CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 13

Samson Zhou

Presentation Schedule

- September 25: Team DAP, Team Bokun, Team Jason
- September 27: Galaxy AI, Team STMI
- September 29: Jung, Anmol, Chunkai

Last Time: L_2 Heavy-Hitters

- Goal: Given a set S of m elements from $[n]$ that induces a frequency vector $f \in R^n$ and a threshold parameter $\varepsilon \in (0,1)$, output a list that includes:
  - The items from $[n]$ that have frequency at least $\varepsilon \cdot ||f||_2$
  - No items with frequency less than $\frac{\varepsilon}{2} \cdot ||f||_2$

Last Time: CountSketch, i.e., CountMin and the Power of Random Signs

- Initalization: Create b buckets of counters and use a random hash function $h: [n] \to [b]$ and a uniformly random sign function $s: [n] \to \{-1, +1\}$, i.e., $\Pr[s(i) = +1] = \Pr[s(i) = -1] = \frac{1}{2}$
- Algorithm: For each insertion (or deletion) to x_i, change the counter $h(x_i)$ by $s(x_i)$ (or $-s(x_i)$)

c_1	c_2	c_3	c_4
0	0	0	0

- At the end of the stream, output the quantity $s(x_i) \cdot h(x_i)$ as the estimate for x_i

CountSketch

- Given a set S of m elements from $[n]$, let $\widehat{f_i}$ be the estimated frequency for f_i

- Suppose $h(i) = a$ so that $\widehat{f_i} = s(i) \cdot c_a$
- Note that c_a includes the signed number $s(j) \cdot f_j$ of occurrences of any j with $h(j) = a = h(i)$, including f_i itself

CountSketch

- Suppose $h(i) = a$ so that $c_a = \widehat{f_i}$
- Note that c_a includes the signed number $s(j) \cdot f_j$ of occurrences of any j with $h(j) = a = h(i)$, including f_i itself

- $c_a = \sum_{j:h(j)=a} s(j) \cdot f_a$
- Estimated frequency f_i of i is $\widehat{f_i} = s(i) \cdot c_a$
- $s(i)\ c_a = s(i) \cdot s(i) \cdot f_i + \sum_{j \neq i,\text{ with } j:h(j)=a} s(i) \cdot s(j) \cdot f_j$

CountSketch Error Analysis

- $c_a = s(i) \cdot s(i) \cdot f_i + \sum_{j \neq i, \text{ with } j:h(j)=a} s(i) \cdot s(j) \cdot f_j$
- Since $s(i) \in \{-1, +1\}$, we have $s(i) \cdot s(i) = 1$
- What is the expected error for f_i?

CountSketch Error Analysis

- $c_a = s(i) \cdot s(i) \cdot f_i + \sum_{j \neq i, \text{ with } j : h(j) = a} s(i) \cdot s(j) \cdot f_j$
- What is the expectation of the error term for f_i?
- $\mathrm{E}\left[\sum_{j \neq i, \text{ with } j : h(j) = a} s(i) \cdot s(j) \cdot f_j\right] = \Sigma_{j \neq i} \mathrm{E}\left[s(i) \cdot s(j) \cdot f_j \cdot I_{h(j) = h(i)}\right]$

CountSketch Error Analysis

- $c_a = s(i) \cdot s(i) \cdot f_i + \sum_{j \neq i, \text{ with } j : h(j) = a} s(i) \cdot s(j) \cdot f_j$
- What is the expectation of the error term for f_i?
- $\mathrm{E}\left[\sum_{j \neq i, \text{ with } j : h(j) = a} s(i) \cdot s(j) \cdot f_j\right] = 0$

CountSketch Error Analysis

- $c_a = s(i) \cdot s(i) \cdot f_i + \sum_{j \neq i, \text{ with } j:h(j)=a} s(i) \cdot s(j) \cdot f_j$
- What is the expectation of the error term for f_i?
- $\mathrm{E}\left[\sum_{j \neq i, \text{ with } j:h(j)=a} s(i) \cdot s(j) \cdot f_j\right] = 0$
- What is the variance of the error term for f_i?

CountSketch Error Analysis

- Variance is at most the 2nd moment of the error term
- $\mathrm{E}\left[\left(\sum_{j \neq i, \text{ with } j:h(j)=a} s(i) \cdot s(j) \cdot f_j\right)^2\right]$

CountSketch Error Analysis

- Variance is at most the 2nd moment of the error term
- $\mathrm{E}\left[\left(\sum_{j\neq i,\text{ with } j:h(j)=a} s(i)\cdot s(j)\cdot f_j\right)^2\right]=\sum_{j\neq i}\mathrm{E}\left[\left|f_j\right|^2\cdot I_{h(j)=h(i)}\right]$

CountSketch Error Analysis

- Variance is at most the 2nd moment of the error term
- $\mathrm{E}\left[\left(\sum_{j\neq i,\text{ with } j:h(j)=a} s(i)\cdot s(j)\cdot f_j\right)^2\right] = \sum_{j\neq i}\mathrm{E}\left[\left|f_j\right|^2\cdot I_{h(j)=h(i)}\right]$
  $= \sum_{j\neq i}\mathrm{E}\left[I_{h(j)=h(i)}\right]\cdot\left|f_j\right|^2$

CountSketch Error Analysis

- Variance is at most the 2nd moment of the error term
- $\mathrm{E}\left[\left(\sum_{j\neq i,\text{ with } j:h(j)=a} s(i)\cdot s(j)\cdot f_j\right)^2\right] = \sum_{j\neq i}\mathrm{E}\left[\left|f_j\right|^2\cdot I_{h(j)=h(i)}\right]$

$$= \sum_{j\neq i}\mathrm{E}\left[I_{h(j)=h(i)}\right]\cdot\left|f_j\right|^2$$

$$= \sum_{j\neq i}\Pr[h(j)=h(i)]\cdot\left|f_j\right|^2$$

CountSketch Error Analysis

- Variance is at most the 2nd moment of the error term
- $\mathrm{E}\left[\left(\sum_{j\neq i,\text{ with } j:h(j)=a} s(i)\cdot s(j)\cdot f_j\right)^2\right] = \sum_{j\neq i}\mathrm{E}\left[\left|f_j\right|^2\cdot I_{h(j)=h(i)}\right]$

$$= \sum_{j\neq i}\mathrm{E}\left[I_{h(j)=h(i)}\right]\cdot\left|f_j\right|^2$$

$$= \sum_{j\neq i}\Pr[h(j)=h(i)]\cdot\left|f_j\right|^2$$

$$= \sum_{j\neq i}\frac{1}{b}\cdot\left|f_j\right|^2 \leq \frac{\|f\|_2^2}{b}$$

CountSketch Error Analysis

- Variance is at most the 2nd moment of the error term
- $\mathrm{E}\left[\left(\sum_{j\neq i,\text{ with } j:h(j)=a} s(i)\cdot s(j)\cdot f_j\right)^2\right] = \sum_{j\neq i}\mathrm{E}\left[|f_j|^2\cdot I_{h(j)=h(i)}\right]$

$$= \sum_{j\neq i}\mathrm{E}\left[I_{h(j)=h(i)}\right]\cdot |f_j|^2$$

$$= \sum_{j\neq i}\Pr[h(j)=h(i)]\cdot |f_j|^2$$

$$= \sum_{j\neq i}\frac{1}{b}\cdot |f_j|^2 \leq \frac{\|f\|_2^2}{b}$$

- Set $b = \frac{9k^2}{\varepsilon^2}$, then the variance is at most $\frac{\varepsilon^2\|f\|_2^2}{9k^2}$

CountSketch Error Analysis

- Set $b = \frac{9k^2}{\varepsilon^2}$, then the variance is at most $\frac{\varepsilon^2 \|f\|_2^2}{9k^2}$
- By Chebyshev's inequality, the error for f_i is at most $\frac{\varepsilon}{k}\|f\|_2$ with probability at least $\frac{2}{3}$
- How to ensure accuracy for all $i \in [n]$?

CountSketch Error Analysis

- By Chebyshev's inequality, the error for f_i is at most $\frac{\varepsilon}{k}\|f\|_2$ with probability at least $\frac{2}{3}$

- How to ensure accuracy for all $i \in [n]$?

- Repeat $\ell := O(\log n)$ times to get estimates $e_1, \ldots, e_\ell$ for each $i \in [n]$ and set $\widehat{f_i} = \mathrm{median}(\mathrm{e}_1, \ldots, \mathrm{e}_\ell)$

CountSketch Error Analysis

- Claim: For all estimated frequencies $\widehat{f_i}$ by CountSketch, we have

$$f_i - \frac{\varepsilon \|f\|_2}{k} \leq \widehat{f_i} \leq f_i + \frac{\varepsilon \|f\|_2}{k}$$

CountSketch Summary

- CountSketch solves the L_2 heavy-hitters problem: Given a set S of m elements from $[n]$ that induces a frequency vector $f \in R^n$ and a threshold parameter $\varepsilon \in (0, 1)$, output a list that includes:
  - The items from $[n]$ that have frequency at least $\varepsilon \cdot \|f\|_2$
  - No items with frequency less than $\frac{\varepsilon}{2} \cdot \|f\|_2$

- Space usage: $O\left(\frac{1}{\varepsilon^2} \log^2 n\right)$ bits of space