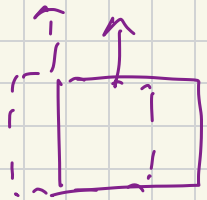
L13: Streaming Frequency Approximation

Oct 1, 2025
Data Mining
Jeff M. Phillips

Input $A = \langle a_1, a_2, \ldots, a_i, \ldots, a_m \rangle$

$a_i \in [U]$

all IP addresses



space limited

space ~ m, U

too big

Goal Summary

- update w/ insertion
- small space poly(log(m), log(U))

counter

label

Input $a_i \in [v]$

e.g. how many packets with IP address j?

frequency $f_j = |\{a_i \in A \mid a_i = j\}|$ $j \in [v]$

F_0 = # distinct items : size of $|\{f_j \mid f_j > 0\}|$

F_1 = total number items m

$F_2 = \sqrt{\sum_j f_j^2}$

typical $F_1 \gg F_2$ ↳ lots of skew

Approx Freq $\hat{f}_j$ for all $j \in [v]$

$|f_j - \hat{f}_j| \leq \varepsilon \cdot m$

$\varepsilon = (0,1)$

$\varepsilon = 0.01$

Sampling for Freq. Approx

$k = \frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}$ samples in stream

$\hat{f}_j = |\{ a_i \in S \mid a_i = j \}| \cdot \frac{m}{k}$

δ = Prob failure

$|\hat{f}_j - f_j| \leq \varepsilon m$ with prob $\geq 1 - \delta$

ε	$1/\varepsilon^2$
0.1	100
0.01	10,000
0.001	1,000,000

MAJORITY

- if some $f_j > \frac{m}{2}$ $\Rightarrow$ output j
- if not $\Rightarrow$ output anything

1 counter 1 label type $A = \{a_1, a_2, \ldots a_i \ldots\}$

3 4 2 2 2 5 2 2 6 2 3 3 2 2

2 3 3 2 2 2 2 2 2 2 2 2 2 2

< 1 0 1 2 3 2 3 4 3 4 3 < 3 4

1. $C = 1$ $L = a_1$
2. for $i = 2$ to m

   if $(a_i = L)$ $C{+}{+}$

   else $C{-}{-}$

   if $(C < 0)$ $L = a_i$; $C = 1$
3. Return L

Frequency Approx — Misra-Gries

k-1 Counters $C[1 \dots k]$ k-1 Labels $L[1 \dots k]$

Set $k = 1/\varepsilon$

0. Init $C[1 \dots k-1] = 0$
   $L[1 \dots k-1] = \phi$

$\hat{f}_j = C[j]$ if $L[j] = j$
ow. $\hat{f}_j = 0$

$$0 \le f_j - \hat{f}_j \le \frac{m}{k} = \varepsilon \cdot m$$

1. for $i = 1$ to m
   - if $\exists (L[j] = a_i)$ $C[j]$ ++
   - else $\exists (L[j] = \phi)$ $L[j] = a_i$; $C[j] = 1$
   - else Decrease all counters
     $\forall j$ $C[j]$ --
     if $[C[j] = 0]$ $L[j] = \phi$

Deletion?

How many times?
↳ $\frac{m}{k}$

Misra-Gries

- Trouble w/ Deletions
- Changes state (not guarantee it order stream changes)
- Deterministic
- Undercount
- Error $\leq \varepsilon m$

$$0 \leq f_j - \hat{f}_j \leq \varepsilon m$$

CountMin

- Handle Deletions (turnstile)
- State is same w/ diff order
- Randomized
- Overcount

$$0 \leq \hat{f}_j - f_j \leq \varepsilon m$$

Count Min Sketch

t hash functions h_j s.t.

$h_j : [v] \to [k]$

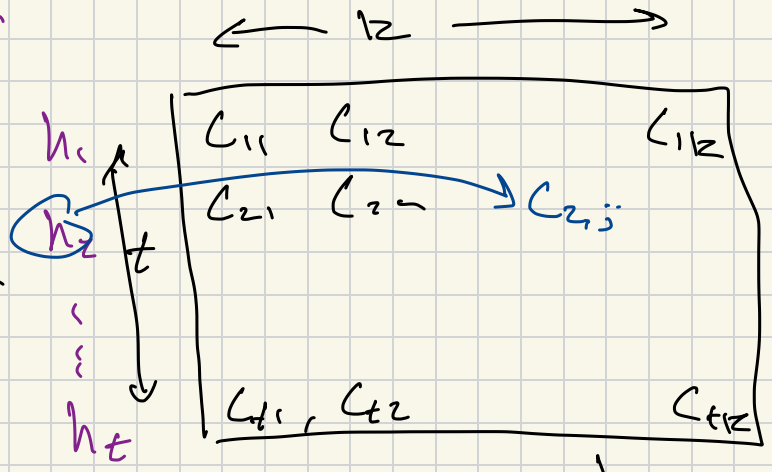
0. Set all $C_{i,j} = 0$ $\forall i,j$

1. for $i = 1$ to m

   for $j = 1$ to t

   $C_{j, h_j(a_i)} ++$

$k \cdot t$ counters



counters

$k = 2/\varepsilon$ $\quad t = \log 1/\delta$

$\hat{f}_q = \min_{j \in [v]} C_{j, h_j(q)}$ w.p. $1-\delta$

$q \in [v]$ $\quad 0 \le \hat{f}_q - f_q \le \varepsilon m$

A - Priori Algorithm

Domain [U]

Market basket

[U] All item is

$R_1 = \{1, 2, 3, 4, 5\}$ $\{6, 7, 9\}$

$R_2 = \{2, 6, 7, 9\}$ $\{3, 6, 9\}$

$R_3 = \{1, 3, 5, 6\}$ $\{6, 7, 8\}$

$\{2, 6, 9\}$

$\{7, 8\}$

$\{1, 2, 6\}$

$\{0, 3, 5, 6\}$

$\{0, 2, 4\}$

$\{7, 4\}$

1st pass

1	2	3	4	5	6	7	8	9
3	5	4	3	3	8	4	2	4

freq items $\geq$ 3 time

(2,3)	(2,6)	(2,7)	(2,9)	(3,6)	(3,7)	(3,9)	(6,7)	(6,9)	(7,9)
1	2	1	2	3	0	1	3	4	2