L15 : CountMin Sketch

(and friends)

March 3, 2025

Jeff M. Phillips | Data Mining

Stream

$A = \langle a_1, a_2, a_3, \dots a_i, \dots a_n \rangle$

$a_i \in [m]$ ← domain

Compute Statistics on A

- small space
- one pass

n too large — $\log n$ ← counter

m too large — $\log m$ ← label

S sketch(A)

frequency $j \in [m]$

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

$$F_1 = \sum_j f_j = n$$

$$F_2 = \sqrt{\sum_j f_j^2} \qquad F_2 << F_1$$

$F_0 =$ number of distinct elements in A.

Refresh

Frequency Approximation

$\forall j \in [m] \rightarrow \hat{f}_j$ so $|f_j - \hat{f}_j| \leq \varepsilon n = \varepsilon F_1$

$\leq \varepsilon F_2$ (Count Sketch)

MG: $f_j - \varepsilon n \leq \hat{f}_j \leq f_j$

CountMin $f_j \leq \hat{f}_j \leq f_j + \varepsilon n$

w.p. $1-\delta$

$k = 1/\varepsilon$

$k-1$ counters

$k-1$ labels

C_1	C_2	C_3	...	C_{k-1}
L_1	L_2	L_3	...	L_{k-1}

$a_i \in A$

if $a_i = L_j$

else: if $(C_j = 0)$ → else decrease all counters

$L_j = a_i \quad C_j = 1$

Sketch Data Structure

Data Structure $S(A)$

- update $S(A)$ w/ a_i
- query $S(A)(j) \Rightarrow \hat{f}_j$

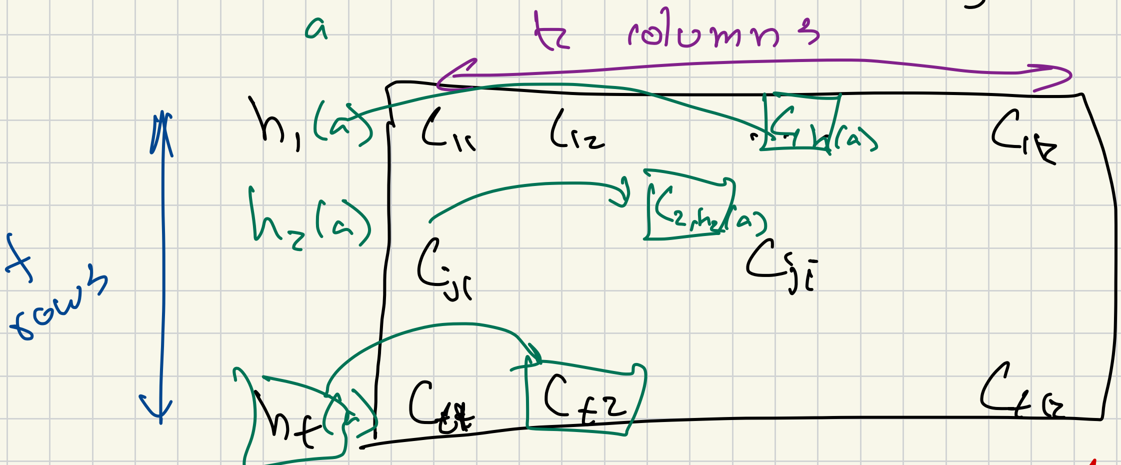
Count Min Sketch

randomness

$h_j \sim \mathcal{H}$

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

$t = \log_2 \frac{1}{\delta}$ hash functions $\quad h_j : [m] \to [k]$



uniform

$\forall j \in [t]$

$C_{j, h_j(q)} \geq f_q$

$q \in [m]$

insert $(a_i \in A)$

for $j=1$ to t

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

query (q) $\Longrightarrow \hat{f}_j$

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

<u>Show</u> CM $f_g \le \hat{f}_g \le f_g + \varepsilon n$ w.p. $\ge 1-\delta$

$f_g \le \hat{f}_g$ each counter $C_{j,h_j(g)}$ includes count of f_g

$\hat{f}_g \le f_g + \boxed{w_g}$ $\quad w \le \varepsilon n = \varepsilon F_1 = \varepsilon \cdot \sum_j f_j$ $\quad \boxed{k = 2/\varepsilon}$

Some $s \in [m]$ $\quad w_g(s) = Y_s = \begin{cases} f_s & \text{if } h_j(s) = h_j(g) \text{ w.p. } 1/k \\ 0 & \text{o.w.} \end{cases}$

some $j \in [t]$

total overcount $w = \sum_{s \in [m]} w(s) = \sum_{s \in [m]} Y_s = X$

$E[X] = E[\sum_s Y_s] = \sum_s E[Y_s] = \sum_s f_s/k = \frac{1}{k} \sum_s f_s = \frac{1}{k} F_1 = \frac{\varepsilon n}{2}$

<u>Markov Ineq</u>

R.V. $X \ge 0$

$Pr(X > \alpha) < \frac{E(X)}{\alpha}$ $\quad \Big| \quad \alpha = E[X] \cdot 2 = \varepsilon n$ $\quad \Rightarrow Pr[X > \varepsilon n] \le \frac{E(X)}{E[X] \cdot 2} = \frac{1}{2}$

t hash functions $\qquad t = \log_2(1/\delta)$

1 hash $\quad h_j \sim \mathcal{H}$

$$\Pr[w_j = X > \varepsilon n] \leq \frac{1}{2}$$

$$\Pr[\text{all hash } h_j \text{ has } w_j > \varepsilon n] = \left(\frac{1}{2}\right)^t$$

t independent hash functions

$$\left(\frac{1}{2}\right)^t = \left(\frac{1}{2}\right)^{\log_2(1/\delta)} = 2^{-\log_2(1/\delta)} = 2^{\log_2(\delta)}$$

$$\delta = \frac{1}{2^{10}} = \frac{1}{1024} \qquad = \delta = \frac{1}{32}$$

$$t = 10 \qquad \Rightarrow t = 5$$

Compare MG vs Count Min

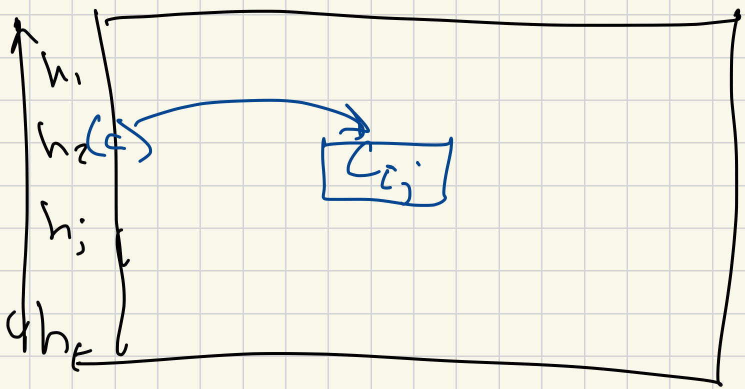
	MG	Count Min
Space	$1/\varepsilon$ counters + labels	$e/\varepsilon \cdot \log_2 \frac{1}{\delta}$ counters + $\log \frac{1}{\delta}$ hash fxn
	Deterministic	Randomized w.p. $1-\delta$
Bias:	under count	over count.
	heavy hitters	most query guesses of MG
Deletions	X	can handle deletions (linear sketch)

Count sketch

unbiased εF_2

k columns

t hash functions



$k = 4/\varepsilon^2$

$t = 2 \log_2 \frac{1}{\delta}$

$h_j \sim \mathcal{H}$ $\quad h_j : [m] \to [k]$

sign $s_j \sim \mathcal{S}$ $\quad s_j : [m] \to \{-1, +1\}$

insert(a_i)

for $j = 1$ to t

$C_{j,h_j(a_i)} \mathrel{+}= s_j(a_i)$

$E[C_{ji}] = 0$

query(q) $\quad q \in [m]$

$\hat{f}_q = \text{median}(C_{j,h_j(q)})$

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

$E[\hat{f}_j] = f_j$