Exercise 3.1.3

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?

Step-by-step explanation

Jaccard similarity between two sets is defined to be the ratio of the cardinality of the intersection to the union, that is

$$\mathrm{SIM}(S,T) = \frac{|S \cap T|}{|S \cup T|}. \qquad (1)$$

I will make the following assumptions from my interpretation of the statement. By choosing subsets at "random", the authors refer to uniform distribution among the possible ${}^{n}C_{m}$ subsets of size and chosen independently. Without loss of generality, we may assume U = 1, 2, 3....., n and m<n is fixed. Furthermore, we may assume that S = {1, 2..., m}

What we are to compute then is E[SIM(S, T)] for the exercise.

Subsets of size 1

For singletons, the similarity only has two possibilities: either they are the same set in which case the value is one, or not and the value is zero. Two randomly chosen items from a set of n possibilities are equal with probability (1/n). So we have

$$\mathrm{SIM}(S,T) = \begin{cases} 1 & \text{w.p. (with probability) } \frac{1}{n} \\ 0 & \text{w.p. } \frac{n-1}{n}. \end{cases}$$

Taking the expected value we get the answer is (1/n)

Subsets of size 2

As we noticed in the previous section, if T=S, then the similarity will be one. If they are disjoint, then it will be zero. In this case, they may have just one shared element. There are ${}^{n}C_{2}$ possible subsets for T. Among them, only one has similarity one. How many subsets share just one element? Either T can be of the form {1, k}, k ≠ 2 or {2,k'}, k' ≠ 1. There are 2*(n-2) of these. The similarity for these sets is 1/3 because they share one element in common, and the size of their union is four minus the size of the intersection. So the Jaccard similarity has probability mass function

$$\mathrm{SIM}(S,T) = \begin{cases} 1 & \text{w.p. } \frac{1}{\binom{n}{2}} \\ \frac{1}{3} & \text{w.p. } 2 \cdot \frac{(n-2)}{\binom{n}{2}} \\ 0 & \text{w.p. } 1 - \frac{1}{\binom{n}{2}} - 2 \cdot \frac{(n-2)}{\binom{n}{2}}. \end{cases}$$

Subsets of size 3

Exercise left to the reader. The answer is

$$\mathrm{SIM}(S,T) = \begin{cases} 1 & \text{w.p. } \frac{1}{\binom{n}{3}} \\ \frac{1}{2} & \text{w.p. } \frac{3(n-3)}{\binom{n}{3}} \\ \frac{1}{5} & \text{w.p. } \frac{1}{\binom{n}{3}} \cdot 3 \cdot \binom{n-3}{2} \\ 0 & \text{w.p. } 1 - \frac{1}{\binom{n}{3}} - \frac{3(n-3)}{\binom{n}{3}} - \frac{1}{\binom{n}{3}} \cdot 3 \cdot \binom{n-3}{2}. \end{cases}$$

Solution

Working through the previous subsection really drove the point home for me. One idea is to iterate over how many elements they **don't** have in common. We'll index by sizes j = 0,1,2...,m where (j=0) corresponds to identical subsets and (j=m) corresponds to disjoint subsets. If the intersection is the size (m-j), there are ${}^{m}C_{m-j}$ ways to pick (m-j) elements from subset S to be in the intersection. If T is one such subset, then the Jaccard similarity is (m-j)/(m+j). So we have to count how many such T there are. For the remaining j elements, they must be chosen from the remaining (n-m) part of U that is in (U/S). There are ${}^{n-m}C_{j}$ such choices. Thus we arrive at the following probability mass function for Jaccard similarity

$$\mathrm{SIM}(S,T) = \begin{cases} 1 & \text{w.p. } \frac{1}{\binom{n}{m}} \\ \frac{m-1}{m+1} & \text{w.p. } \frac{1}{\binom{n}{m}} \binom{m}{m-1} \binom{n-m}{1} \\ \frac{m-2}{m+2} & \text{w.p. } \frac{1}{\binom{n}{m}} \binom{m}{m-2} \binom{n-m}{2} \\ \vdots & \\ \frac{m-(m-1)}{m+(m-1)} & \text{w.p. } \frac{1}{\binom{n}{m}} \binom{m}{m-(m-1)} \binom{n-m}{m-1} \\ 0 & \text{w.p. } 1 - \text{sum of above :P} \end{cases} \tag{2}$$

Taking expected value we get the expectation value for two independently with uniformly randomly chosen subsets of size m of n objects is

$$\mathbb{E}[\mathrm{SIM}(S,T)] = \frac{1}{\binom{n}{m}} \sum_{j=0}^{m-1} \frac{m-j}{m+j} \binom{m}{m-j} \binom{n-m}{j}. \qquad (3)$$

Exercise 3.5.3

Exercise 3.5.3: Prove that if i and j are any positive integers, and i < j, then the Li norm between any two points is greater than the Lj norm between those same two points.

Exercise 5.1.4

Construct, for any integer n, a Web such that, depending on β, any of the n nodes can have the highest PageRank among those n. It is allowed for there to be other nodes in the Web besides these n.

Exercise 4.3.3:

As a function of n, the number of bits and m the number of members in the set S, what number of hash functions minimizes the false- positive rate?

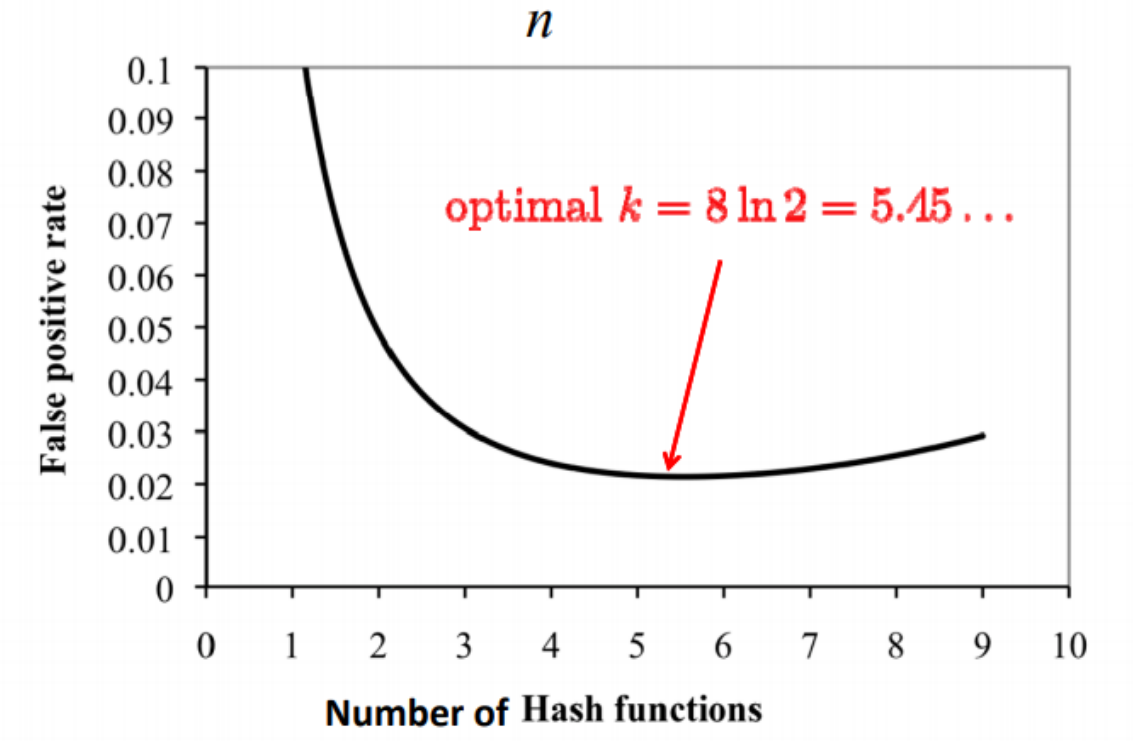
> **Exercise 4.3.3:** As a function of n, the number of bits and m the number of members in the set S, what number of hash functions minimizes the false-positive rate?

Had downvote on solution:

After all members of S have been hashed to a Bloom filter, the probability that a specific bit is still 0 is

$$p' = (1 - \frac{1}{m})^{kn} \simeq e^{-\frac{kn}{m}} = p$$

Number of bits per member $\frac{m}{n} = 8$



Optimal number of hash functions

Using $k_{opt} = \frac{m}{n} ln(2)$ the false positive rate is

$$1 - p^{\frac{m}{n} ln(2)} = 0.5^{\frac{m}{n} ln(2)} \simeq (0.6185)^{\frac{m}{n}}, where ln(2) = 0.6931$$

In practice, k should be an integer. May choose an integer value smaller than k_{opt} to reduce hashing over head

m/n denotes
bits per entry False positive rate
m/n =6 k=4 p_{error} = 0.0561

m/n =8 k=6 p_{error} = 0.0215

m/n =12 k=8 p_{error} = 0.00314

m/n =16 k=11 p_{error} = 0.000458

OR:

2. To find the k that minimizes the false positive rate we must differentiate g with respect to k and set the outcome to 0.
Let $g = (1 - e^{-\frac{km}{n}})^k$. g is continuous and can be differentiated to find the minima.

L3S Research Center
Large Scale Data Mining, SS 2016
Dr. Avishek Anand
Solution to Assignment 2, due: **28 April 2016**

$$\frac{dg}{dk} = \frac{d}{dk}\{1 - e^{-\frac{km}{n}}\}^k = 0 \quad (1)$$

Using the chain rule we find that $k = \frac{n}{m}ln(2)$

Exercise 6.4.3 :

Suppose item i appears exactly s times in a file of n baskets, where s is the support threshold. If we take a sample of n/100 baskets, and lower the support threshold for the sample to s/100, what is the probability that i will be found to be frequent? You may assume that both s and n are divisible by 100.

Exercise 9.3.1:

Figure 9.8 is a utility matrix, representing the ratings, on a 1–5 star scale, of eight items, a through h, by three users A, B, and C. Compute the following from the data of this matrix. (a) Treating the utility matrix as boolean, compute the Jaccard distance between each pair of users.

(b) Repeat Part (a), but use the cosine

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Figure 9.8: A utility matrix for exercises

distance.

(c) Treat ratings of 3, 4, and 5 as 1 and 1, 2, and blank as 0. Compute the Jaccard distance between each pair of users.

(d) Repeat Part (c), but use the cosine distance.

(e) Normalize the matrix by subtracting from each nonblank entry the average value for its user.

(f) Using the normalized matrix from Part (e), compute the cosine distance between each pair of users

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	1	0	1	1	1

(a) Jaccard distance is used for measuring dissimilarity which is (1 - Jaccard Coefficient)

$\Rightarrow$ **Jaccard distance = 1 - Jaccard Coefficient**

Jaccard Coefficient between any 2 variables can be calculated as -

$$\text{Coefficient} = \frac{|A \cap B|}{|A \cup B|}$$

A = 11010010
B = 01110000
C = 00010111

So. Distance between A & B

$A \cap B = 01010000 \Rightarrow |A \cap B| = 2$
$A \cup B = 11110010 \Rightarrow |A \cup B| = 5$

$$d_{A,B} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

$$= 1 - \frac{2}{5} = \frac{3}{5} = \boxed{0.60}$$

distance between A & C

$A \cap C = 00010010 \Rightarrow |A \cap C| = 2$
$A \cup C = 11010111 \Rightarrow |A \cup C| = 6$

$$d_{A,C} = 1 - \frac{|A \cap C|}{|A \cup C|}$$

$$= 1 - \frac{2}{6} = \frac{4}{6} \Rightarrow \boxed{d_{A,C} = 0.66}$$

distance between B & C

$B \cap C = 00010000 \Rightarrow |B \cap C| = 1$
$B \cup C = 01110111 \Rightarrow |B \cup C| = 6$

$$d_{B,C} = 1 - \frac{|B \cap C|}{|B \cup C|}$$

$$= 1 - \frac{1}{6} = \frac{5}{6} \Rightarrow \boxed{d_{B,C} = 0.833}$$

(b) Cosine distance between 2 vectors $\{a,b,c..\}$ and $\{x,y,z,..\}$

$$= 1 - \left[\frac{a.x + b.y + c.z + \dots}{\sqrt{a^2+b^2+c^2+\dots} \cdot \sqrt{x^2+y^2+z^2+\dots}} \right]$$

So for given vectors

A = 11010010
B = 01110000
C = 00010110

} Considering 3,4,5 as 1 and 1,2 & blanks as 0

Distance between A & B

$$= 1 - \left[\frac{1.0+1.1+0.1+1.1+0.0+0.0+1.0+0.0}{\sqrt{1^2+1^2+1^2+1^2+0^2+0^2+0^2+0^2} \cdot \sqrt{0^2+1^2+1^2+1^2+0^2+0^2+0^2+0^2}} \right]$$

$$= 1 - \left[\frac{2}{\sqrt{4} \cdot \sqrt{3}} \right] = 1 - \frac{1}{\sqrt{3}} = \boxed{0.4226}$$

Distance between A & C

$$= 1 - \left[\frac{1.0+1.0+0.0+1.1+0.0+0.1+1.1+0.0}{\sqrt{1^2+1^2+1^2+1^2+0^2} \cdot \sqrt{1^2+1^2+1^2+0^2+\dots}} \right]$$

$$= 1 - \frac{2}{\sqrt{4} \cdot \sqrt{3}} = 1 - \frac{1}{\sqrt{3}} = \boxed{0.4226}$$

Between B & C

$$= 1 - \left[\frac{0.0+1.0+1.0+1.1+0.0+0.1+0.1+0.0}{\sqrt{1^2+1^2+1^2+0^2+\dots} \cdot \sqrt{1^2+1^2+1^2+0^2+\dots}} \right]$$

$$= 1 - \left[\frac{1}{\sqrt{3} \cdot \sqrt{3}} \right] = 1 - \frac{1}{3} = \boxed{0.667}$$

3. Average (A) $= \frac{4+5+5+1+3+2}{8} = \frac{20}{8} = 2.5$

Average (B) $= \frac{3+4+3+1+2+1}{8} = \frac{14}{8} = 1.75$

Average (C) $= \frac{2+1+3+4+5+3}{8} = \frac{18}{8} = 2.25$

	a	b	c	d	e	f	g	h
A	1.5	2.5		2.5	-1.5		0.5	-0.5
B		1.25	2.25	1.25	-0.75	0.25	-0.75	
C	-1.25		+1.25	0.75		1.75	2.75	0.75

d. Normalized Matrix

	a	b	c	d	e	f	g	h
A	1.5	2.5		2.5	-1.5		0.5	-0.5
B		1.25	2.25	1.25	-0.75	0.25	-0.75	
C	-0.25		-1.25	0.75		1.75	2.75	0.75

Cosine distance

A & B

$$Cd_{AB} = 1 - \left[\frac{(2.5)(1.25) + (2.5)(1.25) + (-1.5)(-0.75) + (0.5)(-0.75)}{\sqrt{(1.5)^2 + (2.5)^2 + (2.5)^2 + (-1.5)^2 + (0.5)^2 + (-0.5)^2} \cdot \sqrt{(1.25)^2 + (2.25)^2 + (1.25)^2 + (-0.75)^2 + (0.25)^2 + (-0.75)^2}} \right]$$

$$= 1 - \left[\frac{7}{\sqrt{17.5} \cdot \sqrt{9.375}} \right] = 1 - \frac{7}{(3.06)(4.18)}$$

$$= 1 - 0.547 = \boxed{0.452} \checkmark$$

A & C

$$Cd_{AC} = 1 - \left[\frac{(1.5)(-0.25) + (2.5)(0.75) + (0.5)(2.75) + (-0.5)(0.75)}{\sqrt{(1.5)^2 + (2.5)^2 + (1.5)^2 + 2(0.5)^2} \cdot \sqrt{(-0.25)^2 + (-1.25)^2 + (0.75)^2 + (1.75)^2 + (2.75)^2 + (0.75)^2}} \right]$$

$$= 1 - \left[\frac{2.5}{\sqrt{17.5} \cdot \sqrt{13.375}} \right] = 1 - 0.163 = \boxed{0.836} \checkmark$$

B & C

$$Cd_{BC} = 1 - \left[\frac{(2.25)(-1.25) + (1.25)(0.75) + (0.25)(1.75) + (-0.75)(2.75)}{\sqrt{2(1.25)^2 + (2.25)^2 + 2(-0.75)^2 + (0.25)^2} \cdot \sqrt{2(0.75)^2 + (0.25)^2 + (-1.25)^2 + (1.75)^2 + (2.75)^2}} \right]$$

$$= 1 - \left[\frac{-3.5}{\sqrt{9.375} \cdot \sqrt{13.375}} \right] = 1 + 0.312$$

$$= \boxed{1.312} \checkmark$$