大数据算法 HW2

PB18111697 王章瀚

2021 年 5 月 30 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)

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***[1] 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:

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

证明: 对于不等式左边有:

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

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

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

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

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：

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

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

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

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 kerenel 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 $proj_{\vec{x}_i}(q_i)$ 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 $\phi(x_{i+1})0$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

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

     这仅仅是一个关于 α 的二次方程, 取最值时有

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

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.