Min Hashing

Note Title 1/25/2016

K-grams

Documents

↓

Sets

$\{1,2,7\} \subset \{1,2,\ldots 100\}$

Min Hashing

today

Sets → Vectors

$JS(S_1, S_2)$

$\hat{JS}(\tilde{v}_1, v_2)$

LSH

- generalize

X = large # sets

+ $g \in$ every if $g \in X$

+ are any two $g, g' \in X$ same

Set → Vector $v = (v_1, v_2, \ldots, v_k)$ iid

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{2,3\}|}{|\{1,2,3,4\}|} \quad A = \{1,2,3\}, \; B = \{2,3,4\}$$

$$= \frac{1}{4}$$

Matrix

	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	0	0
6	0	0	1	1

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

$S_2 = \{3\}$

$S_3 = \{2, 3, 4, 6\}$

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

← very sparse

Step 1 Permute Rows

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

Step 2

$m(S_1) = 2$

$m(S_2) = 3$

$m(S_3) = 2$

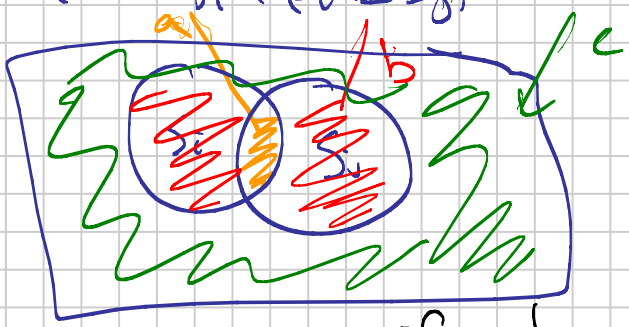
$m(S_4) = 6$

$$\hat{J}S_1 = (S_i, S_j) = \begin{cases} 1 & \text{if } m(S_i) = m(S_j) \\ 0 & \text{otherwise} \end{cases}$$

$$E\left[\hat{J}S_1(S_i, S_j)\right] = JS(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j| = |S_i \cap S_j| + |S_i \Delta S_j|}$$

$X = S_0 \cup S_1$

$= |S_0 \cap S_1| + |S_0 \Delta S_1|$



$$Pr[m_i = m_j] = \frac{a}{a+b}$$

first non-zero

$$\hat{JS}_k(A,B) = \frac{1}{k}\sum_{i=1}^{k} \begin{cases} 1 & \text{if } m_i(A) = m_i(B) \\ 0 & \text{otherwise} \end{cases}$$

k ↑ # hash fcn

$\varepsilon \in (0,1)$ $\quad 0.1$

$$Pr\left[\left|JS(A,B) - \hat{JS}_k(A,B)\right| \geq \varepsilon\right] \leq 2\exp\left(\frac{-2\varepsilon^2}{k \cdot 1^2}\right)$$

$$k \approx \frac{1}{2\varepsilon^2} \Rightarrow 200 \Rightarrow 1000 = k$$

Fast Min-Hashing

$\{0,1\} \leftarrow$ rand-bit()

$[0,1] \leftarrow$ rand-Unif()

$h : [n] \to [100\, n \log n]$

for $x \in S$

  for $j = 1$ to k (9) ← all entrees in vector

    if $(h_j(x) < v_j)$ (y_j)

      $v_j \leftarrow h_j(x)$ (y_j)

      $v_j \leftarrow x$

$(v_1 = \infty, v_2 = \infty, \ldots, v_k = \infty)$

$S = \{a, b, c\}$

$\{a, b, a, c\}$

	h_1	h_2
a	7	2
b	3	10
c	22	1

v (∞, ∞)

(7, 2) a

(3, 2) b

(3, 1) c

Runtime $O(|S| \cdot k)$

Document

n words

g - word-grams

k - hash fxns

$\binom{n}{g} \cdot k \approx n^g \cdot k$

$n \cdot g \cdot k$ too big

$(n/g) \cdot k$ too small

$(n-g) \cdot k$

a b c d e f g h i j k

$(n-g+1)$

$O(n \cdot k)$