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 $w : S_1 \to \mathbb{R}$ 和集合 S_2 及权重函数 $w : S_2 \to \mathbb{R}$ 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：

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

$\because$ S_1 和 $w_1: S_1 \to R$ 是 A_1 的 (k,ε)-coresets,

$\therefore$ 对所有 $C \subseteq \mathbb{R}^d$, $|C| \leq k$, $|D(A_1, C) - D(S_1, w_1, C)| \leq \varepsilon D(A_1, C)$

其中 $D(A_1, C) = \sum_{x \in A_1} \min_{c \in C} \|x-c\|^2$, $D(S_1, w_1, C) = \sum_{x \in S_1} w(x) \cdot \min_{c \in C} \|x-c\|^2$

$\because$ S_2 和 $w_2: S_2 \to R$ 是 A_2 的 (k,ε)-coresets,

$\therefore$ 对所有 $C \subseteq \mathbb{R}^d$, $|C| \leq k$, $|D(A_2, C) - D(S_2, w_2, C)| \leq \varepsilon D(A_2, C)$

其中 $D(A_2, C) = \sum_{x \in A_2} \min_{c \in C} \|x-c\|^2$, $D(S_2, w_1, C) = \sum_{x \in S_2} w(x) \cdot \min_{c \in C} \|x-c\|^2$

$\because$ A_1、A_2 互不相交,

$$\therefore D(A_1 \cup A_2, C) = \sum_{x \in A_1 \cup A_2} \min_{c \in C} \|x-c\|^2$$

$$= \sum_{x \in A_1} \min_{c \in C} \|x-c\|^2 + \sum_{x \in A_2} \min_{c \in C} \|x-c\|^2$$

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

$$D(S_1 \cup S_2, w_1+w_2, C) = \sum_{x \in S_1 \cup S_2} (w_1+w_2)(x) \min_{c \in C} \|x-c\|^2$$

$$= \sum_{x \in S_1 - S_2} w_1(x) \min_{c \in C} \|x-c\|^2 + \sum_{x \in S_2 - S_1} w_2(x) \min_{c \in C} \|x-c\|^2 + \sum_{x \in S_1 \cap S_2} (w_1(x)+w_2(x)) \min_{c \in C} \|x-c\|^2$$

$$= \sum_{x \in S_1} w_1(x) \min_{c \in C} \|x-c\|^2 + \sum_{x \in S_2} w_2(x) \min_{c \in C} \|x-c\|^2$$

$$= D(S_1, w_1, C) + D(S_2, w_2, C)$$

$\therefore$ 对所有 $C \subseteq \mathbb{R}^d$, $|C| \le k$,

$$\left| D(A_1 \cup A_2, C) - D(S_1 \cup S_2, w_1 + w_2, C) \right|$$

$$= \left| D(A_1, C) + D(A_2, C) - D(S_1, w_1, C) - D(S_2, w_2, C) \right|$$

$$\le \left| D(A_1, C) - D(S_1, w_1, C) \right| + \left| D(A_2, C) - D(S_2, w_2, C) \right|$$

$$\le \varepsilon D(A_1, C) + \varepsilon D(A_2, C)$$

$$= \varepsilon D(A_1 \cup A_2, C)$$

故 $S_1 \cup S_2$ 及 $w_1 + w_2 : S_1 \cup S_2 \to \mathbb{R}$ 是 $A_1 \cup A_2$ 的 (k, ε)-coreset.

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。

(1) 设 k 个中心点 $\{c_1, c_2, \cdots, c_k\} \subseteq R^d$ 是来自整个欧式空间 $\mathbb{R}^d$ 的最优解对应的数据集划分为 $\{C_1, C_2, \cdots, C_k\}$，目标函数 $cost = \sum_{j=1}^{k} \sum_{a \in C_j} D(a, c_j)$

设 k 个中心点 $\{c_1', c_2', \cdots, c_k'\} \subseteq A$ 是来自输入数据集 A 的最优解对应的数据集划分为 $\{C_1', C_2', \cdots, C_k'\}$，目标函数 $cost' = \sum_{j=1}^{k} \sum_{a \in C_j'} D(a, c_j')$

其中 $c_j' = \underset{a \in C_j}{\arg\min}\ D(a, c_j)$，则

$$
\begin{aligned}
cost' = \sum_{j=1}^{k} \sum_{a \in C_j'} D(a, c_j') &\le \sum_{j=1}^{k} \sum_{a \in C_j} D(a, c_j') \\
&\le \sum_{j=1}^{k} \sum_{x \in C_j} \Big(D(a, c_j) + D(c_j, c_j') \Big) \\
&\le \sum_{j=1}^{k} \sum_{x \in C_j} \Big(D(a, c_j) + D(c_j, a) \Big) \\
&= 2\, cost
\end{aligned}
$$

(2) 假设与 (1) 同，但目标函数 $cost = \sum_{j=1}^{k} \sum_{a \in C_j} D^2(a, c_j)$，$cost' = \sum_{j=1}^{k} \sum_{a \in C_j'} D^2(a, c_j')$.

$$
\begin{aligned}
cost' = \sum_{j=1}^{k} \sum_{a \in C_j'} D^2(a, c_j') &\le \sum_{j=1}^{k} \sum_{a \in C_j} D^2(a, c_j') \\
&\le \sum_{j=1}^{k} \sum_{a \in C_j} \Big(D(a, c_j) + D(c_j, c_j') \Big)^2 \\
&\le \sum_{j=1}^{k} \sum_{a \in C_j} \Big(D(a, c_j) + D(c_j, a) \Big)^2 \\
&= 4\, cost
\end{aligned}
$$

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 $c_1, ..., c_k$ 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 $\{1, ..., k\}$ 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 $O(\binom{n}{k}nk)$ 时间内找到最优的聚类。（注意到，$\binom{n}{k} \leq O(n^k)$，而后者在 $k \ll n$ 时远远小于 k^n。）

∵ k个中心点均来自于输入数据集 A，∴ 共有 $\binom{n}{k}$ 种可能的聚类中心，

计算一种聚类的划分及代价需 $O(nk)$ 的时间，

选择其中代价最小的聚类即可.

故可在 $O(\binom{n}{k}nk)$ 时间内找到最优的聚类.

Exercise 4 *20 分*

考虑 k-means 问题。令 $A=\{a_1,\dots,a_n\}\subseteq\mathbb{R}^d$ 为一个含有 n 个点的集合。对于 A 的任意一个 k-划分 $A_1,\dots,A_k$，定义

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

这里的 $\mu(A_i)=\frac{1}{|A_i|}\sum_{a\in A_i}a$。

令 $\varepsilon\in(0,1)$。令 $d'\geq\Omega(\frac{\log n}{\varepsilon^2})$ 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 $f:\mathbb{R}^d\to\mathbb{R}^{d'}$ 满足对于 A 的所有的 k-划分 $A_1,\cdots,A_k$，下面的式子成立：

$$|D(\{A_i\}_{i=1,\dots,k})-D(\{f(A_i)\}_{i=1,\dots,k})|\leq\varepsilon\cdot D(\{A_i\}_{i=1,\dots,k}),$$

这里 $f(A_i)=\{f(x)\mid x\in A_i\}$。这里的 f 是 A 与集合 $f(A)=\{f(x)\mid x\in A\}$ 之间的双射。

由 JL 引理，对 $\forall\ P\subseteq\mathbb{R}^d$，$|P|=n$，$\varepsilon\in(0,1)$，$d'\geq\Omega\left(\frac{\log n}{\varepsilon^2}\right)$，

$\exists$ 线性映射 $f\ \mathbb{R}^d\to\mathbb{R}^{d}$，s.t. 对所有 $x,y\in P$，

$$(1-\varepsilon)\|x-y\|^2<\|f(x)-f(y)\|^2\leq(1+\varepsilon)\|x-y\|^2$$

对 $\forall\ \{C_i\}_{i=1,\cdots,k}$，有 $D(f(A),\{f(A_i)\}_{i=1,\cdots,k})$

$$=\sum_{i=1}^{k}\sum_{a\in A_i}\|f(a)-\mu(f(A_i))\|^2$$

$$\sum_{a\in A_i}\|a-b\|^2=\sum_{a\in A_i}\|a-\mu(A_i)\|^2+|A_i|\cdot\|b-\mu(A_i)\|^2$$

$$\sum_{b\in A_i}\sum_{a\in A_i}\|a-b\|^2=|A_i|\sum_{a\in A_i}\|a-\mu(A_i)\|^2+|A_i|\sum_{b\in A_i}\|b-\mu(A_i)\|^2$$

$$\sum_{a\in A_i}\|a-\mu(A_i)\|^2=\frac{1}{2|A_i|}\sum_{b\in A_i}\sum_{a\in A_i}\|a-b\|^2$$

$$\therefore\ D(A,\{A_i\}_k)=\sum_{i=1}^{k}\sum_{a\in A_i}\|a-\mu(A_i)\|^2=\sum_{i=1}^{k}\frac{1}{2|A_i|}\sum_{b\in A_i}\sum_{a\in A_i}\|a-b\|^2$$

$$\therefore\ |D(A,\{A_i\}_k)-D(f(A),\{f(A_i)\}_k)|$$

$$=\left|\sum_{i=1}^{k}\frac{1}{2|A_i|}\sum_{b\in A_i}\sum_{a\in A_i}\left(\|a-b\|^2-\|f(a)-f(b)\|^2\right)\right|$$

$$\leq\varepsilon\left|\sum_{i=1}^{k}\frac{1}{2|A_i|}\sum_{b\in A_i}\sum_{a\in A_i}\|a-b\|^2\right|$$

$$=\varepsilon D(A,\{A_i\}_k)$$

Exercise 5 *20 分*

考虑讲义 Lecture 21 中的算法 JL-DimRed-k-means。假设在算法第 3 步中，我们对降维之后的数据集合 $f(A)$ 使用的是一个 α-近似算法（针对 $f(A)$ 上的 partition-based k-means 问题）。证明 JL-DimRed-k-mean 算法以很大的概率，输出的解是（关于原数据集 A 上 partition-based k-means 问题）最优解的 $(\alpha+\varepsilon)$-近似。

对 JL-DimRed-k-means，$\varepsilon' = \frac{\varepsilon}{b\alpha}$，$d' = \Omega\left(\frac{\log n}{\varepsilon^2}\right)$

设 JL-DimRed-k-means 输出的解对应的划分是 $\{A_i\}_k$，

原数据集 A 上 partition-based k-means 的最优解是 $\{A_i^*\}_k$

由 exercise 4，有 $|D(A, \{A_i\}_k) - D(f(A), \{f(A_i)\}_k)| \le \varepsilon' D(A, \{A_i\}_k)$

$-\varepsilon' D(A, \{A_i\}_k) \le D(A, \{A_i\}_k) - D(f(A), \{f(A_i)\}_k) \le \varepsilon' D(A, \{A_i\}_k)$

$\Rightarrow \quad D(f(A), \{f(A_i)\}_k) \le (1+\varepsilon') D(A, \{A_i\}_k)$

$\qquad D(A, \{A_i\}_k) \le \frac{1}{1-\varepsilon'} D(A, \{A_i\}_k)$

$$D(f(A), \{f(A_i^*)\}_k) = \min D(f(A), \{f(A_i)\}_k)$$
$$\le \min (1+\varepsilon') D(A, \{A_i\}_k) = (1+\varepsilon') D(A, \{A_i^*\}_k)$$

由 α-近似算法 $\quad D(f(A), \{f(A_i)\}_k) \le \alpha D(f(A), \{f(A_i^*)\}_k)$

$$\therefore D(A, \{A_i\}_k) \le \frac{1}{1-\varepsilon'} D(A, \{A_i\}_k)$$
$$\le \frac{\alpha}{1-\varepsilon'} D(f(A), \{f(A_i^*)\}_k)$$
$$\le \frac{\alpha(1+\varepsilon')}{1-\varepsilon'} D(A, \{A_i^*\}_k)$$

将 $\varepsilon' = \frac{\varepsilon}{b\alpha}$ 代入，$\quad D(A, \{A_i\}_k) \le \frac{\alpha(b\alpha+\varepsilon)}{b\alpha-\varepsilon} D(A, \{A_i^*\}_k)$

$$\frac{\alpha(b\alpha+\varepsilon)}{b\alpha-\varepsilon} = \frac{\alpha(b\alpha-\varepsilon+2\varepsilon)}{b\alpha-\varepsilon} = \left(\alpha + \frac{2\varepsilon}{b-\frac{\varepsilon}{\alpha}}\right)$$

$\because \alpha \ge 1$，$0 < \varepsilon < 1$，$\therefore 0 < \frac{\varepsilon}{\alpha} < 1 \quad \therefore \frac{2}{b-\frac{\varepsilon}{\alpha}} \le 1$

$$\therefore D(A, \{A_i\}_k) \le (\alpha+\varepsilon) D(A, \{A_i^*\}_k)$$

Exercise 6 附加题 *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：

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

设 $x = a-c$，$y = b-c$，则原不等式可写为

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

设 $y = kx$，则原不等式可写为

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

即证 $|1-k^2| \leq \frac{12}{\varepsilon}(1-k)^2 + 2\varepsilon$

① $1-k^2 \geq 0$，即 $-1 \leq k \leq 1$

即证 $\frac{12}{\varepsilon}(k^2-2k+1) + 2\varepsilon \geq 1-k^2$

$$\Leftrightarrow \left(\frac{12}{\varepsilon}+1\right)k^2 - \frac{24}{\varepsilon}k + \frac{12}{\varepsilon} + 2\varepsilon - 1 \geq 0$$

$$\Delta = \left(-\frac{24}{\varepsilon}\right)^2 - 4\times\left(\frac{12}{\varepsilon}+1\right)\times\left(\frac{12}{\varepsilon}+2\varepsilon-1\right) = -8\varepsilon - 92 < 0$$

又 $\varepsilon \in (0,1)$，故 $\Delta < 0$，故不等式成立.

② $1-k^2 < 0$，即 $k > 1$ 或 $k < -1$

即证 $\frac{12}{\varepsilon}(k^2-2k+1) + 2\varepsilon \geq k^2 - 1$

$$\Leftrightarrow \left(\frac{12}{\varepsilon}-1\right)k^2 - \frac{24}{\varepsilon}k + \frac{12}{\varepsilon} + 2\varepsilon + 1 \geq 0$$

$$\Delta = \left(-\frac{24}{\varepsilon}\right)^2 - 4\left(\frac{12}{\varepsilon}-1\right)\left(\frac{12}{\varepsilon}+2\varepsilon+1\right) = 8\varepsilon - 92 < 0$$

又 $\varepsilon \in (0,1)$，故 $\Delta < 0$，故不等式成立

综上，不等式得证.