Massive Data Algorithmics

The Streaming Model

Lecture 15: Frequent Items via Sketching

The Problem

- Frequecy
  - Input: the stream $\sigma =< a_1, \cdots, a_m >$ where $a_i \in [n]$
  - Frequent number of item j: $f_j = |\{i : a_i = j\}|$,
  - Frequency vector: $F = (f_1, f_2, \cdots, f_n)$
  - Frequency moments: $F_k = ||F||_k^k = \sum_{j=1}^n f_j^k$
  - F_0: the number of distint items
  - F_1: the number of items (i.e. m)
- Problem: Estimating the frequent number
  - Input: the stream $\sigma =< a_1, \cdots, a_m >$ where $a_i \in [n]$, and k
  - Output for a query a: $\hat{f}_a$, an estimator of f_a

Misra-Gries Algorithm

- Let $\hat{f}_a$ be the output of Misra-Gries Algorithm. We showed $f_a - m/k \leq \hat{f}_a \leq f_a$.
- Let $MG(\sigma)$ denote the data structure (the set of $k-1$ keys and their counters) computed by Misra-Gries algorithm.
- One drawback of this data structure is that there is no general way to compute $MG(\sigma_1 \cdot \sigma_2)$ from $MG(\sigma_1)$ and $MG(\sigma_2)$ where $\sigma_1 \cdot \sigma_2$ is the concatenation of σ_1 and σ_2.

Sketch

- A data structure $DS(\sigma)$ computed in streaming fashion by processing a stream σ is called a sketch if there is a space-efficient combining algorithm COMB such that, for every two stream σ_1 and σ_2, we have

$$\mathrm{COMB}(DS(\sigma_1), DS(\sigma_2)) = DS(\sigma_1 \cdot \sigma_2)$$

The Algorithm

Initialize :

$C[1 \dots k] \leftarrow \vec{0}$, where $k := 3/\varepsilon^2$;

Choose a random hash function $h : [n] \to [k]$ from a 2-universal family ;

Choose a random hash function $g : [n] \to \{-1, 1\}$ from a 2-universal family ;

Process (j, c):

$C[h(j)] \leftarrow C[h(j)] + cg(j)$;

Output :

On query a, report $\hat{f}_a = g(a)C[h(a)]$;

Analysis

- $X = \hat{f}_a$
- Let Y_j be the indicator for the event "$h(j) = h(a)$".
- Token j contributes to the counter $C[h(a)]$ iff $h(j) = h(a)$ and the amount of the contribution is its frequency times sign $g(j)$.
- $X = g(a)\sum_{j=1}^{n} f_j g(j) Y_j = f_a + \sum_{j\in[n]-\{a\}} f_j g(a) g(j) Y_j$
- $E(g(j)Y_j) = E(g(j))E(Y_j) = 0 \cdot E(Y_j) = 0$
- $E(X) = f_a + \sum_{j\in[n]-\{a\}} f_j g(a) E(g(j)Y_j) = f_a$
- We need to show that X is unlikely to deviate too much from its mean. For this, we analyze its variance.

Analysis

- For $j \in [n] - \{a\}$, we have

$$E(Y_j^2) = E(Y_j) = \Pr(h(j) = h(a)) = 1/k$$

- For all $i, j \in [n]$ and $i \neq j$, we have

$$E(g(i)g(j)Y_iY_j) = E(g(i))E(g(j))E(Y_iY_j) = 0 \cdot 0 \cdot E(Y_iY_j) = 0$$

- $Var(X) = 0 + g(a)^2 Var(\sum_{j \in [n]-\{a\}} f_j g(j) Y_j) = E(\sum_{j \in [n]-\{a\}} f_j^2 Y_j^2 + \sum_{i,j \in [n]-\{a\}, i \neq j} f_i f_j g(i) g(j) Y_i Y_j) - (\sum_{j \in [n]-\{a\}} f_j E(g(j)Y_j))^2 = \sum_{j \in [n]-\{a\}} \frac{f_j^2}{k} + 0 - 0 = \frac{||F||_2^2 - f_a^2}{k}$

Analysis

- $\Pr(|\hat{f}_a - f_a| \geq \varepsilon\sqrt{||F||_2^2 - f_a^2}) = \Pr(|X - E(X)|) \geq \varepsilon\sqrt{||F||_2^2 - f_a^2}) \leq \frac{Var(X)}{\varepsilon^2(||F||_2^2 - f_a^2)} = \frac{1}{k\varepsilon^2} = \frac{1}{3}$
- For $j \in [n]$, let us define F_{-j} to be the $(n-1)$-dimensional vector obtained by dropping the j-th entry of F. Then, $||F_{-j}||_2^2 = ||F||_2^2 - f_j^2$.
- We can rewrite the above statement in the following more memorable form

$$\Pr(|\hat{f}_a - f_a| \geq \varepsilon||F_{-a}||_2) \leq \frac{1}{3}$$

The Final Sketch

- Apply the median trick to bring its probability down to δ

Initialize :

$C[1\ldots t][1\ldots k] \leftarrow \vec{0}$, where $k := 3/\varepsilon^2$ and $t := O(\log(1/\delta))$;

Choose t independent hash functions $h_1, \ldots h_t : [n] \rightarrow [k]$, each from a 2-universal family ;

Choose t independent hash functions $g_1, \ldots g_t : [n] \rightarrow [k]$, each from a 2-universal family ;

Process (j, c):

for $i = 1$ **to** t **do** $C[i][h_i(j)] \leftarrow C[i][h_i(j)] + c g_i(j)$;

Output :

On query a, report $\hat{f}_a = \text{median}_{1 \le i \le t}\, g_i(a) C[i][h_i(a)]$;

Analysis

- $\Pr(|\hat{f}_a - f_a| \geq \varepsilon ||F_{-a}||_2) \leq \delta$
- space: $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} (\log n + \log m))$

The Algorithm

Initialize :

$C[1\dots t][1\dots k] \leftarrow \vec{0}$, where $k := 2/\varepsilon$ and $t := \lceil \log(1/\delta) \rceil$;

Choose t independent hash functions $h_1, \dots h_t : [n] \to [k]$, each from a 2-universal family

Process (j, c):

for $i = 1$ **to** t **do** $C[i][h_i(j)] \leftarrow C[i][h_i(j)] + c$;

Output :

On query a, report $\hat{f}_a = \min_{1 \le i \le t} C[i][h_i(a)]$;

Analysis

- space: $O(\frac{1}{\varepsilon}\log\frac{1}{\delta}(\log n + \log m))$
- It is clear $f_a \leq \hat{f}_a$
- We analyze the excess in (i.e. $\hat{f}_a - f_a$) for counter $C[i][h_i(a)]$.
- Let X_i be this excess.
- For $j \in [n] - \{a\}$, let $Y_{i,j}$ be the indicator of the event "$h_i(j) = h_i(a)$".
- j makes a contribution to the counter iff $Y_{i,j} = 1$ and when it does contribute, it causes f_j to be added to this counter.
- $X_i = \sum_{j\in[n]-\{a\}} f_j Y_{i,j}$
- $E(Y_{i,j}) = 1/k$
- $E(X_i) = \sum_{j\in[n]-\{a\}} \frac{f_j}{k} = \frac{||F||_1 - f_a}{k} = \frac{||F_{-a}||_1}{k}$

Analysis

- $\Pr(X_i \geq \varepsilon ||F_{-a}||_1) \leq \frac{||F_{-a}||_1}{k\varepsilon ||F_{-a}||_1} = \frac{1}{2}$
- $\Pr(\hat{f}_a - f_a \geq \varepsilon ||F_{-a}||_1) = \Pr(\min(X_1, \cdots, X_t) \geq \varepsilon ||F_{-a}||_1) = \Pr(\wedge_{i=1}^{t}(X_i \geq \varepsilon ||F_{-a}||_1)) = \Pi_{i=1}^{t} \Pr(X_i \geq \varepsilon ||F_{-a}||_1)] \leq \frac{1}{2^t} \leq \delta$
- $f_a \leq \hat{f}_a \leq f_a + \varepsilon ||F_{-a}||_1$
- The deviation $\varepsilon ||F_{-a}||_1$ is weaker than the deviation $\varepsilon ||F_{-a}||_2$ of the count sketch.

Comparison of Frequency Estimation Methods

Method	$\hat{f}_a - f_a \in \cdots$	Space	Error Probability
Misra-Gries	$\left[-\varepsilon\\|\mathbf{f}_{-a}\\|_1,\ 0\right]$	$O\left(\frac{1}{\varepsilon}(\log m + \log n)\right)$	0 (deterministic)
Count Sketch	$\left[-\varepsilon\\|\mathbf{f}_{-a}\\|_2,\ \varepsilon\\|\mathbf{f}_{-a}\\|_2\right]$	$O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\cdot(\log m + \log n)\right)$	δ (overall)
Count-Min Sketch	$\left[0,\ \varepsilon\\|\mathbf{f}_{-a}\\|_1\right]$	$O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\cdot(\log m + \log n)\right)$	δ (upper bound only
Count/Median	$\left[-\varepsilon\\|\mathbf{f}_{-a}\\|_1,\ \varepsilon\\|\mathbf{f}_{-a}\\|_1\right]$	$O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\cdot(\log m + \log n)\right)$	δ (overall)

References

- **Data Stream Algorithms** (Chapter 4)
  Lecture notes by A. Chakrabbarti and D. College