X_i$，$S_1 = B_i$，并注意 $|B_i| \geq |X_i^{\text{cor}}|$，可得：

$$\|\bar{\mu}_i - \mu_i\|^2 \leq \frac{\rho}{1-\rho} \cdot r_i^2$$

□

Theorem 2.4. 该算法返回的聚类，其代价最多为：

$$\frac{\Delta_2^2(X)}{1-\rho},$$

并且以至少 $1-O(\rho)$ 的概率成功，算法运行时间为 $O(nd)$，其中：

$$\rho = \frac{100\varepsilon^2}{1-\varepsilon^2}$$

Proof. 该解的总损失至多为：

$$\sum_{i,\, x\in X_i} \|x-\bar{\mu}_i\|^2 = \sum_i \left(\Delta_1^2(X_i) + n_i\|\bar{\mu}_i - \mu_i\|^2\right) \le \frac{\Delta_2^2(X)}{1-\rho}$$

□

References

[1] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22, 2013.