大数据算法 HW2

PB18111697 王章瀚

2021 年 5 月 19 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)

考虑通过简单的均匀采样得到一个具有常数近似比的初始解的概率:

$$Pr\left[\sum_{i=1}^{n}\|p_i - C(p_i)\|^2 \le (1+\epsilon)\sum_{i=1}^{n}\|p_i - C^{opt}(p_i)\|^2\right]$$

根据 Markov's 不等式:

$$Pr\left[\sum_{i=1}^{n}\|p_i - C(p_i)\|^2 \le (1+\epsilon)\sum_{i=1}^{n}\|p_i - C^{opt}(p_i)\|^2\right] \ge 1 - \frac{E\left[\sum_{i=1}^{n}\|p_i - C(p_i)\|^2\right]}{(1+\epsilon)\sum_{i=1}^{n}\|p_i - C^{opt}(p_i)\|^2}$$

下面计算 $E\left[\sum_{i=1}^{n}\|p_i - C(p_i)\|^2\right]$:

$$E\left[\sum_{i=1}^{n}\|p_i - C(p_i)\|^2\right] = \sum_{i=1}^{n}E\left(p_i^2 + C^2(p_i) - 2\langle p_i, C(p_i)\rangle\right)$$

$$=$$

参考文献 Fast k-Means Algorithms with Constant Approximation`https://link.springer.com/content/pdf/10.1007%2F11602613.pdf` 所给出的 Algorithm1, 可以做到常数近似比. 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

近似比证明

它的近似比

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)

考虑映射后的点集, 记为 $\phi(P^+)$ 和 $\phi(P^-)$. 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.

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 $\phi(q_i) = \arg\min_{q_i \in Q} proj_{\phi(x_i)}(\phi(q_i)) = \frac{K(\phi(q_i), \phi(x_i))}{\sqrt{K(\phi(x_i), \text{``}\phi(x_i))}}$ 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 空间就由 $\|\phi(x_i)\|_2^2$ 给出, 通过核函数和两点内积的关系就可以推得原空间的 polytope distance.