L12: Streaming : Count-Min Sketch and Others

- Count-Min Sketch
  - ↳ proof
- Count Sketch

Jeff M. Phillips

February 21, 2018

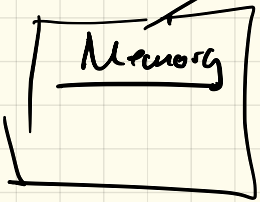
- Frequent Itemset (Apriori Alg)
- Bloom Filters

Streaming Model

$a_i \in [m]$

Input $A = \langle \underbrace{a_1, a_2,}_{A_i} \dots a_i, \dots a_n \rangle$

$A_i = \{a_1, a_2, \dots a_i\}$



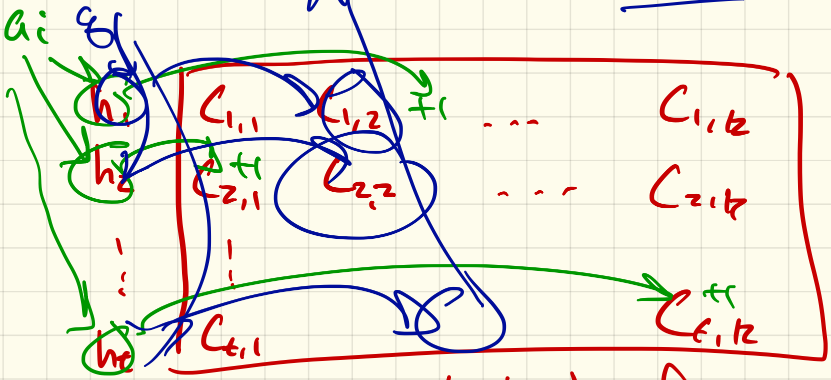
m, n very large
↳ use $O(\log n + \log m)$ space

Frequency $f_j = |\{a_i \in A \mid a_i = j\}|$

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

CM: $f_j \le \underbrace{\hat{f}_j \le f_j + \varepsilon n}_{\text{hold w.p. } 1-\delta}$ ← also handle substractions "turnstyle"

Count-Min Sketch



t hash fxns

$t \cdot k$ counters

$k = 2/\varepsilon$

$t = \log(1/\delta)$

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

space$(C_{ij}) = O(\log n)$

space$(h_j) = O(\log m)$

Initialize $C_{ij} = 0 \quad \forall i, j$

for $a_i \in A$

  for $j = 1$ to t

    $C_{j, h_j(a_i)} = C_{j, h_j(a_i)} + 1$

Query $\hat{f}_q$? $q \in [m]$

$\hat{f}_q = \min_{j' \in [t]} C_{j', h_{j'}(q)}$

Clear $f_g \le \hat{f}_g$ ← only overcounts

$\hat{f}_g \le f_g + w$ $p \in [m]$

R.V. $Y_{p,j} = \begin{cases} f_p & \text{w.p. } 1/k \\ 0 & \text{otherwise} \end{cases}$

$E[Y_{p,j}] = f_p/k$

jth row
Prob $C_{j,h_j(g)}$
the overcount from
$p \in [m]$

R.V. $X_j = \sum'_{\substack{p \in [m] \\ \neq g}} Y_{p,j}$ ← in jth row, total over count on $C_{j,h_j(g)}$

$E[X_j] = E\left(\sum'_{p \neq g} Y_{p,j}\right) = \sum'_{p \neq g} f_p/k \le \frac{m}{k} = \frac{\varepsilon m}{2}$

$\Pr[X > \alpha] = \frac{E[X]}{\alpha} = \frac{1}{2}$ Prob all t rows $> \varepsilon n$ error $(1/2)^t$

$\alpha = E[X] \cdot 2$

Count Sketch

$$\begin{array}{cc|cccc|}
s_1 & h_1 & C_{1,1} & C_{1,2} & \cdots & C_{1,k} \\
s_2 & h_2 & & & & \\
\vdots & \vdots & & & & \\
s_t & h_t & C_{t,1} & & \cdots & C_{t,k}
\end{array}$$

$$\boxed{k = \frac{1}{\varepsilon^2}}$$

$$t = \log\left(\frac{2}{\delta}\right)$$

$$h_j : [m] \to [k] \quad \text{(random)}$$

$$\text{(random)} \quad s_j : [m] \to \{-1, +1\}$$

for $a_i \in A$

  for $j \in [t]$

$$C_{j,h_j(a_i)} = C_{j,h_j(a_i)} + s_j(a_i) \cdot 1$$

$$E[C_{ij}] = 0$$

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

$$F_2 = \sqrt{\sum_{p \in [m]} f_p^2}$$

Bloom Filter

Data Structure S for sets.

Stream for $a_i \in A$

put $a_i \xrightarrow[\text{into}]{} S$

Query is $q \in [m]$ in S?

- if $q \in S \to$ always return true
- if $q \notin S$ (not in) $\to$ usually return false

Init $B[q] = 0 \ \forall q$

for $a_i \in A$

  for $j = 1$ to k

    set $B[h_j(a_i)] = 1$

k hash fcns $h_1, h_2, \ldots h_k$

1 array of bits of m bits $B[\ ]$

$k \approx \frac{m}{n} \ln(2)$

A-Priori Algorithm (Frequent Itemsets)

<u>Input</u>: $A = \{a_1, a_2 \ldots a_n\}$

$a_i = \{x_1, x_7, x_{14}\} \subset [m]$

Market Basket Analysis

$\hookrightarrow$ beer + diapers

$\varepsilon = 0.05$

Find all tuples $\{x_1, x_2, x_7\}$ w/ cooccur in at least $\boxed{\varepsilon n}$ baskets

If $\{x_1, x_2, x_7\}$ cooccur in 5% then each of x_1, x_2, and x_7 must each occur $\geq$ 5%

Frequent Itemsets : Apriori

Find tuples in at least 1/3 sets

0	1	2	3	4	5	6	7	8	9
2	3	5	4	3	3	8	4	2	4

2,3	2,6	2,7	2,9	3,6	3,7	3,9
1	3	1	2	3	0	1

6,7	6,9	7,9
3	4	2

3, 6, 9

$T_1 = \{1, 2, 3, 4, 5\}$

$T_2 = \{2, 6, 7, 9\}$

$T_3 = \{1, 3, 5, 6\}$

$T_4 = \{2, 6, 9\}$

$T_5 = \{7, 8\}$

$T_6 = \{1, 2, 6\}$

$T_7 = \{0, 3, 5, 6\}$

$T_8 = \{0, 2, 4\}$

$T_9 = \{2, 4\}$

$T_{10} = \{6, 7, 9\}$

$T_{11} = \{3, 6, 9\}$

$T_{12} = \{6, 7, 8\}$