IS622 Week 4 HW

Brian Chu | Sept 20, 2015

Exercise 3.1.3: Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T?

$\text{SIM}(S,T) = \frac{|S \cap T|}{|S \cup T|}$

Since subsets S and T each have m elements, $S \cup T = 2m - (S \cap T)$

To simplify this, I will assume S has already been selected. Therefore the probability of each m element also being in T is $\frac{m}{n}$.

If $E[X] = xp(x)$, then $E[S \cap T] = \sum_{i=1}^{m} \frac{m}{n}$
$= m * \frac{m}{n} = \frac{m^2}{n}$

$\text{SIM}(S,T) = (\frac{m^2}{n})/(2m - \frac{m^2}{n})$
$= (\frac{m^2}{n}) * \frac{1}{(2m - \frac{m^2}{n})}$
$= \frac{m^2}{2mn - m^2}$
$= \frac{1}{\frac{2n}{m} - 1}$
$= \frac{m}{2n-m}$

Exercise 3.3.3 : In Fig. 3.5 is a matrix with six rows.

(a) Compute the minhash signature for each column if we use the following three hash functions: h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x)=5x+2 mod6.

Compute hash functions

$$\begin{array}{cc} & \begin{matrix} S_1 & S_2 & S_3 & S_4 & 2x+1mod6 & 3x+2mod6 & 5x+2mod6 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 2 & 2 \\ 0 & 1 & 0 & 0 & 3 & 5 & 1 \\ 1 & 0 & 0 & 1 & 5 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 & 5 & 5 \\ 0 & 0 & 1 & 1 & 3 & 2 & 4 \\ 1 & 0 & 0 & 0 & 5 & 5 & 3 \end{bmatrix} \end{array}$$

Row 0

$$\begin{array}{cc} & \begin{matrix} S_1 & S_2 & S_3 & S_4 \end{matrix} \\ \begin{matrix} h_1 \\ h_2 \\ h_3 \end{matrix} & \begin{bmatrix} \infty & 1 & \infty & 1 \\ \infty & 2 & \infty & 2 \\ \infty & 2 & \infty & 2 \end{bmatrix} \end{array}$$

Row 1

$$\begin{array}{cc} & \begin{matrix} S_1 & S_2 & S_3 & S_4 \end{matrix} \\ \begin{matrix} h_1 \\ h_2 \\ h_3 \end{matrix} & \begin{bmatrix} \infty & 1 & \infty & 1 \\ \infty & 2 & \infty & 2 \\ \infty & 1 & \infty & 2 \end{bmatrix} \end{array}$$

Row 2

$$\begin{array}{cc} & \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{array}{c} h_1 \\ h_2 \\ h_3 \end{array} & \begin{bmatrix} 5 & 1 & \infty & 1 \\ 2 & 2 & \infty & 2 \\ 0 & 1 & \infty & 0 \end{bmatrix} \end{array}$$

Row 3

$$\begin{array}{cc} & \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{array}{c} h_1 \\ h_2 \\ h_3 \end{array} & \begin{bmatrix} 5 & 1 & 1 & 1 \\ 2 & 2 & 5 & 2 \\ 0 & 1 & 5 & 0 \end{bmatrix} \end{array}$$

Row 4

$$\begin{array}{cc} & \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{array}{c} h_1 \\ h_2 \\ h_3 \end{array} & \begin{bmatrix} 5 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 0 & 1 & 4 & 0 \end{bmatrix} \end{array}$$

Row 5 = final signature matrix

$$\begin{array}{cc} & \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{array}{c} h_1 \\ h_2 \\ h_3 \end{array} & \begin{bmatrix} 5 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 0 & 1 & 4 & 0 \end{bmatrix} \end{array}$$

(b) Which of these hash functions are true permutations?

Function 3 is a true permutation because each input (0-5) is mapped to a different output. Functions 1 and 2 have repeated outputs for different input values.

(c) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

$SIM(S_1, S_2) = \frac{1}{3} = 0.33, true = \frac{0}{4} = 0$
$SIM(S_1, S_3) = \frac{1}{3} = 0.33, true = \frac{0}{4} = 0$
$SIM(S_1, S_4) = \frac{1}{3} = 0.33, true = \frac{1}{4} = 0.25$
$SIM(S_2, S_3) = \frac{2}{3} = 0.67, true = \frac{0}{4} = 0$
$SIM(S_2, S_4) = \frac{2}{3} = 0.67, true = \frac{1}{4} = 0.25$
$SIM(S_3, S_4) = \frac{2}{3} = 0.67, true = \frac{1}{4} = 0.25$

The estimated and true Jaccard similarities were not too similar although this is only a very small matrix representation.

Exercise 3.5.5: Compute the cosines of the angles between each of the following pairs of vectors.

(a) (3,-1,2) and (-2,3,1)
$x \cdot y = -7$
$L_2 norms = \sqrt{14}, \sqrt{14}$
cosine angle $= \frac{-7}{\sqrt{(14)}\sqrt{(14)}} = -0.5$
cosine distance = acos(-0.5) * 180/pi = 120

(b) (1,2,3) and (2,4,6)
$x \cdot y = 28$
$L_2 norms = \sqrt{14}, \sqrt{56}$
cosine angle $= \frac{28}{\sqrt{(14)}\sqrt{(56)}} = 1$
cosine distance = 0

(c) (5,0,-4) and (-1,-6,2)

$x \cdot y = -13$
$L_2 norms = \sqrt{4}1, \sqrt{4}1$
cosine angle $= \frac{-13}{\sqrt{(41)}\sqrt{(41)}} = -0.317$
cosine distance $= 108.5$

(d) (0,1,1,0,1,1) and (0,0,1,0,0,0)

$x \cdot y = 1$
$L_2 norms = \sqrt{4}, \sqrt{1}$
cosine angle $= \frac{1}{\sqrt{(4)}\sqrt{(1)}} = 0.5$
cosine distance $= 60$

Exercise 3.7.1 : Suppose we construct the basic family of six locality-sensitive functions for vectors of length six. For each pair of the vectors 000000, 110011, 010101, and 011100, which of the six functions makes them candidates?

000000 and 110011
000000 and 010101
000000 and 011100
110011 and 010101
110011 and 011100
010101 and 011100

I didn't quite understand the objective and approach for this question.