中国科学技术大学
计算机科学与技术学院

“大数据算法” 作业 3
2023 年春

注: 本次作业**不**计入成绩，无需提交。

习题 1

给定一个由集合 $[n] = \{1, \cdots, n\}$ 中的 m 个整数组成的数据流 $\mathcal{S}$, 令 $\mathrm{HH}_k = \{i \in [n] : f_i > \frac{m}{k}\}$ 为 k-频繁项集, 其中 f_i 为 $\mathcal{S}$ 中元素 i 的频率 (即出现次数)。
修改 Misra-Gries 算法以找到一个集合 H 使得

$$\mathrm{HH}_k(\mathcal{S}) \subseteq H \subseteq \mathrm{HH}_{2k}(\mathcal{S}).$$

你的算法应该只用一遍扫描 (one pass), 并且使用最多 $O(k(\log n + \log m))$ 比特的空间。

习题 2

设 X 和 Y 为有限集, Y^X 表示从 X 到 Y 的所有函数构成的集合。我们将这些函数视为哈希函数。如果对于从函数族 $\mathcal{H} \subseteq Y^X$ 中均匀随机选择的函数 h, 以下属性成立, 我们称 $\mathcal{H}$ 是强 2-全域 (strongly 2-universal) 的:

$$\forall x, x' \in X \ \forall y, y' \in Y \left(x \neq x' \Rightarrow \Pr_h[h(x) = y \wedge h(x') = y'] = \frac{1}{|Y|^2} \right).$$

给定一个由 X 中的元素组成的数据流 $\mathcal{S}$, 并假设 $\mathcal{S}$ 中包含的不同元素最多为 s 个。设 $\mathcal{H} \subseteq Y^X$ 是一个强 2-全域的哈希函数族, 其中 $|Y| = cs^2$ ($c > 0$ 为某个常数)。假设我们使用一个随机函数 $h \in \mathcal{H}$ 来哈希。
证明碰撞 (即 $\mathcal{S}$ 中两个不同的元素哈希到相同位置的事件) 的概率最多为 $1/(2c)$。

习题 3

在严格的十字转门模型 (strict turnstile model) 下, 对于一个来自 $[n]$ 的整数数据流 $\mathcal{S}$, 考虑以下两个问题:
(k, ℓ_2)**-点查询问题**: 给定一个查询 query(i), $1 \leq i \leq n$, 找到一个值 $\tilde{x}_i \in [x_i - \frac{\|\boldsymbol{x}\|_2}{k}, x_i + \frac{\|\boldsymbol{x}\|_2}{k}]$。
(k, ℓ_2)**-频繁项问题**: 给定一个查询 query(), 找到一个集合 $L \subset [n]$ 使得 $|L| = O(k^2)$, 并且如果 $x_i > \frac{\|\boldsymbol{x}\|_2}{k}$ 则 $i \in L$。
假设存在一个解决 (k, ℓ_2)-点查询问题的算法 $\mathcal{A}$, 其失败概率为 $\frac{\delta}{n}$, 并且使用 s 比特的空间。给出一个解决 $(\frac{k}{4}, \ell_2)$-频繁项问题的算法 $\mathcal{A}'$, 其失败概率为 δ, 并且使用尽可能少的空间。

习题 4

在本课程中, 我们学习了一个 KMV 算法 (记为 $\mathcal{A}$), 用来估计数据流中不同元素的个数。算法 $\mathcal{A}$ 的失败概率为 $\frac{1}{3}$。对于该问题, 设计一个新算法, 使得对于任意的 $\delta < \frac{1}{3}$, 其失败概率最多为 δ。你的算法可以使用 $\mathcal{A}$ 作为一个子程序。
