L11: Streaming : Frequent Items and Quantiles

Jeff M. Phillips

February 12, 2020

Big Data

size $|X| = n$
very large

too big to fit one 1 computer.

- **Parallelism**: More computers
  MapReduce

- **Sampling**: $P \sim X \overset{iid}{\sim} \mu$ $\qquad |P| \ll |X|$
  data set $\qquad P \sim \mu$

- **Streaming**: $X = \langle x_1, x_2, \ldots, x_i, \ldots, x_n \rangle$
  Read data in one pass
  Maintain small space summary

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

$a_i \in \mathbb{R}$

$A_i = \langle a_1, a_2, \ldots, a_i \rangle$

$\text{mean}(A_i) = \frac{S_i}{i}$ $\quad$ maintain $\quad S_i = \sum_{j=1}^{i} a_i$

i

$\text{variance}(A_i) = \frac{Q_i}{i} + \left(\frac{S_i}{i}\right)^2 \quad Q_i = \sum_{j=1}^{i} a_i^2$

Reservoir Sampling

Maintain Random Sample $B \subset A$

w/o replacement
uniform
$|B| = k$

1. Keep first k $B = A_k$
2. for $j = k+1$ to n
   with prob $\frac{k}{j}$ $\longrightarrow$ replace some $b_i \in B$ w/ a_j
   otherwise keep B

ε - error

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

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

$= 1, 2, \ldots m$

n too large

m too large

IP addresses

k-grams

label $j \in [m]$ $\underline{\log m}$ bits

counter value $\in 1 \ldots n$ $\underline{\log n}$ bits

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

$j \in [m]$

apx freq $\hat{f}_j$: $|f_j - \hat{f}_j| \leq \varepsilon n$

Is any $j \in [m]$ have $f_j > \phi n$

or $\hat{f}_j > \phi n - \varepsilon n$

MAJORITY

$A = \langle a_1 \dots a_n \rangle$
$a_i \in [m]$

If some $f_j > \frac{n}{2} \rightarrow$ output j
else output anything.

1 counter, 1 label

if $(a_i = \text{label})$
counter $c = c + 1$

else
$c = c - 1$

if $c < 0$
$\ell = a_i$

Majority

Majority(A)

```
Set c = 0 and ℓ = ∅
for i = 1 to n do
    if (a_i = ℓ) then
        c = c + 1
    else
        c = c − 1
    if (c < 0) then
        c = 1, ℓ = a_i
return ℓ
```

Misra - Greis Freq. Apx

$k-1$ counters, $k-1$ labels

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

\# decrements

apx freq

forall $j \in [m]$

$k = 1/\varepsilon$

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

if j not in set $L = \{l_1, l_2 .. l_{k-1}\}$
$\hat{f}_j = 0$

large counts

storing $\hat{f}_j$ = 1 counter
1 label j

counters $C = \{c_1, .. c_{k-1}\}$

for ($a_i \in A$)
  if $a_i \in L$ matches a label l_j
    $c_j = c_j + 1$
  else ($a_i \notin L$)
    for $j \in [1 .. k-1]$ $c_j = c_j - 1$
  if ($c_j \in C$ has $c_j = 0$)
    $l_j = a_i$, $c_j = 1$

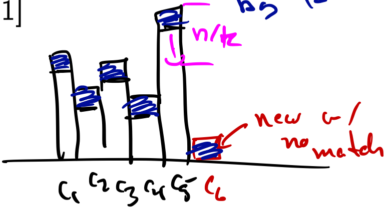
Misra-Gries

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

<u>Misra-Gries(A)</u>

  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

Total count decreases by k



$k = 6$

How many times?

<u>n/k times</u>

Streaming Median

$A = \langle a_1, \ldots a_n \rangle \quad a_i \in \mathbb{R}$

Maintain median(A)

$\mathrm{rank}_A(v) = |\{ a_i \in A \mid a_i \leq v \}|$

$v \in \mathbb{R}$

quantile estimate $Q_A(v) \rightarrow \frac{1}{\varepsilon} \log\log \frac{1}{\varepsilon}$

space

for $v \in \mathbb{R}$ $\quad \left| Q_A(v) - \frac{\mathrm{rank}_A(v)}{n} \right| \leq \varepsilon$

$\frac{1}{\varepsilon} \log^{3/2} \frac{1}{\varepsilon}$

Frugal Median

Assume A iid

order random

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**
    $\ell \leftarrow \ell - 1$.
return ℓ.

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.