L11 Streaming (Misra-Greis)

Monday, February 24, 2020 6:06 AM

Big Data (size $|X| = n$, very large)

- to big to fit on one computer

Strategies

- Parallelism: more computers
- Sampling: $X \overset{iid}{\sim}_{data} M$ $|P| << |X|$, $P \sim M$
- Streaming: $X = \langle x_1, x_2, \ldots, x_n \rangle$ (there is some order)
  - read data in one pass
  - maintain small space summary

Streaming:

- Data $A = \langle a_1, a_2, \ldots, a_n \rangle$
  - $A_i = \langle a_1, a_2, \ldots, a_i \rangle$ where $A_i \subset A$
    - mean$(A_i) = S_i / i$ maintain $S_i = \sum_{j=1}^{i} a_i$ and i
    - variance$(A_i) = \frac{Q_i}{i} - \left(\frac{S_i}{i}\right)^2$ where $Q_i = \sum_{j=1}^{i} a_i^2$

Reservoir Sampling

- maintain a random sample $B \subset A$ without replacement
  - uniform
  - $|B| = K$

1. Keep the first K things: $B = A_K$
2. For $j = K+1$ to n . . .
   - with probability K/j . . .
     - replace some $b_j \in B$ with a_j
     - ε-error $\Rightarrow$ $K = \frac{1}{\varepsilon^2}$

- Consider a stream $A = \langle a_1, a_2, \ldots, a_n \rangle$, $a_i \in [m] = 1, 2, \ldots, m$
  - n is too large
  - m is too large
  - we can store some label $j \in [m]$ with $\log m$ bits.
  - we can also keep a count of observed values with $\log n$ bits.
  - we would like to find the frequency of j

- we can also keep a count of observed values with $\log n$ bits.
- we would like to find the frequency of j

$$\text{frequency } f_j \quad j \in [m] = |\{a_i \in A \mid a_i = j\}|$$

- we can also track some approximation of frequency $\hat{f}_j$ such that $|f_j - \hat{f}_j| \leq \epsilon n$
  - Then a question of interest may be, ...
  - Q: Does any $j \in [m]$ have $f_j > \phi(n)$ or $\hat{f}_j > \phi n - \epsilon n$

Majority Problem:

- We have some stream $A = \langle a_1, a_2, \ldots, a_n \rangle$, $a_i \in [m]$
- if some $f_j > n/2$ → output j
  - else, output anything

} can be done with one counter and one label

Algorithm:

- set $c = 0$, $\ell = \phi$
- for $i = 1$ to n
  - 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 (Frequency Approximation):

- $k-1$ counters, $k-1$ labels
  - $f_j - \frac{n}{k} \leq \hat{f}_j \leq f_j$ for all $j \in [m]$

    approx of frequency

    if $k = 1/\epsilon$, then $\frac{n}{k} = n\epsilon$
  - if j not in the set of labels $L = \{\ell_1, \ell_2, \ldots, \ell_{k-1}\}$, then $\hat{f}_j = 0$

- for $(a_i \in A)$ for each element in A
  - if $a_i \in L$ • check if a_i matches a label
    - $c_j = c_j + 1$ • if so, increment its count
  - else $(a_i \notin L)$ • if it didn't match a label

- $c_j = c_j + 1$ • if so, increment its count
- else ($a_i \notin L$) • if it didn't match a label
  - for all $j \in [1, \ldots, K-1]$ then for each label
    - $c_j = c_j - 1$ decrement its count
  - if ($c_j \in C$ has $c_j \leq 0$) check if any label has a count ≤ 0
    - $l_i = a_i$, $c_j = 1$ replace that label with a_i
- return L (all labels) and C (all counts)

/