中国科学技术大学
计算机科学与技术学院

《大数据算法》作业
2022 年春
截止日期：2022 年 6 月 9 日 23:59

Exercise 1 *20 分*
证明（关于欧氏 k-means 问题的）coresets 满足下面的可组合性质 (composability)：
令 $A_1, A_2 \subseteq \mathbb{R}^d$ 是两个互不相交的集合。假设集合 S_1 及权重函数 $w_1 : S_1 \to \mathbb{R}$ 和集合 S_2 及权重函数 $w_2 : S_2 \to \mathbb{R}$ 分别是 A_1 和 A_2 的 (k, ε)-coresets。那么 $S_1 \cup S_2$ 及函数 $w_1 + w_2 : S_1 \cup S_2 \to \mathbb{R}$ 是 $A_1 \cup A_2$ 的 (k, ε)-coreset。
注： 这里 $w_1 + w_2$ 的定义如下：

$$(w_1 + w_2)(x) = \begin{cases} w_1(x) & \text{如果 } x \in S_1 \setminus S_2, \\ w_2(x) & \text{如果 } x \in S_2 \setminus S_1, \\ w_1(x) + w_2(x) & \text{如果 } x \in S_2 \cap S_1 \end{cases}$$

证明. Let

$$D(A, C) = \sum_{x \in A} \min_{c \in C} \|x - c\|^2$$
$$D(A, w, C) = \sum_{x \in A} w(x) \min_{c \in C} \|x - c\|^2$$

and

$$w(x) = \begin{cases} w_1(x) & x \in S_1 \setminus S_2, \\ w_2(x) & x \in S_2 \setminus S_1, \\ w_1(x) + w_2(x) & x \in S_2 \cap S_1 \end{cases}$$

Then, $\forall C \subseteq \mathbb{R}^d$ with $|C| \leq k$, we have

$$D(A_1 \cap A_2, C) = D(A_1, C) + D(A_2, C)$$

and

$$\begin{aligned} D(S_1 \cap S_2, w, C) &= D(S_1 \setminus S_2, w_1, C) + D(S_1 \cap S_2, w_1 + w_2, C) + D(S_2 \setminus S_1, w_2, C) \\ &= D(S_1 \setminus S_2, w_1, C) + (D(S_1 \cap S_2, w_1, C) + D(S_1 \cap S_2, w_2, C)) + D(S_2 \setminus S_1, w_2, C) \\ &= (D(S_1 \setminus S_2, w_1, C) + D(S_1 \cap S_2, w_1, C)) + (D(S_1 \cap S_2, w_2, C) + D(S_2 \setminus S_1, w_2, C)) \\ &= D(S_1, w_1, C) + D(S_2, w_2, C) \end{aligned}$$

Since set S_1 with weight function w_1 and set S_2 with weight function w_2 are (k, ε)-coresets of A_1 and A_2 respectively. $\forall C \subseteq \mathbb{R}^d$ with $|C| \leq k$, we have

$$|D(A_1, C) - D(S_1, w_1, C)| \leq \varepsilon D(A_1, C)$$
$$|D(A_2, C) - D(S_2, w_2, C)| \leq \varepsilon D(A_2, C)$$

Therefore, $\forall C \subseteq \mathbb{R}^d$ with $|C| \leq k$, there is

$$\begin{aligned}
|D(A_1 \cap A_2, C) - D(S_1 \cap S_2, w, C)| &= |(D(A_1, C) + D(A_2, C)) - (D(S_1, w_1, C) + D(S_2, w_2, C))| \\
&= |(D(A_1, C) - D(S_1, w_1, C)) + (D(A_2, C) - D(S_2, w_2, C))| \\
&\leq |D(A_1, C) - D(S_1, w_1, C)| + |D(A_2, C) - D(S_2, w_2, C)| \\
&\leq \varepsilon D(A_1, C) + \varepsilon D(A_2, C) \\
&= \varepsilon (D(A_1, C) + D(A_2, C)) \\
&= \varepsilon D(A_1 \cap A_2, C)
\end{aligned}$$

Consequently, set $S_1 \cap S_2$ with weight function w is a (k, ε)-coreset for all $C \subseteq \mathbb{R}^d$ with $|C| \leq k$.
Q.E.D. □

Exercise 2 *20 分*

- 对于欧氏 k-median 问题，我们可以限制 k 个中心点 $c_1, \ldots, c_k$ 都是来自于输入数据集 A 中的，也可以允许它们是来自整个欧氏空间 $\mathbb{R}^d$ 的。证明在这两种情况下，问题的最优解所对应的目标函数值的比值不超过 2。
- 对于欧氏 k-means 问题，我们可以限制 k 个中心点 $c_1, \ldots, c_k$ 都是来自于输入数据集 A 中的，也可以允许它们是来自整个欧氏空间 $\mathbb{R}^d$ 的。证明在这两种情况下，问题的最优解所对应的目标函数值的比值不超过 4。

---

k-median

证明. Let $C = \{c_1, c_2, \cdots c_k\}$ be the optimal solution to the k-median clustering, and let $S_1, S_2, \cdots S_k$ be the corresponding clusters. Similarly, let C' be the optimal solution to the k-median clustering when $\forall c' \in C'$, there is $c' \in A$. Choose $b_1, b_2, \cdots b_k$ s.t.

$$b_i = \text{argmin}_{b_i \in S_i} \|b_i - c_i\|$$

Since C' is the optimal solution, we have

$$\begin{aligned}
D(A, C') = \sum_{i=1}^{k} D(S_i, C') &\leq \sum_{i=1}^{k} D(S_i, b_i) \\
&= \sum_{i=1}^{k} \sum_{a \in S_i} \|a - b_i\| \\
&\leq \sum_{i=1}^{k} \sum_{a \in S_i} (\|a - c_i\| + \|c_i - b_i\|) \\
&\leq \sum_{i=1}^{k} (\sum_{a \in S_i} \|a - c_i\| + \sum_{a \in S_i} \|c_i - a\|) \\
&= 2 \sum_{i=1}^{k} \sum_{a \in S_i} \|a - c_i\| \\
&= 2D(A, C)
\end{aligned}$$

Therefore

$$\frac{D(A,C')}{D(A,C)} \leq 2$$

Q.E.D. □

k-means

证明. Let $C = \{c_1, c_2, \cdots c_k\}$ be the optimal solution to the k-means clustering, and let $S_1, S_2, \cdots S_k$ be the corresponding clusters. Similarly, let C' be the optimal solution to the k-means clustering when $\forall c' \in C'$, there is $c' \in A$. Choose $b_1, b_2, \cdots b_k$ s.t.

$$b_i = \text{argmin}_{b_i \in S_i} \|b_i - c_i\|^2$$

As c_i is the centroid of S_i $(1 \leq i \leq k)$, there is

$$c_i = \frac{1}{|S_i|} \sum_{a \in S_i} a$$

Since C' is the optimal solution, we have

$$\begin{aligned}
D(A,C') = \sum_{i=1}^{k} D(S_i, C') &\leq \sum_{i=1}^{k} D(S_i, b_i) \\
&= \sum_{i=1}^{k} \sum_{a \in S_i} \|a - b_i\|^2 \\
&= \sum_{i=1}^{k} (\sum_{a \in S_i} \|a - c_i\|^2 + |S_i| \|c_i - b_i\|^2) \\
&\leq \sum_{i=1}^{k} (\sum_{a \in S_i} \|a - c_i\|^2 + \sum_{a \in S_i} \|c_i - a\|^2) \\
&= 2 \sum_{i=1}^{k} \sum_{a \in S_i} \|a - c_i\|^2 \\
&= 2D(A,C)
\end{aligned}$$

Therefore

$$\frac{D(A,C')}{D(A,C)} \leq 2$$

Q.E.D. □

Exercise 3 *20 分*

考虑平面 $\mathbb{R}^2$ 上的 k-median 问题，其中我们要求 k 个中心点 $c_1, \ldots, c_k$ 都是来自于输入数据集 A 中的。考虑枚举所有可能的聚类并从中选出具有最小代价的聚类。我们可以将所有的 n 个点进行标号，每个标号是 $\{1, \ldots, k\}$ 中的一个数。注意到所有可能的标号数是 k^n，这对应着高昂的时间。

证明我们可以在 $n^{O(k)}$ 时间内找到最优的聚类。

证明. Since $\forall 1 \leq i \leq k, c_i \in A$, we can enumerate all possible set of centroids, compute the cost and choose the one with the optimal cost. In practice, we can keep two data structures *cost_min* and *vector_min*. *vector_min* is a n-bit vector where each bit is set as 1 if the n-th point is chosen as one of the k centroids and as 0 otherwise. *cost_min* is the minimum cost found so far associated with a *vector_min*. *cost_min* and *vector_min* are only updated after computation of the cost in each enumeration. Therefore, the time complexity of the problem is

$$\begin{aligned} T &= O(\mathrm{C}_n^k \cdot kn) \\ &= O(n^k \cdot kn) \\ &= n^{O(k)} \end{aligned}$$

Q.E.D. □

Exercise 4 *10 分*

令 a, b, c 为任意三个实数。证明对于任意的 $\varepsilon \in (0, 1)$，下面的不等式（即推广的三角不等式）成立：

$$\big||a-c|^2 - |b-c|^2\big| \leq \frac{12}{\varepsilon} \cdot |a-b|^2 + 2\varepsilon \cdot |a-c|^2$$

证明. Since a, b and c are real numbers, let $x = a - c$ and $y = b - c$, and the inequality can be rewritten as

$$|x^2 - y^2| \leq \frac{12}{\varepsilon}(x-y)^2 + 2\varepsilon x^2$$

that is

$$2x^2\varepsilon^2 - |x^2 - y^2|\varepsilon + 12(x-y)^2 \geq 0$$

When $x^2 \geq y^2$, the inequality turns to

$$2x^2\varepsilon^2 - (x^2 - y^2)\varepsilon + 12(x-y)^2 \geq 0$$

Since

$$\begin{aligned} \frac{x^2 - y^2}{2 \cdot 2x^2} - 1 &= \frac{x^2 - y^2 - 4x^2}{4x^2} \\ &= -\frac{3x^2 + y^2}{4x^2} \\ &\leq 0 \end{aligned}$$

,i.e. $\frac{x^2-y^2}{4x^2} \leq 1$, and $\varepsilon \in (0, 1)$, the minimum value of the function reaches when $\varepsilon = \frac{x^2-y^2}{4x^2}$. And the minimum value of the function is

$$\begin{aligned} \frac{4 \cdot 2x^2 \cdot 12(x-y)^2 - (x^2-y^2)^2}{4 \cdot 2x^2} &= \frac{96x^2(x-y^2) - (x-y)^2(x+y)^2}{8x^2} \\ &= \frac{(x-y)^2}{8x^2}(96x^2 - (x+y)^2) \\ &\geq \frac{(x-y)^2}{8x^2}(96x^2 - 2(x^2+y^2)) \\ &\geq \frac{(x-y)^2}{8x^2}(96x^2 - 2(x^2+x^2)) \\ &\geq 11.5(x-y)^2 \\ &\geq 0 \end{aligned}$$

Therefore, we have

$$2x^2\varepsilon^2 - (x^2 - y^2)\varepsilon + 12(x-y)^2 \geq 0$$

holds if $x^2 \geq y^2$.
When $y^2 \geq x^2$, the inequality turns to

$$2x^2\varepsilon^2 - (y^2 - x^2)\varepsilon + 12(x-y)^2 \geq 0$$

if $y^2 \geq 5x^2$, we have

$$\begin{aligned}\frac{y^2 - x^2}{2 \cdot 2x^2} - 1 &= \frac{y^2 - x^2 - 4x^2}{4x^2} \\ &= \frac{y^2 - 5x^2}{4x^2} \\ &\geq 0\end{aligned}$$

So the minimum value of the function reaches when ε is close enough to 1. And because

$$\begin{aligned}2x^2 - (y^2 - x^2) + 12(x-y)^2 &= 36x^2 + 11y^2 - 24xy \\ &= 36(x^2 - \frac{2}{3}xy + \frac{1}{9}y^2) + 7y^2 \\ &= 36(x - \frac{1}{3}y)^2 + 7y^2 \\ &\geq 0\end{aligned}$$

we have

$$2x^2\varepsilon^2 - (y^2 - x^2)\varepsilon + 12(x-y)^2 \geq 0$$

holds when $y^2 \geq 5x^2$.
If $x^2 \leq y^2 \leq 5x^2$, we have

$$\begin{aligned}\frac{y^2 - x^2}{2 \cdot 2x^2} - 1 &= \frac{y^2 - x^2 - 4x^2}{4x^2} \\ &= \frac{y^2 - 5x^2}{4x^2} \\ &\leq 0\end{aligned}$$

the minimum value of the function reaches when $\varepsilon = \frac{y^2 - x^2}{4x^2}$. And the minimum value of the function is

$$\begin{aligned}\frac{4 \cdot 2x^2 \cdot 12(x-y)^2 - (y^2 - x^2)^2}{4 \cdot 2x^2} &= \frac{96x^2(x-y^2) - (x-y)^2(x+y)^2}{8x^2} \\ &= \frac{(x-y)^2}{8x^2}(96x^2 - (x+y)^2) \\ &\geq \frac{(x-y)^2}{8x^2}(96x^2 - 2(x^2 + y^2)) \\ &\geq \frac{(x-y)^2}{8x^2}(96x^2 - 2(y^2 + y^2)) \\ &\geq \frac{(x-y)^2}{8x^2}(96x^2 - 4y^2) \\ &\geq 9.5(x-y)^2 \\ &\geq 0\end{aligned}$$

So we have

$$2x^2\varepsilon^2 - (y^2 - x^2)\varepsilon + 12(x-y)^2 \geq 0$$

holds when $x^2 \leq y^2 \leq 5x^2$.
Q.E.D. □

Exercise 5 *30 分*
考虑 k-means 问题。令 $A = \{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ 为一个含有 n 个点的集合。对于 A 的任意一个 k-划分 $C_1, \cdots, C_k$，定义

$$D(A, \{C_i\}_{i=1,\ldots,k}) := \sum_{i=1}^{k} \sum_{a \in C_i} \|a - \mu(C_i)\|^2,$$

这里的 $\mu(C_i) = \frac{1}{|C_i|} \sum_{a \in C_i} a$。
令 $\varepsilon \in (0,1)$。令 $k' \geq \Omega(\frac{\log n}{\varepsilon^2})$ 为 JL 引理 (Johnson-Lindenstrauss Lemma) 中将 A 中的点通过随机投影降维之后的维度。
证明存在一个线性映射 $f : \mathbb{R}^d \to \mathbb{R}^{k'}$ 满足对于 A 的所有的 k-划分 $C_1, \cdots, C_k$，下面的式子成立：

$$|D(A, \{C_i\}_{i=1,\ldots,k}) - D(f(A), \{f(C_i)\}_{i=1,\ldots,k})| \leq \varepsilon \cdot D(A, \{C_i\}_{i=1,\ldots,k}),$$

这里 $f(C_i) = \{f(x) \mid x \in C_i\}$，$f(A) = \{f(x) \mid x \in A\}$。这里的 f 是 A 与 $f(A)$ 之间的双射。

证明. Let the linear projection f be defined as the random projection in JL-Lemma. Since it is is linear, we have

$$\begin{aligned}
\mu(f(C_i)) &= \frac{1}{|C_i|} \sum_{a \in f(C_i)} a \\
&= \frac{1}{|C_i|} \sum_{a \in C_i} f(a) \\
&= f(\frac{1}{|C_i|} \sum_{a \in C_i} a) \\
&= f(\mu(C_i))
\end{aligned}$$

In other words, $\mu(f(C_i))$ is exactly the same as the result if the k centroids are projected by the same f defined above. Consider that the projection f is performed on the union of A and the k centroids. According to the JL-Lemma, with probability $\geq 1 - \frac{3}{2(n+k)}$,$\forall x_i, x_j \in A \cup \{\mu(C_i) | 1 \leq i \leq k\}$, we have

$$|\|x_i - x_j\|^2 - \|f(x_i) - f(x_j)\|^2| < \varepsilon \|x_i - x_j\|^2$$

Moreover, because

$$\begin{aligned}
D(f(A), \{f(C_i)\}_{i=1,\ldots,k}) &= \sum_{i=1}^{k} \sum_{a \in f(C_i)} \|a - \mu(f(C_i))\|^2 \\
&= \sum_{i=1}^{k} \sum_{a \in C_i} \|f(a) - \mu(f(C_i))\|^2 \\
&= \sum_{i=1}^{k} \sum_{a \in C_i} \|f(a) - f(\mu(C_i))\|^2
\end{aligned}$$

there is

$$
\begin{aligned}
|D(A,\{C_i\}_{i=1,\ldots,k}) - D(f(A),\{f(C_i)\}_{i=1,\ldots,k})| &= |\sum_{i=1}^{k}\sum_{a\in C_i}\|a-\mu(C_i)\|^2 - \sum_{i=1}^{k}\sum_{a\in C_i}\|f(a)-f(\mu(C_i))\|^2| \\
&= |\sum_{i=1}^{k}\sum_{a\in C_i}(\|a-\mu(C_i)\|^2 - \|f(a)-f(\mu(C_i))\|^2)| \\
&\leq \sum_{i=1}^{k}\sum_{a\in C_i}|\|a-\mu(C_i)\|^2 - \|f(a)-f(\mu(C_i))\|^2| \\
&\leq \sum_{i=1}^{k}\sum_{a\in C_i}\varepsilon\|a-\mu(C_i)\|^2 \\
&= \varepsilon\sum_{i=1}^{k}\sum_{a\in C_i}\|a-\mu(C_i)\|^2 \\
&= \varepsilon D(A,\{C_i\}_{i=1,\ldots,k})
\end{aligned}
$$

Q.E.D. □