Streaming : Frequent Items (etc)

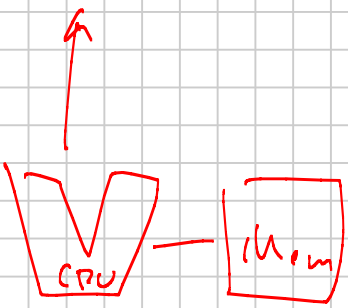
Note Title 2/24/2016

- Count-Min Sketch
  Count Sketch
- Apriori Algorithm (Frequent Itemsets)
- Bloom Filters

Streaming $A = \langle a_1, a_2, \ldots, a_i, \ldots, a_m \rangle$ (A_i)

$a_i \in [n]$

$f_j = |\{a_i \in A \mid a_i = j\}|$, $j \in [n]$



Data Structure

$S_A(j) \to \hat{f}_j$

$f_j \le \hat{f}_j \le f_j + \varepsilon m$

($\hat{f}_j \le f_j$ always true; $\hat{f}_j \le f_j + \varepsilon m$ true w.p. $\ge 1-\delta$)

hashing

$h_r : [n] \to [k]$

$h_r \sim \mathcal{H}$

$r = \{1, 2, \ldots, t\}$

$h_1(a_i) \to j \in [k]$

Count-Min Sketch

Update

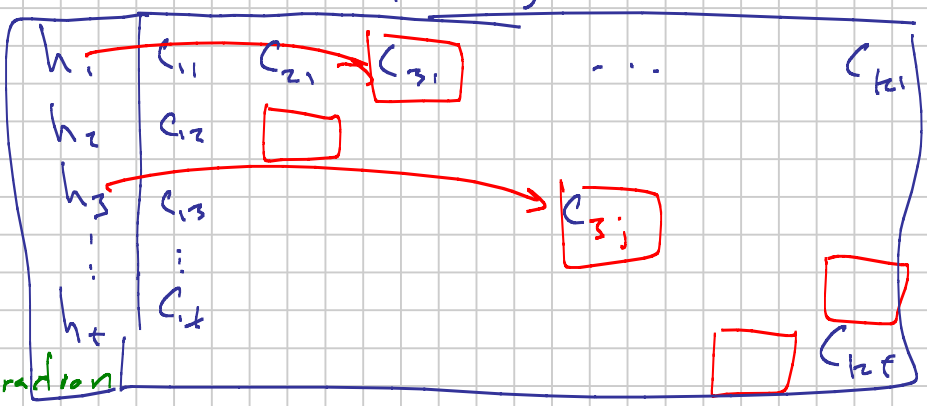
for $i = 1$ to m

for $r = 1$ to t

$C_{r\, h_r(a_i)}$ ++

j ; ++ or −− on subtraction

Init: $C_{rj} = 0$



Query

$S_A(q) = \hat{f}_q$

$k = \frac{2}{\varepsilon}$ $\quad t = \log_2 \frac{1}{\delta} \approx 5 - 10$

$\arg\min_{r \in [t]} C_{r, h_r(q)} \rightarrow f_q \le \hat{f}_q \le f_q + w$

If too large, i.e.

Random Variable $Y_{rj} \equiv$ Amount of counts from q (which hash) in $C_{r h_r(q)}$ (query)

$[n] \setminus q$

$$Y_{rj} = \begin{cases} f_j & \text{with prob } 1/k \\ 0 \end{cases}$$

$E[Y_{rj}] = f_j / k$

$$E\left[\sum_{j \in [n] \setminus q} Y_{rj}\right] = \sum_j E[Y_{rj}] = \frac{\sum_j f_j}{k} = \frac{m}{k} \le \frac{\varepsilon m}{2}$$

Expected overcount in row r is $\frac{m}{k} = \frac{\varepsilon m}{2}$

Markov Inequality $Pr[X > \varepsilon m] \le \frac{E[X]}{\varepsilon m} = \frac{1}{2}$ if $X > 0$

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

$$Pr\left[\arg\min_{r \in [t]} \sum_j Y_{rj} > \varepsilon m\right] \le \left(\frac{1}{2}\right)^t = \left(\frac{1}{2}\right)^{\log_2 \frac{1}{\delta}} = \delta$$

$\frac{\varepsilon m}{2} / \varepsilon m =$

Space for Count-Min

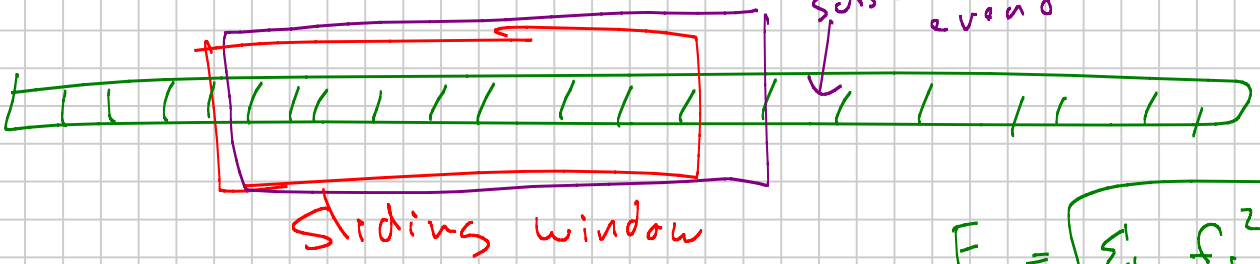
$$\left(\#\text{counters} = kt = \frac{2}{\varepsilon}\log\frac{1}{\delta}\right)\log m + \underbrace{\left(t = \log\frac{1}{\delta}\right)\log n}_{\text{space of hash}}$$

$\log\frac{1}{\delta} \approx 5\text{–}10\times$ bigger

MG sketch $\frac{1}{\varepsilon}(\log n + \log m)$

CM-sketch can do subtraction

turnstile model



$$F_2 = \sqrt{\sum_{j\in[n]} f_j^2}$$

Count-Sketch

$|f_q - \hat{f}_q| \le \varepsilon F_2$

sign hash function s_i | hash function h_i

$s_r : [n] \to \{-1, +1\}$

C_{11} C_{21} . . . . C_{k1}

C_{12} $t = 2\log(1/\delta)$

$k = O(1/\varepsilon^2)$

C_{1t} C_{kt}

Update

for $i=1$ to m

for $r=1$ to t

$$C_{r h_r(a_i)} = C_{r h_r(a_i)} + s_r(a_i)$$

Query

$$s_A(q) = \hat{f}_q$$

$$\text{median}_{r\in[t]} \left\{ C_{r, h_r(q)} \cdot s_r(q) \right\}$$

$E[\hat{f}_q] = f_q$

Frequent Itemset (Apriori Algorithm)

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

$a_i = \{ b_{i1}, b_{i2}, b_{i3}, \ldots \}$ $b_{ij} \in [n]$

"market basket"

Goal: All subsets of $[n]$ which co-occur more than ϕm baskets.

If $\{milk, eggs\}$ occurs ϕm times then both $\{milk\}$ and $\{eggs\}$ occurs $\geq \phi m$ times

→ Only Return Maximal Sets

→ Apriori Algorithm.

Round 1 Find all individual items occuring more than ϕm times

$M = \bigcup_{a_i} |a_i|$ → Heavy-Hitters M, $\phi M - \varepsilon M$

↳ $\{b_1, b_2, b_3, \ldots b_k\} \subset [n]$

Round 2 Pairs $\{\{b_1, b_2\}, \{b_1, b_3\}, \ldots \{b_{k-1}, b_k\}\}$ usually $< n$

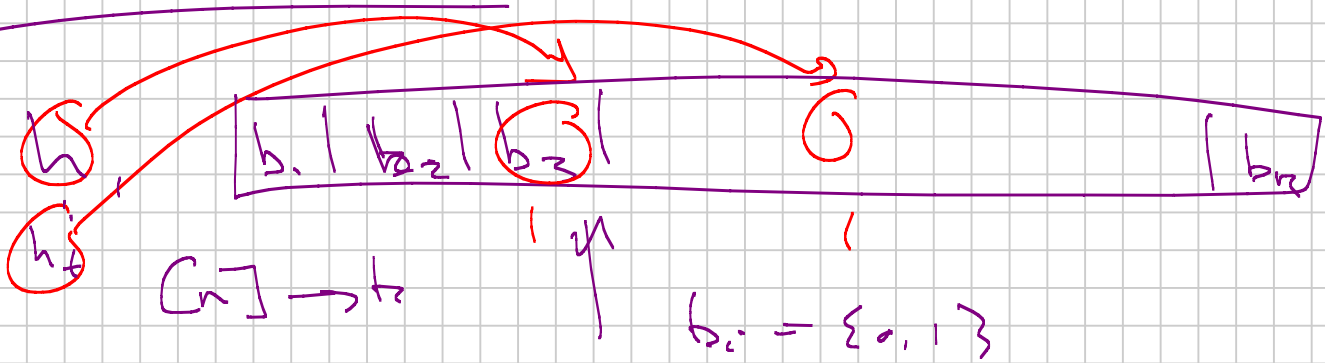
Counters for all pairs (or HH from δm-sm)

Set of Heavy pairs

$\{\{b_1, b_j\} \ldots \{\ , \ \}\}$

Round 3

Bloom Filter



Update

for $x \in S$

for $j = 1$ to k

Set $b_{h_j(x)} = 1$

IS x in S?

True iff all $j \in [k]$

$(b_{h_j(x)} == 1)$