lec 12

Aside if X rv $\in \mathbb{Z}$ $\quad E(X) = \sum_{k} k \cdot P(X=k)$

if $X \geq 0$ $= \sum_{k=1}^{\infty} P(X \geq k)$

$X \in \mathbb{R}$ then $E(X) = \int_{-\infty}^{\infty} u \cdot P(u)\, du$

if $X \geq 0$ $= \int_{0}^{\infty} Pr(X \geq u)\, du$

$X = \begin{cases} 0 & \text{w.p. } 1-p \\ 1 & \text{w.p. } p \end{cases}$

$Var(X) = E(X^2) - E(X)^2$ → $Var(X) = p(1-p)$

if X, Y ind $\Rightarrow$ $Var(X+Y) = Var(X) + Var(Y)$

Document duplicate detection

- Hashing for exact-duplicate detection:
  - Pick $h: Docs \to \{0,1\}^b$ from 2-wise independent family.
  - $\Pr(h(A) = h(B)) = \begin{cases} 1 & \text{if doc. } A \text{ is } \textbf{identical} \text{ to doc. } B \\ 2^{-b} & \text{if } A \neq B \end{cases}$

→ only for exact matches

Consider → a document is a sequence of words. Lets transform into set that captures some of this

("quick brown fox jumped") → {"quick brow", "brown fox", ...}

→ so here we care abt set similarity

→ similarity → not

if $A = B \Rightarrow Sim(A,B) = 1$

if $A \cap B = \emptyset \Rightarrow Sim(A,B) = 0$

→ we will use Jaccard Similarity: $Sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$

MinHash

- Pick $h: [u] \to [m]$ (want $m \gg$ size of any set)
- The basic sketch: $sk(S) = \min\{h(x) \mid x \in S\}$.
- To compare sketches of two sets S_1, S_2:

$$Y = \begin{cases} 0 & \text{if } sk(S_1) \neq sk(S_2) \\ 1 & \text{if } sk(S_1) = sk(S_2) \end{cases}$$

- What is $E(Y)$?
  - $Y = 1$ if
    - The element in $x \in S_1 \cup S_2$ with minimum $h(x)$ lies in $S_1 \cap S_2$.
    - The elements $x_1 \in S_1, x_2 \in S_2$ with minimum hash values have $h(x_1) = h(x_2)$ but $x_1 \neq x_2$.
  - $E(Y) \leq \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} + \frac{|S_2|}{m} = sim(S_1, S_2) + o(1)$.

→ want to in big numbers

→ in expectation

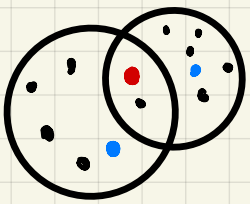
$E(Y) = Pr(Y=1) \Rightarrow Pr(sk(S_1) = sk(S_2))$

→ we want minimal val in $S_1 \cup S_2$ to be in $S_1 \cap S_2$ → so we need

$CoPr = Sim(A,B)$

or we could just map different things that hash to same min val...

$Coprr < \frac{|S_2|}{m}$



But the above isn't great → Y is only 0,1 so we will get its value in expectation by reducing variance

MinHash

- Y is an ***unbiased*** estimate of $sim(S_1, S_2)$ but it has a ***large variance***. How can variance be reduced?
- The sketch: $SK(S) = (sk_1(S), sk_2(S), \dots, sk_t(S))$
- Estimate is $\widetilde{sim}(SK(S_1), SK(S_2)) = (\sum_{i=1}^{t} Y_i)/t$, where:

$$Y_i = \begin{cases} 0 & \text{if } sk_i(S_1) \neq sk_i(S_2), \\ 1 & \text{if } sk_i(S_1) = sk_i(S_2). \end{cases}$$

- $E\left(\widetilde{sim}(SK(S_1), SK(S_2))\right) = sim(S_1, S_2) + o(1)$.
- $\text{Var}\left(\widetilde{sim}(SK(S_1), SK(S_2))\right) = \frac{1}{t} sim(S_1, S_2)(1 - sim(S_1, S_2)) + o(1)$
- Space?
  - $t \cdot \log m$ bits.

output per sketch

each sketch is a hash function

$$\text{Var}\left(\frac{1}{t}\sum_{i=1}^{t} Y_i\right) \overset{\text{ind}}{=} \frac{1}{t^2}\sum_{i=1}^{t} \text{Var}(Y_i) \overset{\text{iid}}{=} \frac{1}{t}\text{Var}(Y_i) \overset{\text{bernoulli}}{=} \frac{1}{t} sim(S_1, S_2)(1 - sim(S_1, S_2))$$

can reduce by ↑ t

how do we reduce space

$\Pr(Y_i = 1)$

MinHash

- Reducing space: pick hash functions
  - $h: [u] \to [u]$
  - $g: [u] \to \{0,1\}$
  - Basic sketch: $sk(S) = g(x)$, where $x = \text{argmin}_{x \in S}(h(x))$.
  - $SK(S) = (sk_1(S), sk_2(S), \dots, sk_t(S))$

$$Y_i = \begin{cases} 0 & \text{if } sk_i(S_1) \neq sk_i(S_2), \\ 1 & \text{if } sk_i(S_1) = sk_i(S_2). \end{cases}$$

- $E(Y_i) = sim(S_1, S_2) + (1 - sim(S_1, S_2))/2 = (1 + sim(S_1, S_2))/2$.
- $\widetilde{sim}(SK(S_1), SK(S_2)) = (\sum_{i=1}^{t}(2Y_i - 1))/t$.
- $E\left(\widetilde{sim}(SK(S_1), SK(S_2))\right) = sim(S_1, S_2)$.
- $\text{Var}\left(\widetilde{sim}(sk(S_1), sk(S_2))\right) = \frac{4}{t} \cdot \left(\frac{1+sim(S_1,S_2)}{2}\right)\left(\frac{1-sim(S_1,S_2)}{2}\right)$.
- Space?
  - t bits.

Case 1) if min x in intersection then $\Pr(Y_i = 1) = sim(S_1, S_2)$

Case 2) if outside its 50/50 (same as 0).

$\Rightarrow (1 - sim(S_1, S_2))\frac{1}{2}$

$$\Rightarrow E(Y_i) = sim(S_1, S_2) + \frac{(1 - sim(S_1, S_2))}{2} = \frac{(1 + sim(S_1, S_2))}{2}$$

$$\Rightarrow Y = \sum_{i=1}^{t} \frac{(2Y_i - 1)}{t}$$ → in exp it is $sim(S_1, S_2)$

$$\text{Var}(Y) = \frac{Y}{t} \cdot \left(\frac{1+sim(S_1,S_2)}{2}\right)\left(\frac{1-sim(S_1,S_2)}{2}\right)$$

In practice $(sk_1, \dots, sk_t)$ → min t hash values not min for t hash function

single hash

Distinct Elements in Data Streams

- Input is a stream (multiset) $x_1, x_2, \dots, x_N \in [u]^N$.
- How many ***distinct elements*** were seen?
  - Estimate $n = |\{x_1, x_2, \dots, x_N\}|$: return some $\hat{n} \approx n$.
- Assume we have an ideal hash function $h: [u] \to [0,1]$.
- At time t, maintain $X = \min(h(x_1), \dots, h(x_t))$.

Distinct Elements in Data Streams

- What is $E(X)$ = ? (1-st prob to be above in n times)
  - $\int_0^\infty \Pr(X > u)\, du = \int_0^1 (1-u)^n du = \int_0^1 u^n du = \frac{1}{n+1}$.
- What is $\text{Var}(X) = E\left((X - E(X))^2\right) = E(X^2) - E(X)^2$ = ?
  - $E(X^2) = \int_0^1 \Pr(X^2 > u)\, du = \int_0^1 \Pr(X > \sqrt{u})\, du$
  - $= \int_0^1 (1-\sqrt{u})^n du = \int_0^1 (1-v)^n 2v\, dv = 2\int_0^1 v^n(1-v)\, dv$
  - $= 2\left(\frac{1}{n+1} - \frac{1}{n+2}\right) = \frac{2}{(n+1)(n+2)}$.
  - $\text{Var}(X) = \frac{2}{(n+1)(n+2)} - \frac{1}{(n+1)^2} < \frac{1}{(n+1)^2}$.
- $Y = \frac{(X_1 + \cdots + X_m)}{m}$ (avg. of m independent sketches)
- $\text{Var}(Y) = \frac{1}{m^2}\text{Var}(X_1 + \cdots + X_m) = \frac{1}{m}\text{Var}(X_1) < \frac{1}{m} \cdot \frac{1}{(n+1)^2}$.

- (Markov's Ineq.: $X \geq 0$) $\Pr(X \geq t) \leq \frac{E(X)}{t}$.
- (Chebychev: apply Markov to 2nd moment)

$$\Pr(|X - E(X)| \geq t) \leq \frac{E\left(\left(X - E(X)\right)^2\right)}{t^2} = \frac{Var(X)}{t^2}$$

- Estimate $\frac{1}{Y} \in \left[\frac{n+1}{1+\epsilon}, \frac{n+1}{1-\epsilon}\right]$ whenever $Y \in \frac{1}{n+1} \pm \frac{\epsilon}{n+1}$.
- $\Pr\left(\left|Y - \frac{1}{n+1}\right| > \frac{\epsilon}{n+1}\right) < \frac{\text{Var}(Y)}{\left(\frac{\epsilon}{n+1}\right)^2} < \frac{\frac{1}{m(n+1)^2}}{\left(\frac{\epsilon}{n+1}\right)^2} = \frac{1}{m\epsilon^2}$.

So m grows quad with ε

If $n=1 \to E(X) = \frac{1}{2}$

$n=2 \to E(X) = \frac{1}{3}$

$n=100$ $E(X) = \frac{1}{101}$

$\Pr(\hat{n} \in [\frac{n}{1+\epsilon}, \frac{n}{1-\epsilon}])$

from exp

So as long as Y not too far away, we're good