CS60021: Scalable Data Mining

Streaming Algorithms

Sourangshu Bhattacharya

Count distinct

Handling velocity + volume

- Can we process data without explicitly storing all of it in memory? E.g. in a network switch,
  - which IPs have most packets passing through a switch
  - has traffic pattern changed overnight?

- We have to give up on exact answer, and rely on…
  - approximation: return answer close to truth
  - randomization: be correct only with high probability

Streaming model: sketches

- Data is assumed to come as a stream of values
  - e.g. bytes seen when reading off a tape-drive
  - destination IPs seen by a network switch
- Size of universe/stream is much large compared to available memory
  - typically assume memory is $poly(\log)$
  - Can make limited (possibly single) pass over data
  - Will create a “sketch” : a summary data structure used to answer queries at the end

Streaming problem: distinct count

- Universe is U, number of distinct elements = n, stream size is m
  - Example: $U =$ all IP addresses
    10.1.21.10, 10.93.28,1,.....,98.0.3.1,.....10.93.28.1.....
  - IPs can repeat
  - Want to estimate the number of distinct elements in the stream

Other applications

- Universe = set of all k-grams, stream is generated by document corpus
  - need number of distinct k-grams seen in corpus
- Universe = telephone call records, stream generated by tuples (caller, callee)
  - need number of phones that made > 0 calls

Solutions

- Naïve solution: $O(|S|log(n))$ space
  - Store all elements, sort and count distinct
  - Store a hashmap, insert if not present
- Bit array: O(|n|) space:
  - Bits initialized to 1 only if element seen in stream
- Can we do this in less space ?

Approximations

- (ϵ, δ) $-$approximations
  - Algorithm will use random hash functions
  - Will return an answer $\hat{n}$ such that

$$(1 - \epsilon)n \leq \hat{n} \leq (1 + \epsilon)n$$

  - This will happen with probability $1 - \delta$ over the randomness of the algorithm

First effort

- Stream length: m, universe size: n
- Proposed algo: Given space S, sample S items from the stream
  - Find the number of distinct elements in this set: $\hat{n}$
  - return $\hat{n} \times \frac{m}{S}$

First effort

- Stream length: m, distinct elements: n
- Proposed algo: Given space S, sample S items from the stream
  - Find the number of distinct elements in this set: $\hat{n}$
  - return $\hat{n} \times \frac{m}{S}$
- Not a constant factor approximation
  - $\underbrace{1,1,1,1,\ldots..1}_{m-n+1},2,3,4,\ldots.,n-1$

Linear Counting

- Bit array B of size m, initialized to all zero
- Hash function h: $[n] \rightarrow [m]$
- When seeing item x , set $B[h(x)] = 1$

Linear Counting

- Bit array B of size m, initialized to all zero
- Hash function h: $[n] \rightarrow [m]$
- When seeing item x , set $B[h(x)] = 1$
- $z_m =$ fraction of zero entries
- Return estimate $-\mathrm{m}\log(\frac{z_m}{m})$

Linear Counting Analysis

- Pr[position remaining 0] = $\left(1 - \frac{1}{m}\right)^n \approx e^{-\frac{n}{m}}$
- Expected number of positions at zero = $\mathrm{E}[\mathrm{z_m}] = me^{-n/m}$
- Using tail inequalities we can show this is concentrated
- Typically useful only for $m = \Theta(n)$, often useful in practice

Flajolet Martin Sketch

- Components
  - "random" hash function $h: U \rightarrow 2^{\ell}$ for some large ℓ
  - $h(x)$ is a ℓ −length bit string
  - initially assume it is completely random, can relax
- $zero(v)$ = position of rightmost 1 in bit representation of v

  $= \max\{\, i \,, 2^{i} \ divides\ v \,\}$

  - $zeros(10110) = 1, \quad zeros(110101000) = 3$

Flajolet Martin Sketch

Initialize:

– Choose a “random” hash function $h: U \to 2^\ell$

– $z \leftarrow 0$

Process(x)

– if $zeros\big(h(x)\big) > z,\ \ \mathrm{z} \leftarrow zeros(h(x))$

Estimate:

– return $2^{z+1/2}$

Example

	h(.)
	0110101
	1011010
	1000100
	1111010

Space usage

- We need $\ell \geq C \log(n)$ for some $C \geq 3$, say
  - by birthday paradox analysis, no collisions with high prob

- Sketch : z , needs to have only $O(\log \log n)$ bits !!!
- Total space usage = $O(\log n + \log \log n)$

Intuition

- Assume hash values are uniformly distributed
- The probability that a uniform bit-string
  - is divisible by 2 is ½
  - is divisible by 4 is ¼
  - ….
  - is divisible by 2^k is $\frac{1}{2^k}$
- We don't expect any of them to be divisible by $2^{\log_2(n)+1}$

Formalizing intuition

- $S =$ set of elements that appeared in stream
- For any $r \in [\ell], j \in U,\ \ X_{rj} =$ indicator of $\text{zeros}(h(j)) \geq r$
- $Y_r =$ number of $j \in U$ such that $\text{zeros}(h(j)) \geq r$

$$Y_r = \sum_{j \in S} X_{rj}$$

- Let $\hat{z}$ be final value of z after algo has seen all data

Proof of FM

- $Y_r > 0 \;\leftrightarrow \hat{z} \geq r$, equivalently, $Y_r = 0 \leftrightarrow \hat{z} < r$

Proof of FM

- $Y_r > 0 \leftrightarrow \hat{z} \geq r$, equivalently, $Y_r = 0 \leftrightarrow \hat{z} < r$

- $E[Y_r] = \sum_{j \in S} E\left[X_{rj}\right]$ $\quad X_{rj} = \begin{cases} 1 \;\; with\, prob\; \dfrac{1}{2^r} \\ 0 \qquad\qquad\quad else \end{cases}$

- $E[Y_r] = \frac{n}{2^r}$ $\quad var(Y_r) = \sum_{j \in S} var\left(X_{rj}\right) \;\; \leq \sum_{j \in S} E\left[X_{rj}^2\,\right]$

Proof of FM

- $var(Y_r) \leq \sum_{j \in S} E\left[X_{rj}^2\right] \leq \; n/2^r$

$$\Pr[Y_r > 0] = \Pr[Y_r \geq 1] \leq \frac{E[Y_r]}{1} = \frac{n}{2^r}$$

Proof of FM

- $var(Y_r) \leq \sum_{j \in S} E\left[X_{rj}^2\right] \leq \; n/2^r$

$$\Pr[Y_r > 0] = \Pr[Y_r \geq 1] \leq \frac{E[Y_r]}{1} = \frac{n}{2^r}$$

$$\Pr[Y_r = 0] \leq \Pr\left[\,|Y_r - E[Y_r]| \geq E[Y_r]\right] \leq \frac{var(Y_r)}{E[Y_r]^2} \leq \frac{2^r}{n}$$

Upper bound

Returned estimate $\hat{n} = 2^{\hat{z}+1/2}$

a = smallest integer with $2^{a+1/2} \geq 4n$

$$\Pr[\hat{n} \ \geq 4n\ ] = \Pr[\ \hat{z} \geq a\ ] = \Pr[Y_a > 0] \leq \frac{n}{2^a} \leq \frac{\sqrt{2}}{4}$$

Lower bound

Returned estimate $\hat{n} = 2^{\hat{z}+1/2}$

b = largest integer with $2^{b+1/2} \leq n/4$

$$\Pr\left[\hat{n} \leq \frac{n}{4}\right] = \Pr[\ \hat{z} \leq b\ ] = \Pr[Y_{b+1} = 0] \leq \frac{2^{b+1}}{n} \leq \frac{\sqrt{2}}{4}$$

Understanding the bound

- By union bound, with prob $1 - \frac{\sqrt{2}}{2}$,

$$\frac{n}{4} \leq \hat{n} \leq 4n$$

- Can get somewhat better constants
- Need only 2-wise independent hash functions, since we only used variances

Improving the probabilities

- To improve the probabilities, a common trick: median of estimates
- Create $\widehat{z_1}, \widehat{z_2}, \ldots., \widehat{z_k}$ in parallel
  - return median
- Expect at most $\frac{\sqrt{2}}{4}$ of them to exceed $4n$

Improving the probabilities

- To improve the probabilities, a common trick: median of estimates
- Create $\widehat{z_1}, \widehat{z_2}, \ldots., \widehat{z_k}$ in parallel
  - return median
- Expect at most $\frac{\sqrt{2}}{4}k$ of them to exceed $4n$
- But if median exceeds $4n$, then $\frac{k}{2}$ of them does → using Chernoff bound this prob is $\exp(-\Omega(k))$

Improving the probabilities

- To improve the probabilities, a common trick: median of estimates
- Create $\widehat{z_1}, \widehat{z_2}, \ldots, \widehat{z_k}$ in parallel
  - return median

- Using Chernoff bound, can show that median will lie in $\left[\frac{n}{4}, 4n\right]$ with probability $1 - \exp(-\Omega(k))$.
- Given error prob δ, choose $k = O(\log\left(\frac{1}{\delta}\right))$

Summary

- Streaming model– useful abstraction
  - Estimating basic statistics also nontrivial

- Estimating number of distinct elements
  - Linear counting
  - Flajolet Martin

References:

- Primary reference for this lecture
  - Lecture notes by Amit Chakrabarti: http://www.cs.dartmouth.edu/~ac/Teach/data-streams-lecnotes.pdf