L4 Min Hashing

Wednesday, January 15, 2020 3:17 PM

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

Ex. $A = \{0,1,2,3,6\}$
$B = \{1,2,4,6,8\}$

$$JS(A,B) = \frac{\{1,2,6\}}{\{0,1,2,3,4,6,8\}} = \frac{3}{7}$$

- Data set of sets $\{A_1, A_2, \ldots, A_n\}$ ($n = 1$ million)

- $\underline{\text{Document}}$ $D: \xrightarrow[\text{kgrams}]{}$ $\underline{\text{set}}$ $A_i \xrightarrow[\text{min hashing}]{}$ $\underline{\text{Vector}}$ $V_i \in \mathbb{R}^K$ with the property that as K gets larger $JS(A_i, A_j) \approx JS(V_i, V_j)$ gets closer.

Matrix/Vector Set Representation: (exact, but not space efficient)

- Represent set A_i as bit vector $b_i \in \{0,1\}^n$

Ex. $A_1 = \{1, 2, 5\}$, $n = 6$
$A_2 = \{3\}$, $A_3 = \{2, 4, 3, 6\}$, $A_4 = \{1, 4, 6\}$

	b_1	b_2	b_3	b_4
1	1	0	0	1
2	1	0	1	0
3	0	1	1	0
4	0	0	1	1
5	1	0	0	0
6	0	0	1	1

Thus, we've transformed our set of sets into a matrix

Min Hashing:

S1. Randomly re-order (permute) the rows (the bits).

S2. For each set/column, find the top/first 1 bit. ($b_1 = 1$, $b_2 = 3$, $b_3 = 2$, $b_4 = 1$)

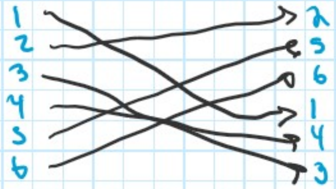
S3. Repeat steps 1,2 K times

$$V_i = \begin{bmatrix} m_1(A_i) \\ m_2(A_i) \\ \vdots \\ m_K(A_i) \end{bmatrix}$$

← each m_i is step 1 & 2 repeated

Ex.

Original Order   Random Order



$\Rightarrow$

b_1	b_2	b_3	b_4
1	0	1	0
1	0	0	0
0	0	1	1
1	0	0	1
0	0	1	1
0	1	1	0

$\Rightarrow m_j(b_i) = (2, 3, 2, 6)$

Then, $\hat{JS}(i, i') = 1$ if $m(i) = m(i')$

Take-away:

$$E[\hat{JS}(i, i')] = JS(A_i, A_{i'})$$

Why Does This Work??

$$P[m(i) = m(i')] = E[\hat{JS}(i, i')] = JS(A_i, A_{i'})$$

T_x = x rows with a 1 in both columns
T_y = y rows with a 1 in exactly 1 column
T_z = z rows with 0 in both columns

	b_1	b_2
T_y	1	0
T_y	0	1
T_x	1	1
T_z	0	0
T_y	0	1

$$JS(A_i, A_{i'}) = \frac{x}{x+y} = P[m(i) = m(i')] = \frac{1}{1+3}$$

Then thinking about only rows,
$P[m(i) = m(i')] = 1$ iff
top row type x
ignoring z.

How Big Should K be??

Chernoff - Hoeffding Bound

- iid RV's $x_1, \dots, x_K$ where $E[x_i] = \mu$

$$M = \frac{1}{K}\sum_{i=1}^{K} x_i \quad \text{and} \quad E[M] = E[x_i]$$

$x_i \in \{0, 1\}$

$$P[|M - E[M]| > \varepsilon] < 2 \cdot \exp\{-2\varepsilon^2 K\}$$

error tolerance 0.05

δ = probability of failure

- we can use algebra to solve for K given δ

$-2\varepsilon^2 K$

we can use algebra to solve for k given δ

↳ $\delta = 0.1 = 2e^{-2\varepsilon^2 k}$

$\Rightarrow 0.05 = e^{-2\varepsilon^2 k}$

$\Rightarrow \ln[0.05] = -2\varepsilon^2 k$

$$\boxed{\therefore k = \frac{\ln[0.05]}{-2\varepsilon^2}}$$

Fast Min Hash Signatures:

- Set A_i to vector $V_i \in \mathbb{Z}^k$
  - k hash functions $h_j : [n] \rightarrow [n']$

don't need to know

instead

$h: \Sigma^3 \rightarrow [n']$

alphabet of chars

Algo

one pass over data

for $x \in A_i$ do
    for $j = 1$ to k
        if $(h_j(x) < V_i(j))$
            $V_j \leftarrow h_j(x)$

$$JS_k(V_i, V_{i'}) = \frac{1}{k} \sum_{j=1}^{k} \begin{cases} 1 & \text{if } V_i(j) = V_{i'}(j) \\ 0 & \text{otherwise} \end{cases}$$