L11: Streaming : Frequent Items and Quantiles

Jeff M. Phillips

February 14, 2018

Streaming Model

Data $A = \langle a_1, a_2, \ldots a_i, \ldots a_n \rangle$

A_i treated as set $\{a_1, a_2, \ldots a_i\}$

Data Type

① $a_i \in [m] = \{1, 2, \ldots m\}$

IP address
3-gram
word
SNP

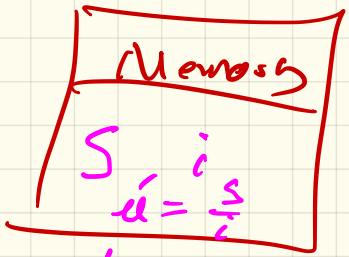
frequent items

$S = \sum_{j=1}^{i} a_j$

② $a_i \in \mathbb{R}$ (scalar) can be sorted

Apx CDF

View a_i



much too small to store all data

Maintain statistics about A_i

Frequent Items

$A = \langle a_1, a_2, \ldots a_n \rangle \qquad a_i \in [m]$

Mem Size $\ll n, m$

Approximate all f_j $= C \cdot (\underbrace{\log n}_{\text{counter}} + \underbrace{\log m}_{\text{label}})$

↳ 1 word

1 IP address

Frequency $f_j \qquad j \in [m]$

$= \#$ times $a_i = j$

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

(counter, label)

$O(\log m + \log n)$ space

↳ How many times saw IP address i.

$F_1 = \sum_{j=1}^{m} f_j \qquad F_2 = \sqrt{\sum_{j=1}^{m} f_j^2}$

$F_0 = \sum_{j=1}^{m} f_j^0 = \#$ distinct items

MAJORITY

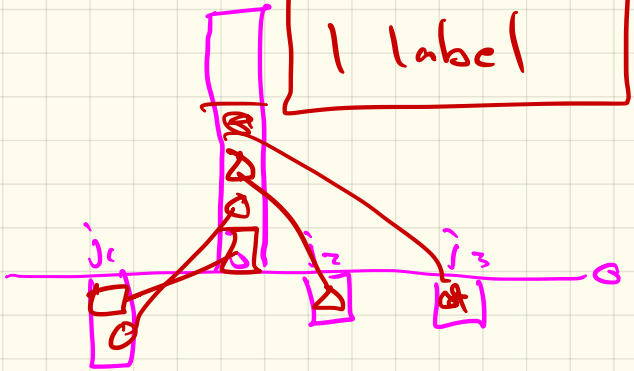
- if (some $f_j > \frac{n}{2}$), output j
  else, output anything

Initialize $L \leftarrow \phi$, $C = 0$

for $i = 1$ to n
  Read a_i
  if $(a_i = L)$ $C = C + 1$
  else $C = C - 1$
  if $(C < 0)$: $C = 1$ & $L = a_i$
end for

return L

Memory
1. counter
1 label



Majority

Majority(A)

Set $c = 0$ and $\ell = \emptyset$
for $i = 1$ **to** m **do**
  if $(a_i = \ell)$ **then**
    $c = c + 1$
  else
    $c = c - 1$
  if $(c < 0)$ **then**
    $c = 1$, $\ell = a_i$
return ℓ

Misra - Gries

For any $j \in [m]$ Return $S(j) = \hat{f}_j$

ε = error $\in [0,1]$

$$f_j - \frac{n}{k} \leq \hat{f}_j \leq f_j$$

$k = \frac{1}{\varepsilon} \Rightarrow \frac{n}{k} = n\varepsilon$

Initialize:

for $i = 1$ to n

- If (a_i = some L_j) $C_j = C_j + 1$
- else (if some $C_j = 0$) $C_j = 1$ $L_j = a_i$
- else Decrement all Counters
  for $l = 1$ to $k-1$ $C_j = C_j - 1$

can occur at most n/k times

end for

Return $S \left[C = C_1, C_2 \ldots C_{k-1} \;/\; L = L_1, L_2 \ldots L_{k-1} \right]$

Memory

$(k-1)$ counters + labels

Misra-Gries

counter array C : $C[1], C[2], \ldots, C[k-1]$
location array L : $L[1], L[2], \ldots, L[k-1]$

Misra-Gries(A)

Set all $C[i] = 0$ and all $L[i] = \emptyset$
for $i = 1$ **to** m **do**
  if $(a_i = L[j])$ **then**
    $C[j] = C[j] + 1$
  else
    if (some $C[j] = 0$) **then**
      Set $L[j] = a_i$ & $C[j] = 1$
    else
      for $j \in [k-1]$ **do** $C[j] = C[j] - 1$
return C, L

Quantiles

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

$q_i \in \mathbb{R}$ ↳ requires $\log m$ space

$\underset{v \in \mathbb{R}}{\text{rank}_A}(v) = |\{a_i \in A \mid a_i \le v\}|$

$\text{cdf}(v) = \frac{\text{rank}_A(v)}{n}$

Summary Q_A s.t. $\forall v \in \mathbb{R}$ $\left| Q_A(v) - \frac{\text{rank}_A(v)}{n} \right| < \varepsilon$

error $\in [0, 1]$

Memory

Space $(\log m) \cdot \frac{1}{\varepsilon} \log\log\frac{1}{\varepsilon}$ $k = \frac{1}{\varepsilon}$

Set $\frac{1}{\varepsilon}$ values sorted order $Q[q_1 \le q_2 \le \dots q_k]$

Merge $Q, Q' \Rightarrow q_1 \le q'_1 \le q_2 \le q'_3 \le q_4 \le q'_4 \le q_5 \le \dots$

Frugal Median

Let $v \in \Pi$ defined so $\frac{rank_\Pi(v)}{n} = \frac{1}{2}$

median $\longrightarrow \hat{q}$

s.t. $\hat{q}$ close to v.

$$\frac{rank_\Pi(\hat{q})}{n} \in \left[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right]$$

Memory

1 label

Set $\ell = a_1$

```
for i=2 to n
    if (a_i > ℓ)
        ℓ = ℓ + 1
    else (a_i < ℓ)
        ℓ = ℓ - 1
return ℓ
```

improve estimation by increment

Frugal Median

Frugal Median(A)

Set $\ell = 0$.
for $i = 1$ **to** m **do**
  if $(a_i > \ell)$ **then**
    $\ell \leftarrow \ell + 1$.
  if $(a_i < \ell)$ **then**
    $\ell \leftarrow \ell - 3$.
return ℓ.

What if estimate is the 25%-quantile?

Frugal Quantile

Frugal Quantile(A, ϕ) e.g. $\phi = 0.75$

Set $\ell = 0$.
for $i = 1$ **to** m **do**
  $r = \text{Unif}(0, 1)$ (at random)
  if $(a_i > \ell$ **and** $r > 1 - \phi)$ **then**
    $\ell \leftarrow \ell + 1$.
  if $(a_i < \ell$ **and** $r > \phi)$ **then**
    $\ell \leftarrow \ell - 1$.
return ℓ.