大数据算法 HW3

PB18111697 王章瀚

2021 年 6 月 28 日

1.

参考"Ke Chen: On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications. SIAM J. Comput. 39(3): 923-947 (2009)"[1] 论文, 阐述如何建立关于线性回归的 coreset (假设任意数据 (x, y), 满足 $x \in [0, \Delta]^d, y \in [0, 1]$).

线性回归的 coreset 建立算法. 假设我们有一个粗糙的算法 $\mathbb{A}$ 能够得到 α 倍近似比的线性回归. 那么要想建立这个 coreset, 记这个粗糙解为 $A = (w_0, b_0)$. 令 $\phi = \log_2(\alpha n)$, $d = \frac{1}{\alpha} f(P, A)$, 我们就可以在这条粗糙解对应的超平面 $w_0^T x + b_0 = 0$ 的两边以 $d, 2d, \cdots, 2^\phi d$ 的距离将空间划分开来, 称区域 $(w_0^T + b_0 \in \pm 2^{t+1} d) \backslash (w_0^T + b_0 \in \pm 2^t)$ 为 A_t. 在 A_t 中分别抽取 m 个点, 其集合记为 N_t, 那么 $S = \bigcup\limits_{t=0}^{\phi} N_t$ 即为 coreset. (下面会考虑 m 应当取怎样的值)

首先证明 $\forall p \in P$, $p \in \bigcup\limits_{i=0}^{\phi} N_i$. 用反证法. 若 $\exists p \notin \bigcup\limits_{i=0}^{\phi} N_i$, 因为

$$\begin{aligned} 2^\phi d &= \alpha n \frac{1}{\alpha} f(P, A) \\ &= n \cdot f(P, A) \\ &= \sum_{i=1}^{n} \|w_0^T x + b_0\|^2 \end{aligned}$$

故 $\|w_0^T p + b_0\|^2 > 2^\phi d = \sum\limits_{i=1}^{n} \|w_0^T x + b_0\|^2$ 矛盾, 至此证毕.

下面考虑 m 的取值. 因为 S 是 P 的一个抽样, 所以有

$$E\left[\sum_{p \in S} \|w^T p + b\|^2\right] = \sum_{p \in P} \|w^T p + b\|^2$$

因此若令 $m = \frac{1}{\epsilon_0^2} \log \frac{1}{\lambda}$, 记 $S_t = A_t \cap S$, $N_t = A_t \cap P$, 每一个 $\|w^T p + b\|^2 \in [z, z + 2^{t+1} d]$. 则由 Hoeffding 不等式可得

$$Pr\left(\left|\frac{1}{|S_t|} \sum_{p \in S_t} \|w^T p + b\|^2 - \frac{1}{|N_t|} \sum_{p \in N_t} \|w^T p + b\|^2\right| \le \epsilon_0 \cdot 2^t d\right) \ge 1 - \lambda$$

[1] ***On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications***: https://epubs.siam.org/doi/pdf/10.1137/070699007

亦即

$$Pr\left(\left|\frac{|N_t|}{|S_t|}\sum_{p\in S_t}\|w^Tp+b\|^2-\sum_{p\in N_t}\|w^Tp+b\|^2\right|\le \epsilon_0\cdot 2^t d|N_t|\right)\ge 1-\lambda$$

考虑近似比:

$$\begin{aligned}|f(S,(w,b))-f(P,(w,b))| &= \left|\sum_{t=0}^{\phi}\frac{|N_t|}{|A_t|}\sum_{p\in A_t}\|w^Tp+b\|^2-\sum_{t=0}^{\phi}\sum_{p\in P}\|w^Tp+b\|^2\right| \\ &\le \sum_{t=0}^{\phi}\left|\frac{|N_t|}{|S|}\sum_{p\in S}\|w^Tp+b\|^2-\sum_{p\in P}\|w^Tp+b\|^2\right| \\ &\le \sum_{t=0}^{\phi}\epsilon_0 2^t d|N_t| \\ &\le \sum_{t=0}^{\phi}\epsilon_0\frac{1}{2}\sum_{p\in N_t}\|w_0^Tp+b_0\|^2 \\ &\le \frac{1}{2}\epsilon_0 f(P,(w_0,b_0)) \\ &\le \frac{1}{2}\alpha\epsilon_0 f(P,(w^*,b^*)) \\ &\le \frac{1}{2}\alpha\epsilon_0 f(P,(w,b))\end{aligned}$$

因此要想满足 ϵ 的近似比, 应令 $\epsilon_0=\frac{2\epsilon}{\alpha}$. 为了使对所有区域, 上面不等式都要满足且概率依然是 $1-\lambda$, 前文所述 x 应该改进为 $\frac{1}{\epsilon^2}\log\frac{\phi+1}{\lambda}=O(\frac{1}{\epsilon^2}\log\frac{\log n}{\lambda})$, 这样就有 $(1-\frac{\lambda}{\phi+1})^{\phi+1}\le 1-\lambda$

此时 $m=O(\frac{\alpha^2}{\epsilon^2}\log\frac{\log n}{\lambda})$, 故

$$|S|=O(\frac{\alpha^2}{\epsilon^2}\log\frac{\log n}{\lambda}\times\log n)\ll O(n)$$

满足 coreset 的大小要求.

2.

对于某一优化问题 A, 假设我们可以构造大小为 $f(\epsilon, n)$ 的 $\epsilon - coreset$, 其中 n 为数据大小. 如果利用 merge-and-reduce 方法建立关于流数据的 $\epsilon - coreset$, 所需的内存空间多大? 给出详细计算过程. (参考问题 1 的论文 appendix B)

结论. 所需内存空间应当是 $Space = O\left(M + \sum_{i=1}^{\lceil \log n \rceil} f\left(\frac{\epsilon}{b(i+1)^2}, 2^{i+1}M\right)\right)$. 特别地, 对于 k-median, 应当是 $O\left(\frac{d^2k^2}{\epsilon^2}\log^8 n\right)$ 的. 下面证明之.

首先考虑 coreset 的两个性质:

Coreset 的取并性质. 若 S_1 和 S_2 分别是 P_1, P_2 的 $(\epsilon) - coreset$, 那么 $S_1 \cup S_2$ 就是 $P_1 \cup P_2$ 的 $(\epsilon) - coreset$.

Coreset 的传递性质. 若 S_1 是 S_2 的 $(\epsilon) - coreset$, S_2 是 S_3 的 $(\delta) - coreset$, 那么 S_1 就是 S_3 的 $((1+\epsilon)(1+\delta) - 1) - coreset$.

现在考察 Merge-and-Reduce 方法本身:

Merge-and-Reduce 方法. 对于流数据 $p_1, p_2, \cdots \in \mathbb{R}^d$, 我们用桶 $B_1, B_2, \cdots$ 来存储. 其中 B_0 大小为 $M = \left\lceil \frac{k^2 d}{\epsilon^2} \right\rceil$, B_i 大小为 $2^{i-1}M$. 当 p_m 插入 B_0 后, B_0 没满则完成, 若满了则把 $B_1, \cdots, B_{t-1}$ 合并入 B_t, 这里 B_t 是第一个空桶, 并称这一步为 p_m 触发了 B_t.
考虑每个桶 B_i 有个 Coreset Q_i, 其中 Q_0 即是 B_0 本身. 一旦 p_m 触发了 B_t, 就使 Q_t 为 $\bigcup_{i=0}^{t-1} Q_i$ 的一个 $(\rho_t) - coreset$, 其中置信系数为 $\lambda_m = \frac{\lambda}{m^2}$, $\rho_t = \epsilon / b((t+1)^2)$, b 是一个充分大的常数. 若令 $Q = \bigcup_{i \geq 0} Q_i$, 则我们有以下结论:

Q 以不少于 $1 - \lambda$ 的概率是已处理数据的 $(\epsilon) - coreset$. 其证明如下:
一方面, 通过重复运用 Coreset 的传递性质可以得到: Q_r 是 B_r 的一个 $(\prod_{l=0}^{r}(1+\rho_l) - 1) - coreset$; 又因为当 b 足够大的时候, $\prod_{l=0}^{r}(1+\rho_l) \leq 1 + \epsilon$, 因此 Q_r 是 B_r 的一个 $\epsilon - coreset$. 又由于 $Q = \bigcup_{i \geq 0} Q_i$, 由 coreset 的取并性质可知: Q 是输入数据的一个 $\epsilon - coreset$.
另一方面, 算法每次计算失败的概率是 $\lambda_m = \frac{\lambda}{m^2}$, 且每次到达一个点之多会触发一次. 因此总的失败概率$\leq \sum_{i=M}^{n} \lambda_i = \sum_{i=M}^{n} (\lambda / i^2) \leq \lambda$. 证毕.

流数据的 coreset 所需空间 首先考虑每个 Q_t 的大小. 因为 Q_t 是 $\bigcup_{i=0}^{t-1} Q_i$ 的 $\frac{\epsilon}{b(t+1)^2} - coreset$. 而 $\left|\bigcup_{i=0}^{t-1} Q_i\right| \leq 2^{t+1}M$, 故 $|Q_t| = O(f(\frac{\epsilon}{b(t+1)^2}, 2^{t+1}M))$. 因此, 总的存储空间的通用表达式为

$$Space = M + \sum_{i=1}^{\lceil \log n \rceil} |Q_i| = O\left(M + \sum_{i=1}^{\lceil \log n \rceil} f\left(\frac{\epsilon}{b(i+1)^2}, 2^{i+1}M\right)\right)$$

特别地, 对于 **k-median** 场景, 根据论文的 **Theorem 4.10**, 且在 $\epsilon > 1/n$, $d \leq n$, 并且 k 是个常数的情况下, 有

$$
\begin{aligned}
|Q_i| &= O\left(\frac{ki^4(i+\log M)^2}{\epsilon^2}\left(dk\log\frac{i^2}{\epsilon} + k\log k + k\log(i+\log M) + \log\frac{n^2}{\lambda}\right)\right) \\
&= O\left(\frac{ki^4(i+\log M)^2}{\epsilon^2}\left(dk\log\frac{i^2}{\epsilon} + \log\frac{n^2}{\lambda}\right)\right) \\
&= O\left(\frac{dk^2 i^6 \log(n)}{\epsilon^2}\right)
\end{aligned}
$$

考虑到 $\sum_{i=1}^{\lceil n \rceil} i^7 = O(\log^8 n)$ 且每个数据需要 $O(d)$ 来存储, 故对于 k-median 的流式 coreset 来说, 总的所需空间为:

$$
O(d(M + \sum_{i=1}^{\lceil \log n \rceil} |Q_i|) = O\left(\frac{d^2k^2\log^8 n}{\epsilon^2}\right)
$$

3.

阅读论文 "Artur Czumaj, Christian Sohler: Sublinear-Time Approximation for Clustering Via Random Sampling. ICALP 2004: 396-407"[2], 阐述论文中的方法能不能扩展到问题 1 中的线性回归问题, 并讨论与 coreset 方法的优缺点对比.

根据论文的思想, 其步骤为:

1. 取 P 的一个子集 S
2. 在 S 上运行近似比为 α 的算法 $\mathbb{A}$ 以求得解

且这里要求满足

1. normalization 后 $f(S,(w_{opt},b_{opt}))$ 是 $f(P,(w_{opt},b_{opt}))$ 的一个近似
2. 在 S 中存在一个目标函数值不超过 $\frac{c}{\alpha}f(S,(w_{opt},b_{opt}))$
3. 对于 P 的解空间中的"解" C, 若 $f(P,(w,b)) > cf(P,(w_{opt},b_{opt}))$, 则 $f(S,(w,b)) > cf(S,(w_{opt},b_{opt}))$

如果满足这些条件, 就容易证明这样的抽样能返回 c 近似比的解. 下面针对线性回归做一些推导与证明.

不妨设 $\|w^Tp+b\|^2 \in [0,\Delta]$, 否则可以由旋转等变换得到(有待商榷, 但不影响证明). 由于 S 是 P 的一个抽样, 故由 hoeffding bound 有:

$$Pr\left(\frac{1}{|S|}\sum_{p\in S}\|w^Tp+b\|^2 \geq (1+\xi_1)\frac{1}{|P|}\sum_{p\in P}\|w^Tp+b\|^2\right) \leq \exp\left(-\frac{\frac{1}{|P|}\sum_{p\in P}\|w^Tp+b\|^2\min\{\xi_1,\xi_1^2\}}{3\Delta}\right)$$

若 w_b, b_b 是一个 α' 坏的解, 即

$$\sum_{p\in P}\|w_b^Tp+b_b\|^2 > \sum_{p\in P}\alpha'\|w_{opt}^Tp+b_{opt}\|^2$$

则由前述 hoeffding bound 的结果可知:

$$Pr\left(\frac{1}{|S|}\sum_{p\in S}\|w_b^Tp+b_b\|^2 \geq (1+\xi_1)\alpha'\frac{1}{|P|}\sum_{p\in P}\|w_{opt}^Tp+b_{opt}\|^2\right) \leq \exp\left(-\frac{\frac{1}{|P|}\sum_{p\in P}\|w_b^Tp+b_b\|^2\min\{\xi_1,\xi_1^2\}}{3\Delta}\right)$$

另一方面, 对于在 S 上运行算法 $\mathbb{A}$ 得到的解 w^*, b^*, 同样由 hoeffding bound 知:

$$Pr\left(\frac{1}{|S|}\sum_{p\in S}\|w_{opt}^Tp+b_{opt}\|^2 \geq (1+\xi_2)\frac{1}{|P|}\sum_{p\in P}\|w_{opt}^Tp+b_{opt}\|^2\right) \leq \exp\left(-\frac{\frac{1}{|P|}\sum_{p\in P}\|w_{opt}^Tp+b_{opt}\|^2\min\{\xi_2,\xi_2^2\}}{3\Delta}\right)$$

因此

$$Pr\left(\frac{1}{|S|}\sum_{p\in S}\|{w^*}^Tp+b^*\|^2 \leq (1+\xi_2)\frac{1}{|P|}\sum_{p\in P}\|w_{opt}^Tp+b_{opt}\|^2\right) \geq 1-\exp\left(-\frac{\frac{1}{|P|}\sum_{p\in P}\|w_{opt}^Tp+b_{opt}\|^2\min\{\xi_2,\xi_2^2\}}{3\Delta}\right)$$

[2] Sublinear-Time Approximation for Clustering Via Random Sampling: https://onlinelibrary.wiley.com/doi/epdf/10.1002/rsa.20157

令 $(1+\xi_1)\alpha' = (1+\xi_2)$, 则结合这两个结论可知, 因为坏解会导致上面不等式的不成立概率较大, 因此 w^*, b^* 必然是一个 α' 好的解. 这其中有一些具体数值需要考虑, 但由于在考试周没有时间写了...

另外, 关于这种方法与 coreset 的对比, 主要有以下几点:

1. 直观上, coreset 方法依据一个粗糙的解来进行划分抽样, 这样更有利于抽样子集给出的结果的精度, 也因此能降低需抽样的量, 而本题的方法则主要通过随机抽样来完成, 没能较好地利用数据特征.
2. 但实际上, coreset 的方法中, 所谓的粗糙算法也不容易寻找, 就算有, 也还需要确定一下要运行到什么程度, 什么样的近似比等问题. 这给工程编码带来一定问题. 此外, 即便有了粗糙解, 要作空间划分等操作也比较麻烦. 相反, 本题的方法则能够非常简单地进行均匀采样, 并在理论上有所保证.