

温冰 10205501432

数据科学算法作业4

$$\begin{aligned}
 3. \text{ 证: } F &= \{(b, 1)\} \\
 F &= \{(b, 1), (a, 1)\} \\
 F &= \{\} \\
 F &= \{(a, 1)\} \\
 F &= \{(a, 1), (d, 1)\} \\
 F &= \{\} \\
 F &= \{(a, 1)\} \\
 F &= \{(a, 1), (f, 1)\} \\
 F &= \{(a, 2), (f, 1)\} \\
 F &= \{(a, 1)\}
 \end{aligned}$$

4. 证: 在  $(\epsilon, \delta)$  近似算法中, 定义伯努利随机变量.

$$Y_i = \begin{cases} 1, & |f_a - f_a| \geq \epsilon \|f\|_2 \\ 0, & \text{其它} \end{cases}$$

选择适当  $k = O(\frac{1}{\epsilon^2})$  s.t.  $P(Y_i = 1) < \frac{1}{3}$

故  $\mu = \sum_{i=1}^k Y_i < \frac{t}{3}$ . 由切诺夫不等式.

$$P\left(\sum_{i=1}^k Y_i > \frac{t}{2}\right) \leq P\left(\sum_{i=1}^k Y_i > (1 + \frac{1}{2})\mu\right) \leq \exp\left(-\frac{\mu}{12}\right) < \delta.$$

$$\text{又 } \mu < \frac{t}{3} \text{ 故 } \exp\left(-\frac{t}{36}\right) \leq \exp\left(-\frac{\mu}{12}\right) < \delta.$$

$$-\frac{t}{36} < \ln \delta \text{ 故 } t = O(\log(1/\delta))$$

5. 证: 由于各哈希函数间相互独立, 故  $E(f_a) = f_a$ . 且

$$\begin{aligned}
 \text{Var}(f_a) &= \text{Var}\left(\frac{1}{t^2 k} \sum_{i=1}^t \sum_{j=1}^k g(a)(l_i)[h_j(a)]\right) \\
 &= \frac{\|f - a\|_2^2}{t^2 k} \text{Var}\left(\sum_{i=1}^t g(a)(l_i)[h_i(a)]\right)
 \end{aligned}$$

$$\text{故 } P[|f_a - f_a| \geq \epsilon \|f\|_2] \leq P[|f_a - f_a| \geq \epsilon \|f\|_2] \\ \leq \frac{\text{Var}(f_a)}{\epsilon^2 \|f\|_2^2} = \frac{1}{k + 2\epsilon^2} < \delta$$

$$\text{故 } k = O\left(\frac{1}{\epsilon^2 \delta}\right)$$

$$\text{又由第四题结论 } t = O\left(\log \frac{1}{\delta}\right)$$

$$\text{故算法整体空间消耗为 } O\left(\frac{\log(1/\delta)}{\epsilon^2 \delta}\right)$$

6. 4. 11)

-	-	-	T	-	F	-	-	$h_1(x)$
-	-	T	-	F	-	-	-	$h_2(x)$
F	-	T	-	-	-	-	-	$h_3(x)$
0	1	2	3	4	5	6	7	

$$f_0 = 1 \quad f_1 = 3 \quad f_2 = 1 \quad f_3 = 2 \quad f_4 = 1 \quad f_5 = 1 \quad f_6 = 1 \\ f_7 = 1 \quad f_8 = 1 \quad f_9 = 3.$$

故 ~~步~~ 频率项为 1 (3次), 9 (3次)

(2) 该算法在估计 "1" "8" "9" 时出现偏差, 其中 "9" 的偏差较大, 为 2. (估计 3 次, 实际 1 次)

(3). 应将计数器个数调整为  $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$ . 此前还应调整  $d$ , s.t.  $P(f_a \leq f_a \leq f_a + \epsilon \|f\|_1) \geq 1 - \delta$ .