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：

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

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。

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。

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：

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

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。
