

大数据算法 HW2

PB18111697 王章瀚

2021 年 5 月 31 日

1.

我们对一个 d 维欧氏空间的点集做 k -center 聚类. 假设 d 和 k 都为常数. 任给 $\epsilon > 0$, 我们是否能给出一个 $(1 + \epsilon)$ 倍近似比解? (35分)

本题参考 *Approximate Clustering via Core-Sets*¹ 中提到的算法完成.

算法说明

Algorithm 1 $(1 + \epsilon)$ -approximation for k -center

```
1: function SUB- $k$ -CENTERS( $\{S_i\}$  : current coreset;  $P$  : Remaining point set;  $k$  : #centers)
2:   for  $i = 1$  to  $O(1/\epsilon^2)$  do
3:     for  $j = 1$  to  $k$  do
4:       Let  $B_j(c_j, r_j) = \text{MEB}(S_j)$ 
5:        $p = \text{argmax}_{p \in P} (\min(\|p - c_j\|, j = 1, \dots, k))$ 
6:       return the best solution of SUB- $k$ -CENTERS( $\{S_1, \dots, S_{j-1}, \{p\} \cup S_j, S_{j+1}, \dots, S_k\}$ ,  $P/\{p\}$ ,  $k$ ),  $\forall j = 1, \dots, k$ 
```

近似比证明

其证明的核心是下面这个引理:

Lemma 1

存在一个子集 $S \subseteq P$, $|S| = O(1/\epsilon^2)$, 使得 S 的 MEB 的半径至少是 $\frac{1}{1+\epsilon}$ 倍 P 的 MEB 半径.

其该证明可以在论文的 Lemma2.3 中找到.

近似比

有了 Lemma 1, 就能够保证上述算法返回的半径是 $(1 + \epsilon)r$ 的. 但具体细节我其实没太弄懂, 并且该论文只证明了 1-center 和 2-center 的情况, 对于 k -center 的情况他表示容易扩展, 但我没有想明白怎么扩展, 就不在这里伪证了.

¹ *Approximate Clustering via Core-Sets*: <http://graphics.stanford.edu/courses/cs468-06-winter/Papers/BHI02.pdf>

另一种方法

这是和同学讨论的一个方法, 有必要记录一下.

算法步骤如下:

1. 先对原有数据集 P 做冈萨雷斯算法, 得到一个近似比为 2 的结果 r^* ,
2. 然后考虑在包围所有点的一个有限范围内, 将 d 维空间划分为边长为 $\frac{\epsilon r^*/2}{n\sqrt{d}}$ 的超立方体.
3. 遍历超立方体的每个顶点构成集合的 k -子集, 找出满足目标函数的最有情况, 并作为最优解返回即可.

近似比证明

每个超立方体边长是 δ . 那么考虑点 p 的簇中心是 c , 则对任意 c 所在超立方体的顶点 q , 有

$$\|p - q\| \leq \|p - c\| + \delta\sqrt{d}$$

所以, 对所有点来说, 有

$$\max(\|p - q\|) \leq \max(\|p - c\| + n\delta\sqrt{d}) \leq \max \|p - c\| + n\delta\sqrt{d} \leq r_{opt} + n\delta\sqrt{d}$$

要想满足 $(1 + \epsilon)$ 的近似比, 只要 $n\delta\sqrt{d} \leq \epsilon r_{opt}$ 即可. 虽然我们无法得知 r_{opt} 是多少, 但我们可以用冈萨雷斯算法得到一个不超过 $2r_{opt}$ 的结果, 将这个结果的半径除以二就能得到一个足够小的半径, 以估计 δ .

假设冈萨雷斯算法给出 r^* , 那么只要 $\delta \leq \frac{\epsilon r^*/2}{n\sqrt{d}}$ 即可满足 $(1 + \epsilon)$ 的近似比.

2.

对于 k -means 聚类, 如果 k 为常数, 且我们假设在最优解中, 每一个 cluster 大小的下限为 αn (n 为点的个数, $0 < \alpha < 1/k$), 我们能否通过简单的均匀采样得到一个具有常数近似比的初始解? (35分)

本题参考 *Fast k-Means Algorithms with Constant Approximation*² 中的 Algorithm1 完成. 其主要思想是用抽样的结果来估计指定的 k 个簇中心给出的目标函数值, 从而在减少计算量的前提下, 给出常数倍近似比.

算法内容

考虑输入原始数据点集为 P , 若每个 cluster 的大小下限为 αn , 则只要从 P 中做大小为 $\frac{8}{\epsilon\alpha}$ 的随机采样 T . 遍历 T 上所有 k 大小的点集作为簇中心, 并计算目标函数. 而后将目标函数值最小的 k 大小点集作为最终结果. 这样, 可以至少有 $\frac{1}{12} - \exp(1 - \frac{1}{\epsilon})$ 的概率得到 $(5 + 2\epsilon)$ 的常数近似比.

算法性质证明

Lemma 1

若 T 是原始数据 P 的一个随机抽样, 大小为 $|T|$. 而 μ_P 是 P 的中心, μ_T 是 T 的中心, 那么有至少 $1 - \delta$ ($\delta > 0$) 的概率使下式成立:

$$\sum_{x_i \in P} \|x_i - \mu_T\|^2 \leq (1 + \frac{1}{\delta|T|}) \sum_{x_i \in P} \|x_i - \mu_P\|^2$$

² *Fast k-Means Algorithms with Constant Approximation*: <https://link.springer.com/content/pdf/10.1007%2F11602613.pdf>

证明: 对于不等式左边有:

$$\begin{aligned}\sum_{x_i \in P} \|x_i - \mu_T\|^2 &= \sum_{x_i \in P} \|x_i - \mu_P + \mu_P - \mu_T\|^2 \\ &= \sum_{x_i \in P} \|x_i - \mu_P\|^2 + |P| \|\mu_P - \mu_T\|^2\end{aligned}$$

考虑其第二项, 由 Markov 不等式有:

$$Pr \left[\|\mu_P - \mu_T\|^2 \geq \frac{1}{\delta|P|} E[\|\mu_P - \mu_T\|^2] \right] \leq \delta$$

因此有至少 $(1 - \delta)$ 的概率能够保证:

$$\begin{aligned}\sum_{x_i \in P} \|x_i - \mu_T\|^2 &= \sum_{x_i \in P} \|x_i - \mu_P + \mu_P - \mu_T\|^2 \\ &= \sum_{x_i \in P} \|x_i - \mu_P\|^2 + |P| \|\mu_P - \mu_T\|^2 \\ &\leq \sum_{x_i \in P} \|x_i - \mu_P\|^2 + \frac{1}{\delta} \|\mu_P - \mu_T\|^2 \\ &= \sum_{x_i \in P} \|x_i - \mu_P\|^2 + \frac{1}{\delta|T|} |T| \|\mu_P - \mu_T\|^2 \\ &\leq \sum_{x_i \in P} \|x_i - \mu_P\|^2 + \frac{1}{\delta|T|} \|\mu_P - x_i\|^2\end{aligned}$$

Lemma 2

令 C_T 表示样本中离样本中心最近的点, 那么可以以至少 $\frac{1}{12}$ 的概率保证

$$\sum_{x_i \in P} \|x_i - C_T\|^2 \leq (5 + 2\epsilon) \sum_{x_i \in P} \|x_i - \mu_P\|^2$$

证明: 考虑三角不等式

$$\|x_i - C_T\|^2 \leq 2(\|x_i - \mu_T\|^2 + \|C_T - \mu_T\|^2)$$

对右式第二项求和有

$$\sum_{x_i \in P} \|C_T - \mu_T\|^2 = |P| \|C_T - \mu_T\|^2 \leq \frac{|P|}{|T|} \sum_{x_i \in P} \sum_{x_i \in T} \|x_i - \mu_T\|^2 = |P| Var(T)$$

这里 $Var(T) = \frac{1}{|T|} \sum_{x_i \in P} \|x_i - \mu_T\|^2$. 因为 T 是原始数据 P 的抽样, 根据概率论基础知识有:

$$E(Var(T)) = \frac{|T| - 1}{|T|} Var(P)$$

再由 Markov 不等式可以得到:

$$Pr[Var(T) \leq 1.5 Var(P)] \geq 1 - \frac{|T| - 1}{1.5|T|} > \frac{1}{3}$$

因此, 下式成立的概率至少是 $\frac{1}{3}$:

$$\sum_{x_i \in P} \|C_T - \mu_T\|^2 \leq 1.5|P| Var(P) = 1.5 \sum_{x_i \in P} \|x_i - \mu_P\|^2$$

记该事件成立为事件 A, 而 Lemma 1 的描述的内容(令 $\delta = \frac{1}{4}$) 为事件 B, 那么有:

$$Pr[AB] = 1 - Pr(\bar{A} \cup \bar{B}) \geq 1 - (Pr(\bar{A}) + Pr(\bar{B})) = Pr(A) + Pr(B) - 1 > \frac{3}{4} + \frac{1}{3} - 1 = \frac{1}{12}$$

近似比证明

为了满足上面的不等式, 我们应当能够抽样 T 使得它能包含每个 cluster 的至少 $\frac{4}{\epsilon}$ 个点.

假设 n_s 是最小 cluster S 的大小, 由题设知, $n_s = \alpha|P| = \alpha k \frac{|P|}{k}$.

设 X_s 是 T 中落于最小 cluster S 的点. 根据 Chernoff Bound 的乘积形式, 我们有:

$$Pr \left[X_s \geq \beta \left(\frac{|T|}{|P|} n_s \right) \right] \geq 1 - \exp \left(-\frac{(1-\beta)^2}{2} \left(\frac{|T|}{|P|} n_s \right) \right)$$

也就是

$$Pr [X_s \geq \beta|T|\alpha] \geq 1 - \exp \left(-\frac{(1-\beta)^2}{2} |T|\alpha \right)$$

因此只需要让 $\beta|T|\alpha = \frac{4}{\epsilon}$ 且 $\beta = \frac{1}{2}$, 也就是 $|T| = \frac{8}{\epsilon\alpha}$, 则样本数量满足要求的概率至少有 $1 - \exp(1 - \frac{1}{\epsilon})$

综上所述, 至少有 $\frac{1}{12} - \exp(1 - \frac{1}{\epsilon})$ 的概率能够满足 $(5 + 2\epsilon)$ 的近似比, 即:

$$\sum_{x_i \in P} \|x_i - C_T\|^2 \leq (5 + 2\epsilon) \sum_{x_i \in P} \|x_i - \mu_P\|^2$$

3.

gilbert's 算法的描述是基于欧氏空间. 如果数据经过某个 kernel function 映射到一个新的空间 Π (比如每个点 p 被映射到 $\phi(p) \in \Pi$), 我们能否利用 gilbert's 算法在空间 Π 中计算 polytope distance? (提示: 在空间 Π 中, 我们可以通过 kernel function K 得到任意两点的内积 $K(\phi(p), \phi(q))$) (30分)

在原本的 gilbert 算法中, 我们需要

1. pick $q_0 \in Q$, 使之最接近原点(实际不一定要最近). 令 $x_1 = q_0$
2. 重复以下步骤: 取 $q_i \in Q$, 满足 $proj_{\bar{x}_i}(q_i)$ 最小. 令 x_{i+1} 为直线段 $\overline{q_i x_i}$ 上离原点最近的点.

考虑映射后的点集, 记为 $\phi(P^+)$ 和 $\phi(P^-)$. 将上述 Gilbert 算法稍作改进, 可以改为核形式:

1. pick $q_0 \in Q$, 使 $\phi(q_0)$ 最接近原点(也就是 $K(\phi(q_0), \phi(q_0))$ 最小(实际不一定要最近). 令 $x_1 = q_0$
2. 重复以下步骤: 取 $\phi(q_i) = \arg \min_{q_i \in Q} proj_{\phi(x_i)}(\phi(q_i)) = \frac{K(\phi(q_i), \phi(x_i))}{\sqrt{K(\phi(x_i), \phi(x_i))}}$ 最小.

令 $\phi(x_{i+1})$ 为直线段 $\overline{\phi(q_i)\phi(x_i)}$ 上离原点最近的点.

- 这里的最接近的点 $\phi(x_{i+1})$ 可以这样给出: 由于直线段 $\overline{\phi(q_i)\phi(x_i)}$ 上的点可以表示为 $\phi(q_i), \phi(x_i)$ 的凸组合, 即 $x_{i+1} = \alpha\phi(q_i) + (1-\alpha)\phi(x_i)$.
为使之最小, 可以最小化

$$\begin{aligned} & \|\alpha\phi(q_i) + (1-\alpha)\phi(x_i)\|^2 \\ &= \alpha^2 K(\phi(x_i), \phi(x_i)) + (1-\alpha)^2 K(\phi(q_i), \phi(q_i)) + 2\alpha(1-\alpha)K(\phi(x_i), \phi(q_i)) \end{aligned}$$

这仅仅是一个关于 α 的二次方程, 取最值时有

$$\alpha = \frac{K(\phi(x_i), \phi(q_i)) - K(\phi(q_i), \phi(q_i))}{K(\phi(x_i), \phi(x_i)) + K(\phi(q_i), \phi(q_i)) - 2K(\phi(x_i), \phi(q_i))}$$

迭代足够多次后, polytope distance 在 Π 空间就可以直接由 $\|\phi(x_i)\|_2^2$ 给出.