

《大数据算法》作业

2023 年春

截止日期: 2023 年 6 月 20 日 17:59

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 $w : S_1 \rightarrow \mathbb{R}$ 和集合 S_2 及权重函数 $w : S_2 \rightarrow \mathbb{R}$ 分别是 A_1 和 A_2 的 (k, ε) -coresets。那么 $S_1 \cup S_2$ 及函数 $w_1 + w_2 : S_1 \cup S_2 \rightarrow \mathbb{R}$ 是 $A_1 \cup A_2$ 的 (k, ε) -coreset。

注: 这里 $w_1 + w_2$ 的定义如下:

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

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, \dots, 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, \dots, 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, \dots, 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 $\{1, \dots, k\}$ 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 $O(\binom{n}{k})$ 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 。)

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 $f: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ 满足对于 A 的所有的 k -划分 A_1, \dots, 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。

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。

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：

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