Min-Hashing

Jaccard Sim

Sets → Matrix → Binary Vector

hashing

↓

Compare

LSH

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$A = \{0, 1, 2, 5, 6\}$

$B = \{0, 2, 3, 5, 7\}$

$$= \frac{|\{0, 2, 5\}|}{|\{0, 1, 2, 3, 5, 6, 7\}|} = \frac{3}{7}$$

$S_1 = \{1, 2, 5\}$

$S_2 = \{3\}$

$S_3 = \{2, 3, 4, 5\}$

$S_4 = \{1, 4, 6\}$

$$JS(S_1, S_3) = \frac{|\{2,5\}|}{|\{1,2,3,4,5\}|} = \frac{2}{5}$$

$$\boxed{E[\hat{JS}(S_i, S_j)] = JS(S_i, S_j)}$$

	S_1	S_2	S_3	S_4
1	1	0	0	1
2	1	0	1	0
3	0	1	1	0
4	0	0	1	1
5	1	0	1	0
6	0	0	0	1

mostly 0s

$\Longrightarrow$ random reorder rows $\times k$ times

	S_1	S_2	S_3	S_4
2	1	0	1	0
5	1	0	1	0
6	0	0	0	1
1	1	0	0	1
4	0	0	1	1
3	0	1	1	0
m	2	3	2	6

$$\hat{JS}(S_i, S_{i'}) = \begin{cases} 1 & \text{if } m(S_i) = m(S_{i'}) \\ 0 & \text{otherwise} \end{cases}$$

k random reorders $j = 1, 2, \dots k$

Set $S_{i=3}$ value $\left(m_1(S_i), m_2(S_i), \dots m_k(S_i)\right)$

$$V_i = V_3 = (2, 7, 10023, \dots, 18)$$

$$S_{i'=1} \Rightarrow V_{i'} = V_1 = (3, 7, 42, \dots, 8)$$

$$\quad\ \ 0 \quad 1 \quad 0 \quad \dots \quad 1$$

$$\hat{JS}_j(S_i, S_{i'}) = \begin{cases} 1 & \text{if } m_j(S_{i'}) = m_j(S_i) \\ 0 & \text{o.w.} \end{cases}$$

S_3, S_1 $\qquad S_1 \qquad S_3$

$$\hat{JS} = \frac{1}{k} \sum_{j=1}^{k} \hat{JS}_j(S_{i'}, S_i)$$

Why $E[\hat{JS}(S_1, S_2)] = JS(S_1, S_2)$

3 types of rows

(T_x) = # rows S_1 and S_2 have 1

(T_y) = # rows exactly 1 of S_1, S_2 have 1
other has 0

(T_z) = # rows S_1 and S_2 have 0.

$$JS(S_1, S_2) = \frac{T_x}{T_x + T_y} \Longrightarrow \text{can ignore } (T_z)$$

Collision $m(S_1) = m(S_2)$ iff of T_x and T_y rows
a T_x is at the top.

$$\text{Prob}[\text{collision}] = E[\hat{JS}] = \frac{T_x}{T_x + T_y} = JS$$

Top k - Stretches

k = 2

	s_1
1	0
5	1
2	0
3	1
6	1
4	0

$m(s_1) = (5, 3)$

Fast Min Hashing Algorithm

Replace re-orders w/ hash fxns.

Hash fxns $h_1, h_2, \ldots, h_k$

h_j : {element of set} $\to [n]$

$= \{1, 2, \ldots 1024\}$

$S \to V = (v_1, v_2, \ldots v_k)$

v_2 replace $m_2(S)$

v_j replace $m_j(S)$

for $i \in S$

  for $j = 1$ to k (all hash)

    if $(h_j(i) \le v_j)$

      $v_j \leftarrow h_j(i)$

$S = \{1, 3, 6\}$

	S
1	1
2	0
3	1
4	0
5	0
6	1

h_1:
$1 \to 2$
$2 \to 4$
$3 \to 6$
$4 \to 4$
$5 \to 1$
$6 \to 3$

h_2:
$1 \to 3$
$2 \to 5$
$3 \to 1$
$4 \to 6$
$5 \to 2$
$6 \to 6$

$V_1 = \infty$

1 $V_1 = 2$

3 $V_1 = 2$ $2 < 6$

6 $V_1 = 2$ $2 < 3$

$V_2 = 1$

hash fxns are fixed for all sets

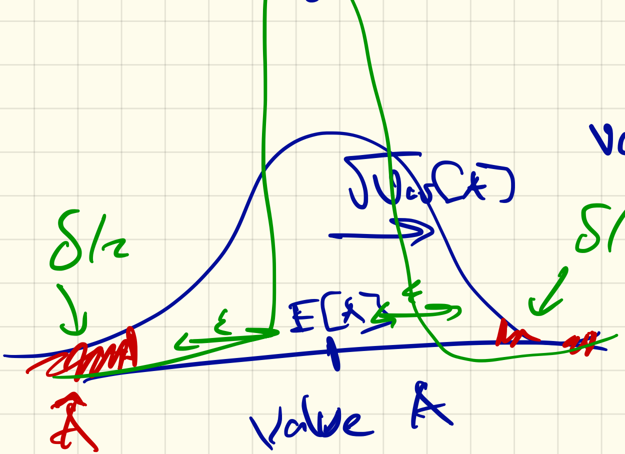
$H : \{h_1, \ldots h_k)$

$f_H(S) \to \mathbb{R}^k \equiv V_S$

Central Limit Theorem

$X = \{X_1, X_2, \ldots X_r\}$ r random variables iid.

$$A = \frac{1}{r}\sum_{i=1}^{r} X_i$$



$$E[A] = E[X_i]$$

$$var(A) = \frac{Var[X_i]}{r}$$

Converge to Normal

Probably Aprox Correct

$$\Pr\left[\,|A - E[x_i]| > \varepsilon\,\right] < \delta$$

estimate

what I want

error tolerance

probability of failure

Want ε, δ small

Algorithm has r steps

Chernoff - Hoeffding Bound

iid $X_1, \ldots, X_r$ $\qquad A = \frac{1}{r}\sum_{i=1}^{r} X_i$

$r = k$

$\hat{JS}$ $\quad \Delta = 1$ $\qquad \Delta = b - a$ s.t. $X_i \in [a, b]$

$\qquad\qquad a = 0, \; b = 1$

$$\Pr\left[\,|A - E[X_i]| > \varepsilon\,\right] \le 2\exp\left(\frac{-2r\varepsilon^2}{\Delta^2}\right)$$

$$\Pr\left[\,|\bar{JS} - JS| > \underbrace{0.1}_{\varepsilon}\,\right] \le 2\exp\left(\frac{-2k(0.1)^2}{1}\right)$$

$\boxed{k = 500}$

$$\le 2\exp\left(\frac{-2k}{100}\right) = 2\exp\left(\frac{-1000}{100}\right) = 2e^{-10}$$