Distinct Elements

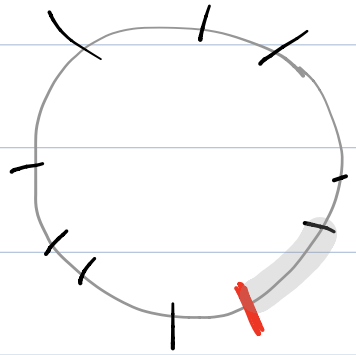
$a_1, a_2, \ldots, a_T$ from universe U, $|U| = m$ elts can be repeated

Streaming alg to determine $n = \#$ distinct elts?

$h: U \to [0,1]$ u.a.r. Ndk: if $a_i = a_j \Rightarrow h(a_i) = h(a_j)$

$$Y = \min_{x \in S} h(x)$$

$$E(Y) = \int_0^1 \Pr(Y \geq z)\, dz = \int_0^1 (1-z)^n dz = \frac{1}{n+1}$$



$$\sigma^2(Y) = \frac{1}{(n+1)^2}$$

Chebychev Inequality

X has mean μ, var σ^2

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$\Rightarrow \Pr(X \in [\mu - k\sigma, \mu + k\sigma]) \geq 1 - \frac{1}{k^2}$$

that in absence of other information

This means X is useful estimate of μ if σ small relative to μ.

How to reduce error?

Generate k indep instances of r.v. $X_1, \ldots, X_k$

and take average

$E(X_i) = \mu \quad Var(X_i) = \sigma^2$

$$E(\bar{X}) = E\left(\frac{1}{k}\sum_{i=1}^{k} X_i\right) = \mu$$

$$Var(\bar{X}) = \frac{1}{k^2} Var\left(\sum_{i=1}^{k} X_i\right) = \frac{1}{k^2} \sum_{i=1}^{k} Var(X_i) = \frac{k\sigma^2}{k^2} = \frac{\sigma^2}{k}$$

indep r.v.'s

So $\sigma(\bar{X}) = \frac{\sigma(X_i)}{\sqrt{k}}$

For minhash $\sigma = \mu$

To reduce $\frac{\sigma}{\mu}$, take avg of several indep samples

$$Y_1 = \min_{x \in S} h_1(x), \ldots, Y_k = \min_{x \in S} h_k(x)$$

$h_i : U \to [0,1]$

h_i's independent

$$\bar{Z} = \frac{1}{k}\sum_{i=1}^{k} Y_i \qquad E(\bar{Z}) = E(Y_1) \qquad \sigma^2(\bar{Z}) = \frac{k\,\sigma^2(Y_1)}{k^2} = \frac{\sigma^2(Y_1)}{k}$$

$\sigma^2\left(\sum_{i=1}^{k} Y_i\right)$

reduced variance

Applying Chebychev to $\bar{Z}$ with $\frac{1}{k} \le \varepsilon^2 \Rightarrow \frac{\sigma(Y_1)}{\sqrt{k}} \le \varepsilon\,\sigma(Y_1)$

$$\bar{Z} \in \left[\mu - c\sigma(\bar{Z}), \mu + c\sigma(\bar{Z})\right] = \left[\frac{1}{n+1} - \frac{c\varepsilon}{n+1}, \frac{1}{n+1} + \frac{c\varepsilon}{n+1}\right] \text{ w.p. } \ge 1 - \frac{1}{c^2}$$

Example: $c = 10$

Pairwise indep hash fns (2-universal)

$h: U \longrightarrow [m]$

$\Pr(h(x)=y) \leq \frac{1}{m}$

$\forall x_1 \neq x_2 \qquad \Pr(h(x_1)=y_1, h(x_2)=y_2) \leq \frac{1}{m^2}$

assuming h selected at random from pairwise indep family

Implementation with pairwise indep family :(

[Flajolet, Martin]

each $h \in \mathcal{H}$

$h: [m] \to [m]$

$U = [m]$

$j \in [m]$

Defn: $\text{zeroes}(j) \triangleq \max\{r \mid 2^r \text{ divides } j\}$

i.e. # trailing 0's in binary representation of j

log m bits

min elt has largest # of leading 0's

instead, look at trailing zeros

leading 0's roughly log n

$\Pr(\text{elt ends w/} \geq k \text{ 0's}) = 2^{-k}$

$\Pr(\text{none have} \geq k) = (1-2^{-k})^n \simeq e^{-2^{-k}n}$

$e^{-\frac{n}{2^k}}$

$k = \log n \Rightarrow$ const

$k \gg \log n \Rightarrow$ big $\Rightarrow e^0$

$k \ll \log n \Rightarrow$ tiny

$n = \#$ of distinct elements

Estimation alg

Choose h from $\mathcal{H}$ u.a.r.

$z := 0$

for each i do

$z := \max(\text{zeroes}(h(a_i)), z)$

output $\tilde{n} := 2^{z+\frac{1}{2}}$ (geometric mean of 2^z & 2^{z+1})

Space used: $2 \log m$ for hash fns $\log\log m$ for z

Claim: with constant prob $\frac{n}{3} \leq \tilde{n} \leq 3n$

Proof: let z^* be final value of z

$$X_{r,j} \triangleq \begin{cases} 1 & \text{zeroes}(h(j)) \geq r \\ 0 & \text{o.w.} \end{cases}$$

$$Y_r \triangleq \sum_{j \mid f_j > 0} X_{r,j}$$

$= \#$ of a_i with at least r trailing zeroes

$$E(Y_r) = \sum_{j \text{ s.t. } f_j > 0} E(X_{r,j}) = \sum_{j \text{ s.th } f_j > 0} \Pr(2^r \text{ divides } h(j)) = \frac{n}{2^r}$$

$$\text{Also } \mathrm{Var}(Y_r) = E(Y_r^2) - E(Y_r)^2 \leq E(Y_r) \qquad (**)$$

$$E(Y_r^2) = \sum_j \sum_k E(X_{r,j} X_{r,k})$$

$X_{r,j}$'s pairwise indep

$$= \sum_j E(X_{r,j}^2) + \sum_j \sum_{k \neq j} E(X_{r,j}) E(X_{r,k})$$

$$\leq \sum_j E(X_{r,j}) + E(Y_r)^2$$

Let a be smallest int s.t. $2^{a+\frac{1}{2}} \geq 3n$ (*)

$$\Pr(\tilde{n} \geq 3n) = \Pr(z^* \geq a)$$

$$= \Pr(Y_a > 0)$$

$$= \Pr(Y_a \geq 1)$$

$$\leq E(Y_a) \quad \text{Markov's Inequality}$$

$$= \frac{n}{2^a}$$

$$\leq \frac{\sqrt{2}}{3} \approx .471 \quad (*)$$

Let b be smallest int s.t. $2^{b+\frac{1}{2}} \leq \frac{n}{3}$ (**)

$$\Pr(\tilde{n} \leq \tfrac{n}{3}) = \Pr(z^* \leq b)$$

$$= \Pr(Y_{b+1} = 0)$$

$$\leq \Pr(|Y_{b+1} - E(Y_{b+1})| \geq E(Y_{b+1}))$$

$$\leq \frac{\mathrm{Var}(Y_{b+1})}{E(Y_{b+1})^2} \quad \text{Chebychev}$$

$$\leq \frac{1}{E(Y_{b+1})} \quad (**)$$

$$= \frac{2^{b+1}}{n} \leq \frac{\sqrt{2}}{3} \quad (**)$$

$Pr(\text{success}) \leq 0.05$ but each of failure conditions happens with prob

< 0.48

so if repeat k times, Prob happens in majority $= e^{-ck}$

for some c by Chernoff.

$\Rightarrow$ median good with prob $1 - e^{-ck}$ for some $c > 0$

Application: Estimating document similarity

Think of document as set of words (shingles)

A definition of similarity

Jaccard Similarity of docs A & B: $\frac{|A \cap B|}{|A \cup B|}$

Cool observation:

$h: \underbrace{U}_{\text{universe of shingles}} \to [0,1]$ at random

$$MH(A) = \min_{x \in A} h(x)$$

$$MH(B) = \min_{x \in B} h(x)$$

$$Pr(MH(A) = MH(B)) = Pr(\text{elt in } A \cap B \text{ yields min hash})$$

$$= \frac{|A \cap B|}{|A \cup B|}$$

To get this, need each elt in $A \cup B$ equally likely to be min

For example: suppose want to flag all docs whose $JS \geq 0.9$

compute k MHs & flag pair if $\geq 0.9-\varepsilon$ of time $MH(A)=MH(B)$

Chernoff bounds $\Rightarrow$

$$X_i = \begin{cases} 1 & \text{if } MH_{h_i}(A) = MH_{h_i}(B) \\ 0 & \text{o.w} \end{cases}$$

if $JS(A,B) \geq 0.9$
$\Rightarrow E(X_i) \geq 0.9$

$$E\left(\sum_{i=1}^{k} X_i\right) \geq 0.9k$$

$JS(A,B) \geq 0.9$

$$Pr\left(\Sigma X_i < (0.9-\varepsilon)k\right) \leq e^{-c\varepsilon^2 k}$$

If $JS(A,B) < 0.9-3\varepsilon$ $\quad Pr\left(\Sigma X_i > (0.9-\varepsilon)k\right) \leq e^{-c'\varepsilon^2 k}$