L12: Streaming

Count-Min Sketch

(and friends)

Stream $A = \langle a_1, a_2, \dots a_n \rangle$ $\underbrace{a_i \in [m]}$

- one pass
- small space

$\Rightarrow S$

n very large but $\log n$ ok count
m very large $\log m$ label

frequency $j \in [m]$

$$f_j = |\{a \in A \mid a = j\}|$$

f_j	3	3	2	1	ϕ
j	a	b	c	d	e

a a b b c a e b c d
$n=10$

$$F_1 = \sum_j f_j = \text{total count } n$$

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

typically $F_1 \gg F_2$

$$F_0 = \sum_j f_j^0 = \#\text{distinct items}$$

$\varepsilon F_1 \gg \varepsilon F_2$

Frequency Approximation

$\forall j \in [m] \to \hat{f}_j$ so

$F_1 = n$

$\to \varepsilon F_1 = \varepsilon^{\%}$

$|f_j - \hat{f}_j| \le \varepsilon F_1$

$\le \varepsilon F_2$

size $\frac{1}{\varepsilon}$ poly$(\log n, \log m)$

$\frac{1}{\varepsilon^2} \cdot$ polylog(n, m)

sketch $S(A)$

data structure

- insert (a_i)
- query $(q \in [m]) \Rightarrow \hat{f}_q$

tradeoff : space $(S(A))$ vs. accuracy ε

CMS: $\boxed{f_j \le \hat{f}_j \le f_j + \varepsilon n}$ w.p. $> 1-\delta$

δ prob of failure

CS: $f_j - \varepsilon F_2 \le \hat{f}_j \le f_j + \varepsilon F_2$

Count-Min Sketch

$t \cdot (k = 2/\varepsilon)$ counters

$t = \log \frac{1}{\delta}$ hash functions $\quad h_j : [m] \to [k]$

a

b

h_1	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$	…	$C_{1,k}$
h_2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	…	$C_{2,k}$
⋮	⋮				
h_t	$C_{t,1}$	$C_{t,2}$	$C_{t,3}$	…	$C_{t,k}$

<u>insert</u> $a \in A \quad a \in [m]$

for $j = 1$ to t

$\quad C_{j,h_j(a)}$++

for each row
hash to counter
increment

<u>query</u> $q \in [m]$

$$\hat{f}_q = \min_{j \in [t]} C_{j,h_j(q)}$$

- $f_q \le \hat{f}_q$: each q always hashes to same $C_{j, h_j(q)}$, only increments

- $\hat{f}_q \le f_q + \boxed{W}$: say $W \le \varepsilon n = \varepsilon F_1 = \varepsilon \sum_j f_j$

  w.p. $> 1-\delta$

1 row / hash fxn

$s \in [m]$ $\quad Y_s = \begin{cases} f_s & \text{if } h(s) = h(q) \quad \text{w.p. } 1/k \\ 0 & \text{otherwise} \end{cases}$

$(k = 2/\varepsilon)$ $\quad X = \sum_{\substack{s \in [m] \\ s \neq q}} Y_s \qquad E[X] = E\left[\sum_s Y_s\right]$

$= \sum_s E[Y_s] = \sum_s \frac{f_s}{k}$

$\le F_1/k$

Expected overcount 1 hash fxn

$\le F_1/k = \frac{\varepsilon n}{2}$ $\Rightarrow \Pr\left[\begin{array}{c}\text{one counter} \\ W > \varepsilon n\end{array}\right] \le 1/2$

Markov Ineq $\quad$ R.V. $X \ge 0$ $E[X] = \mu$

$\Pr[X > \alpha] = \frac{\mu}{\alpha} = \frac{\varepsilon n/2}{\varepsilon n} = \frac{1}{2}$

t hash functions

$t = \log_2\left(\frac{1}{\delta}\right)$

1 hash function h_j

$$p_j = \Pr[w_j > \varepsilon n] \leq 1/2$$

$h_1, \ldots, h_t$ choosen $h_j \in \mathcal{H}$ independently

t hash functions

$$w = \min_j w_j \quad \text{only} \quad w > \varepsilon n$$

if all $w_j > \varepsilon n$

$$\Pr[w > \varepsilon n] = p_j^t \leq 1/(2^t) = 1/2^{\log_2(1/\delta)} = \delta$$

Count Sketch

$k = 4/\varepsilon^2$ $\quad t = \log_2(2/\delta)$

because of median

both chosen randomly from hash family

t : hash fxn $\quad h_j : [m] \to [k]$

t : sign hash fxn $\quad s_j\ [m] \to \{-1, +1\}$

$$
\begin{array}{cc|cccc|}
s_1 & h_1 & C_{1,1} & C_{1,2} & \cdots & C_{1,k} \\
 & \vdots & & & & \\
s_t & h_t & C_{t,1} & & \cdots & C_{t,k}
\end{array}
$$

insert $\quad a \in A, \quad a \in [m]$

for $j = 1$ to t

either add or subtract

$$C_{j,h_j(a)} = C_{j,h_j(a)} + s_j(a)$$

$$|f_j - \hat{f}_j| \le \varepsilon F_2$$

w.p. $\ge 1-\delta$

query $\quad q \in [m]$

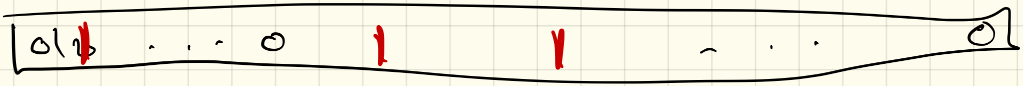
$$\hat{f}_q = \underset{j \in [m]}{\text{median}} \left((C_{j,h_j(q)})(s_j(q)) \right)$$

Bloom Filters

Set data structure B

- if $q \in [m]$ is in B then reports always
- if $q \notin B$, usually says not in B
  ↳ false positives

k bits



t hash functions $h_i : [m] \to [k]$

insert

for $\hat{a} \in [t]$: set $B_{h_i(a)} = 1$

query

if all $h_i(q) = 1 \to$ yes

Apriori Algo.

Frequent Itemsets

Stream / Set $A = \{a_1, a_2, \ldots a_n\}$

$a = \text{set } \{b_1, b_2, \ldots b_t\}$

$b_j \in [m]$