IIT KHARAGPUR

IIT GANDHINAGAR

NPTEL ONLINE CERTIFICATION COURSES

NPTEL

Scalable Data Science

Lecture 10: Frequent Elements: SpaceSaving and CountMin

Anirban Dasgupta

Computer Science and Engineering

IIT GANDHINAGAR

Streaming model revisited

- Data is seen as incoming sequence
  - can be just element-ids, or (id, frequency update) tuple
- Arrival only streams
- Arrival + departure
  - Negative updates to frequencies possible
  - Can represent fluctuating quantities, e.g.

Review: Frequency Estimation in one pass

- Given input stream, length m, want a sketch that can answer frequency queries at the end
  - For give item x, return an estimate of the frequency

- Deterministic algorithm by Misra and Gries
  - f_x = original frequency of item x . Return $\widehat{f_x}$ such that

$$f_x - \epsilon m \leq f_x \leq f_x$$

  - Space = $O(\frac{1}{\epsilon} \log n)$

Space Saving Algorithm

- Keep k counters and items in hand

Initialize:

- Set all counters to 0

Process(x)

- if x is same as any item in hand, increment its counter
- else if number of items $< k$, store x with counter $= 1$
- else replace item with smallest counter by x, increment counter

Query(q)

- If q is in hand return its counter, else 0

Example

Analysis

- Smallest counter value, min, is at most ϵm
  - Counters sum to m, by induction
  - $1/\epsilon$ counters, so average is ϵm, hence smallest is less

Analysis

Claim 1: All items with true count $> \epsilon m$ are present in hand at the end

Analysis

Claim 1: All items with true count $> \epsilon m$ are present in hand at the end

- Smallest counter value, min, is at most ϵm
  - Counters sum to m, by induction
  - $1/\epsilon$ counters, so average is ϵm, hence smallest is less
- True count of an uncounted item is between 0 and min
  - Proof by induction, true initially, min increases monotonically
  - Consider last time the item was dropped

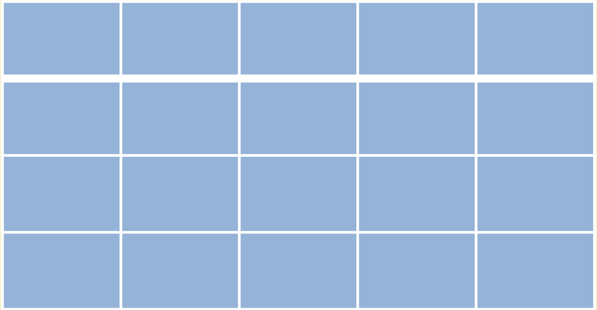
Counter based vs “sketch” based

- Counter based methods
  - Misra-Gries, Space-Saving, ….
  - Work for arrival only streams
  - In practice somewhat more efficient: space, and especially update time

- Sketch based methods
  - “Sketch” is informally defined as a “compact” data structure that allows both inserts and deletes
  - Use hash functions to compute a linear transform of the input
  - Work naturally for arrivals + departure

Count-min sketch

- Model input stream as a vector over U
  - f_x is the entry for dimension x
- Creates a small summary $w \times d$
- Use w hash functions, each maps $U \rightarrow [1, d]$



Count Min Sketch

Initialize

- Choose $h_1, .., h_w,\ A[w, d] \leftarrow 0$

Process(x, c):

- For each $i \in [w],\ A[i, h_i(x)] \mathrel{+}= c$

Query(q):

- Return $\min_i A[i, h_i(x)]$

Example

h1			
h2			

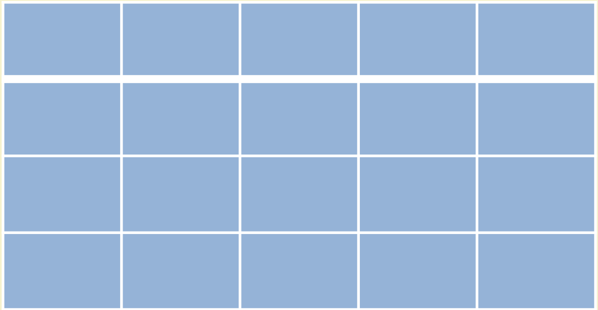
	h1	h2
	2	1
	1	2
	1	3
	3	2

Guarantees

Space = $O(wd)$

Update time = $O(w)$

$x, +c$



Each item is mapped to one bucket per row

Guarantees

- $w = \frac{2}{\epsilon} \quad \mathrm{d} = \log\left(\frac{1}{\delta}\right)$

$Y_1 \ldots . Y_w$ be the w estimates, i.e. $Y_i = A[i, h_i(x)], \quad \widehat{f_x} = \min_i Y_i$

Each estimate $\widehat{f_x}$ always satisfies $\widehat{f_x} \geq f_x$

Guarantees

- $w = \frac{2}{\epsilon} \quad \mathrm{d} = \log\left(\frac{1}{\delta}\right)$

$Y_1 \ldots . Y_w$ be the w estimates, i.e. $Y_i = A[i, h_i(x)], \quad \widehat{f_x} = \min_i Y_i$

Each estimate $\widehat{f_x}$ always satisfies $\widehat{f_x} \geq f_x$

$E[Y_i] = \sum_{y:h_i(y)=h_i(x)} f_y = f_x + \epsilon(m - f_x)/2$

Guarantees

- $w = \frac{2}{\epsilon} \quad \mathrm{d} = \log\left(\frac{1}{\delta}\right)$

$Y_1 \ldots . Y_w$ be the w estimates, i.e. $Y_i = A[i, h_i(x)]$, $\quad \widehat{f_x} = \min_i Y_i$

Each estimate $\widehat{f_x}$ always satisfies $\widehat{f_x} \geq f_x$

$$E[Y_i] = \sum_{y:h_i(y)=h_i(x)} f_y = f_x + \epsilon(m - f_x)/2$$

Applying Markov's inequality,

$$\Pr[\, Y_i - f_x > \epsilon m \,] \leq \frac{\epsilon(m - f_x)}{2\epsilon m} \leq \frac{1}{2}$$

Guarantee

- Since we are taking minimum of $\log\left(\frac{1}{\delta}\right)$ such random variables,

$$\Pr\left[\,\widehat{f_x} > f_x + \epsilon m\right] \le 2^{-\log\left(\frac{1}{\delta}\right)} \le \delta$$

Guarantee

- Since we are taking minimum of $\log\left(\frac{1}{\delta}\right)$ such random variables,

$$\Pr\left[\,\widehat{f_x} > f_x + \epsilon m\right] \le 2^{-\log\left(\frac{1}{\delta}\right)} \le \delta$$

- Hence, with probability $1 - \delta$, for any query x

$$f_x \le \widehat{f_x} \le f_x + \epsilon m$$

Summary

- Two algorithms for frequency estimation
  - Counter based: Space Saving
  - Sketch based: Count-Min
- Guiding principle: use error bounds as design parameters of the data structure
- More to come…

References:

- Primary references for this lecture
  - Lecture slides by Graham Cormode
    http://dmac.rutgers.edu/Workshops/WGUnifyingTheory/Slides/cormode.pdf
  - Lecture notes by Amit Chakrabarti: http://www.cs.dartmouth.edu/~ac/Teach/data-streams-lecnotes.pdf
  - Sketch techniques for approximate query processing, Graham Cormode.
    http://dimacs.rutgers.edu/~graham/pubs/papers/sk.pdf

Thank You!!