CMPSCI 711: More Advanced Algorithms

Vectors 3: Count-Min Sketch and Applications

Andrew McGregor

Last Compiled: February 16, 2018

Point Queries etc.

- *Stream:* m elements from universe $[n] = \{1, 2, \ldots, n\}$, e.g.,

$$\langle x_1, x_2, \ldots, x_m \rangle \quad = \quad \langle 3, 5, 103, 17, 5, 4, \ldots, 1 \rangle$$

  and let f_i be the frequency of i in the stream.
- *Problems:*
  - Point Query: For $i \in [n]$, estimate f_i
  - Range Query: For $i, j \in [n]$, estimate $f_i + f_{i+1} + \ldots + f_j$
  - Quantile Query: For $\phi \in [0, 1]$ find j with $f_1 + \ldots + f_j \approx \phi m$
  - Heavy Hitter Problem: For $\phi \in [0, 1]$, find all i with $f_i \geq \phi m$.

Count-Min Sketch

- Let $H_1, \ldots, H_d : [n] \to [w]$ be 2-wise independent functions.
- As we observe the stream, we maintain $d \cdot w$ counters where

$$c_{i,j} = \text{number of elements } e \text{ in the stream with } H_i(e) = j$$

- For any x, $c_{i,H_i(x)}$ is an over-estimate for f_x and so,

$$f_x \leq \tilde{f}_x = \min(c_{1,H_1(x)}, \ldots, c_{d,H_d(x)})$$

- If $w = 2/\epsilon$ and $d = \log_2 \delta^{-1}$ then,

$$\mathbb{P}\left[f_x \leq \tilde{f}_x \leq f_x + \epsilon m\right] \geq 1 - \delta \ .$$

Count-Min Sketch Analysis (a)

- Define random variables $Z_1, \ldots, Z_k$ such that $c_{i,H_i(x)} = f_x + Z_i$, i.e.,

$$Z_i = \sum_{y \neq x : H_i(y) = H_i(x)} f_y$$

- Define $X_{i,y} = 1$ if $H_i(y) = H_i(x)$ and 0 otherwise. Then,

$$Z_i = \sum_{y \neq x} f_y X_{i,y}$$

- By 2-wise independence,

$$\mathbb{E}\left[Z_i\right] = \sum_{y \neq x} f_y \mathbb{E}\left[X_{i,y}\right] = \sum_{y \neq x} f_y \mathbb{P}\left[H_i(y) = H_i(x)\right] \leq m/w$$

- By Markov inequality,

$$\mathbb{P}\left[Z_i \geq \epsilon m\right] \leq 1/(w\epsilon) = 1/2$$

Count-Min Sketch Analysis (b)

- Since each Z_i is independent

$$\mathbb{P}\left[Z_i \geq \epsilon m \text{ for all } 1 \leq i \leq d\right] \leq (1/2)^d = \delta$$

- Therefore, with probability $1 - \delta$ there exists an j such that

$$Z_j \leq \epsilon m$$

- Therefore,

$$\begin{aligned} \tilde{f}_x &= \min(c_{1,H_1(x)}, \ldots, c_{j,H_j(x)}, \ldots, c_{d,H_d(x)}) \\ &= \min(f_x + Z_1, \ldots, f_x + Z_j, \ldots, f_x + Z_d) \leq f_x + \epsilon m \end{aligned}$$

Theorem

We can find an estimate $\tilde{f}_x$ for f_x that satisfies,

$$f_x \leq \tilde{f}_x \leq f_x + \epsilon m$$

with probability $1 - \delta$ while only using $O(\epsilon^{-1} \log \delta^{-1})$ memory.

Outline

Applications: Range Queries etc.

Variants

Dyadic Intervals

- Define $\lg n$ partitions of $[n]$

$$\begin{aligned}
\mathcal{I}_0 &= \{1,2,3,4,5,6,7,8,\ldots\} \\
\mathcal{I}_1 &= \{\{1,2\},\{3,4\},\{5,6\},\{7,8\},\ldots\} \\
\mathcal{I}_2 &= \{\{1,2,3,4\},\{5,6,7,8\},\ldots\} \\
\mathcal{I}_3 &= \{\{1,2,3,4,5,6,7,8\},\ldots\} \\
\vdots & \quad \vdots \quad \vdots \\
\mathcal{I}_{\lg n} &= \{\{1,2,3,4,5,6,7,8,\ldots,n\}\}
\end{aligned}$$

- *Exercise:* Any interval $[i,j]$ can be written as the union of $\leq 2\lg n$ of the above intervals. E.g., for $n = 256$,

  $[48,107] = [48,48] \cup [49,64] \cup [65,96] \cup [97,104] \cup [105,106] \cup [107,107]$

  Call such a decomposition, the *canonical decomposition*.

Range Queries and Quantiles

- *Range Query:* For $1 \leq i \leq j \leq n$, estimate $f_{[i,j]} = f_i + f_{i+1} + \ldots + f_j$
- *Approximate Median:* Find j such that

$$\begin{aligned} f_1 + \ldots + f_j &\geq m/2 - \epsilon m \quad \text{and} \\ f_1 + \ldots + f_{j-1} &\leq m/2 + \epsilon m \end{aligned}$$

  Can approximate median via binary search of range queries.
- *Algorithm:*
  1. Construct $\lg n$ Count-Min sketches, one for each $\mathcal{I}_i$ such that for any $I \in \mathcal{I}_i$ we have an estimate $\tilde{f}_I$ for f_I such that

$$\mathbb{P}\left[f_I \leq \tilde{f}_I \leq f_I + \epsilon m\right] \geq 1 - \delta \ .$$

  2. To estimate $[i,j]$, let $I_1 \cup I_2 \cup \ldots \cup I_k$ be canonical decomposition. Set

$$\tilde{f}_{[i,j]} = \tilde{f}_{I_1} + \ldots + \tilde{f}_{I_k}$$

  3. Hence, $\mathbb{P}\left[f_{[i,j]} \leq \tilde{f}_{[i,j]} \leq 2\epsilon m \lg n\right] \geq 1 - 2\delta \lg n$.

Heavy Hitters

- *Heavy Hitter Problem:* For $0 < \epsilon < \phi < 1$, find a set of elements S including all i with $f_i \geq \phi m$ but no elements j with $f_j \leq (\phi - \epsilon)m$.
- *Algorithm:*
  - Consider a binary tree whose leaves are $[n]$ and associate internal nodes with intervals corresponding to descendent leaves.
  - Compute Count-Min sketches for each $\mathcal{I}_i$.
  - Going level-by-level from root, mark children I of marked nodes if
  $$\tilde{f}_I \geq \phi m$$
  - Return all marked leaves.
- Can find heavy-hitters in $O(\phi^{-1} \log n)$ steps of post-processing.

Outline

Applications: Range Queries etc.

Variants

CR-Precis: Count-Min with deterministic Hash functions

- Define t functions $H_i(x) = x \bmod p_i$ where p_i is i-th prime number.
- Maintain $c_{i,j}$ as before.
- Define $z_1, \ldots, z_t$ such that $c_{i,H_i(x)} = f_x + z_i$, i.e.,

$$z_i = \sum_{y \neq x : H_i(y) = H_i(x)} f_y$$

- *Claim:* For any $y \neq x$, $H_j(y) = H_j(x)$ for at most $\lg n$ primes p_j.
- Therefore $\sum_i z_t = m \lg n$ and hence,

$$\tilde{f}_x = \min(c_{1,H_1(x)}, \ldots, c_{t,H_t(x)}) = \min(f_x + z_1, \ldots, f_x + z_t) = f_x + \frac{m \lg n}{t}$$

- Setting $t = (\lg n)/\epsilon$ suffices for $f_x \leq \tilde{f}_x \leq f_x + \epsilon m$.
- Requires keeping $t p_t = O(\epsilon^{-2} \operatorname{polylog} n)$ counters.

Count-Sketch: Count-Min with a Twist

- In addition to $H_i : [n] \to [w]$, use hash functions $r_i : [n] \to \{-1, 1\}$.
- Compute $c_{i,j} = \sum_{x : H_i(x) = j} r_i(x) f_x$.
- Estimate $\hat{f}_x = \text{median}(r_1(x) c_{1,H_1(x)}, \ldots, r_d(x) c_{d,H_1(x)})$
- *Analysis:*
  - *Lemma:* $\mathbb{E}\left[r_i(x) c_{i,H_i(x)}\right] = f_x$
  - *Lemma:* $\mathbb{V}\left[r_i(x) c_{i,H_i(x)}\right] \leq F_2/w$
  - *Chebychev:* For $w = 3/\epsilon^2$,

$$\mathbb{P}\left[|f_x - r_i(x) c_{i,H_i(x)}| \geq \epsilon\sqrt{F_2}\right] \leq \frac{F_2}{\epsilon^2 w F_2} = 1/3$$

  - *Chernoff:* With $d = O(\log \delta^{-1})$ hash functions,

$$\mathbb{P}\left[|f_x - \hat{f}_x| \geq \epsilon\sqrt{F_2}\right] \leq 1 - \delta$$

Count-Sketch Analysis

- Fix x and i. Let $X_y = I[H(x) = H(y)]$ and so

$$r(x)c_{H(x)} = \sum_y r(x)r(y)f_y X_y$$

- *Expectation:*

$$\mathbb{E}\left[r(x)c_{H(x)}\right] = \mathbb{E}\left[f_x + \sum_{y\neq x} r(x)r(y)f_y X_y\right] = f_x$$

- *Variance:*

$$\begin{aligned}
\mathbb{V}\left[r(x)c_{H(x)}\right] &\leq \mathbb{E}\left[(\sum_y r(x)r(y)f_y X_y)^2\right] \\
&= \mathbb{E}\left[\sum_y f_y^2 X_y^2 + \sum_{y\neq z} f_y f_z r(y) r(z) X_y X_z\right] \\
&= F_2/w
\end{aligned}$$