大数据算法第三次作业

2025 年 5 月 22 日

Problem 1. 把最优传输问题转化为标准线性规划问题，并写出其对偶问题。

Problem 2. 令 $X = \mathbb{R}^d$，并定义 $\mathcal{H}$ 为 X 上的所有轴对齐矩形（axis-parallel boxes）构成的集合。具体来说：

$$\mathcal{H} = \{h_{a,b} \mid a, b \in X\}$$

其中分类器 $h_{a,b}(x)$ 定义为：

$$h_{a,b}(x) = \begin{cases} 1 & \text{如果} a_i \leq x_i \leq b_i \text{ 对所有} i = 1, \ldots, d \text{ 成立}, \\ -1 & \text{其他情况}. \end{cases}$$

比如，在二维平面上，假如 a=（-1，-1），b=（1，1），那么矩形 $h_{a,b}$ 就表示（(-1，-1)，(-1，1)，(1，1)，(1，-1)））四个点构成的矩形。

证明 $\mathcal{H}$ 的 VC dimension 是 2d

Problem 3. 证明平面上的凸多边形的 VC dimension 是无穷

Problem 4. 设 $a_1, a_2, \ldots, a_n$ 为一个符号流，其中每个符号是集合 $\{1, \ldots, m\}$ 中的一个整数。

1. 均匀随机选择：设计一个算法，从流中均匀随机选择一个符号。你的算法需要多少内存?
2. 加权随机选择：设计一个算法，以与 a_i^2 成正比的概率选择一个符号。

Problem 5. 设 k-median 的优化目标为 $f(P, C)$，相应的距离度量为 g。$A = \{a_1, ..., a_m\}$ 是一个 α, β-近似解，即 $f(P, A) \leq \alpha f(P, C_{opt})$，且 $m \leq \beta k$。以每个 α_j 为圆心，使用以 $2^t R$ 为半径的若干同心圆来划分空间，其中 $R = \frac{1}{\alpha} f(P, A)$，$t = 0, 1, ..., \phi$，$\phi = \log \alpha n$。对于每个 j, t，设同心圆环内部的点集为 N_j^t，我们在每个 N_j^t 中都取 $x = \Theta(\frac{1}{\epsilon_0^2} \log \frac{1}{\lambda})$ 个点组成 coreset。那么对一个固定的 k-median 解 C，依一定的概率，我们有 $|\sum_{p \in S_j^t} \frac{|N_j^t|}{x} g(p, C) - \sum_{p \in N_j^t} g(p, C)| \leq \epsilon_0 2^{t+1} R |N_j^t|$ (Hoeffding Bound).

试证明：

$$|\sum_{j=1}^{m} \sum_{t=0}^{\phi} \sum_{p \in S_j^t} \frac{|N_j^t|}{x} g(p, C) - \sum_{j=1}^{m} \sum_{t=0}^{\phi} \sum_{p \in N_j^t} g(p, C)| \leq \Theta(\epsilon_0) n f(P, A)$$

.

（提示：课上证明了若只考虑 $t \geq 1$ 层的同心圆环内的点集 N_j^t，不等式右侧 $\leq 4\epsilon_0 n f(P, A) = \Theta(\epsilon_0) n f(P, A)$，这个结论成立。现在需证明在考虑 $t = 0$ 层的点集 N_j^0 的情况下，该结论依然成立）

Problem 6. 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 $w : S_1 \to \mathbb{R}$ 和集合 S_2 及权重函数 $w : S_2 \to \mathbb{R}$ 分别是 A_1 和 A_2 的 (k, ε)-coresets。那么 $S_1 \cup S_2$ 及函数 $w_1 + w_2 : S_1 \cup S_2 \to \mathbb{R}$ 是 $A_1 \cup A_2$ 的 (k, ε)-coreset。

注： 这里 $w_1 + w_2$ 的定义如下：

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

Problem 7. 假设 $\alpha \in (0, 1]$。假如我们将（基本的）Morris 算法修改如下：

(a) 初始化 $X \leftarrow 0$。

(b) 对于每次更新，以 $\frac{1}{(1+\alpha)^X}$ 的概率使 X 增加 1。

(c) 对于查询，输出 $\hat{n} = \frac{(1+\alpha)^{X-1}}{\alpha}$。

记 X_n 为上述算法中 n 次更新以后的 X。令 $\hat{n} = \frac{(1+\alpha)^{X_n-1}}{\alpha}$。

- 计算 $\mathbb{E}[\hat{n}]$ 和 $\mathrm{Var}[\hat{n}]$ 。
- 假设 $\epsilon, \delta \in (0, 1)$。基于上述算法，给出一个新算法，使得新算法以至少 $1 - \delta$ 的概率输出一个估计值 $\hat{n}$，满足 $|\hat{n} - n| \leq \epsilon n$。说明你的算法的正确性与（最坏）空间复杂度（即算法使用的比特数）。你的算法只需要满足以至少 $1 - \delta'$ 的概率，其最坏空间复杂度为关于 $\frac{1}{\delta}, \frac{1}{\delta'}, \frac{1}{\epsilon}$ 和 $\log \log n$ 的多项式（即 $\mathrm{poly}(\frac{1}{\delta}, \frac{1}{\delta'}, \frac{1}{\epsilon}, \log \log n)$）。（$\alpha$ 是常数）

提示： 这道题的解答参考课上讲的 Morris 算法和 Morris+ 算法的证明

Problem 思考题.

1. 我们介绍了 rectified flow，知道 rectified flow 试图求动态最优传输的概率路径，从而使得扩散过程的推理步数降低。那么，有没有其他的做法，来让扩散方程的梯度项尽量和时间无关？
2. 大语言模型（如 LLaMA、GPT）通常需要在大规模语料上进行微调（例如 instruction tuning 或 domain adaptation）。为了降低计算成本，研究者尝试用 Coreset 方法从上千万条训练数据中选出一个具有代表性的子集（如 5%）进行训练。如果你无法访问模型梯度信息（例如使用 GPT-4 API），你如何设计一个“无监督”的 Coreset 策略？你会如何定义“代表性”？