Dorina Ismaili & Deniz Gaiser
16.12.2024

Mining Massive Datasets
Homework 7

Exercise 1

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 ? Explain your formula.
The Jaccard similarity of two sets S and T where $|S| = |T| = m$ is given as

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}.$$

The probability that $x \in S$ is equal to $\frac{m}{n}$. The probability that $P(x \in S \cap T) = \frac{m}{n}\frac{m}{n} = \frac{m^2}{n^2}$. Thus the expected size of the intersection is $\mathbb{E}(|S \cap T|) = n\frac{m^2}{n^2} = \frac{m^2}{n}$. The size of union is $|S \cup T| = |S| + |T| - |S \cap T| = 2m - |S \cap T|$. Taking the expectation value it holds that $E|S \cup T| = 2m - \frac{m^2}{n}$. Therefore it holds that

$$\mathbb{E}J(S,T) = \frac{E(|S \cap T|)}{E(|S \cup T|)]}$$

. It holds that $\frac{\frac{m^2}{n}}{2m - \frac{m^2}{n}}$. So the expected value of the Jaccard similarity is

$$\mathbb{E}[J(S,T)] = \frac{m}{2n-m}.$$

Exercise 2

Prove that if the Jaccard similarity of two columns is 0 or 1, then Min-Hashing always gives a correct estimate of the Jaccard similarity.
The Jaccard similarity is given as

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}.$$

The Min-Hashing is a probabilistic algorithm for estimating Jaccard similarity

$$P(MinHash(S) = MinHash(T)) = J(S,T)$$

. In case when $J(S,T) = 1$, it implies $S = T$ as:

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

Using probability estimates it holds that MinHash(S) = MinHash(J) = 1.
For $J(S,T) = 0$, it implies $S \cap T =$, the Min-Hash values for S and T will always differ because there is no overlap between the two sets.
Min-Hashing always gives a correct estimate of the Jaccard similarity.

Exercise 3

Element	S_1	S_2
a	0	1
b	1	0
c	0	1
d	1	1
e	1	0

a) Compute the Jaccard similarity between S_1 and S_2.

To compute the Jaccard similarity we use

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}.$$

From the givens it holds that $|S_1 \cap S_2| = 1$, and $|S_1 \cup S_2| = 5$. Therefore the Jaccard similarity is $J(S_1, S_2) = \frac{1}{5} = 0.2$

b) The probability that two sets S_1 and S_1 hash to the same value under a Min-Hash permutation is exactly equal to their Jaccard similarity:

$$P(MinHash(S_1) = MinHash(S_2)) = J(S_1, S_2) = \frac{1}{5}.$$

This means that for a random permutation of the rows, the two columns will hash to the same value 20% of the time.
There are $5! = 120$ permutations of the rows. Since 20% of these permutations will result in S_1 and S_2 hashing to the same value, the fraction of permutations is: 0.2. The number of such permutations is: $0.2 \times 120 = 24$.

Exercise 4

Extend the implementation of the algorithm for MinHash signatures.

The Python code.

Exercise 5

The Python code.

Exercise 6

d) Conclusion:
From the results we obtain by running the ipynb file, we can conclude that the Jaccard Similarity is oftentimes greater than zero, if there are similarities in the elements of the sets. The signature similarity on the other hand was zero at all times.
This implies, that the hashing algorithm does not approximate the Jaccard similarity enough.

Exercise 7

Consider the S-curve function $1 - (1 - x^r)^b$ which gives the probability that a pair of sets (e.g. of shingles) with Jaccard similarity x becomes a candidate pair when using the LSH technique.

a) Plot the S-curve for all 4 pairs of parameters r = 2, 5 and b = 10, 50.

b) Derive an analytical expression (depending on r and b) for com- puting the x-value where the S-curve function features the highest ascend

Let $f(x) = 1 - (1 - x^r)^b$, the derivative is expressed as $f'(x) = brx^{r-1}(1 - x^r)^{b-1}$. Finding the maximum ascent we set $\frac{d}{dx} f'(x) = 0$, x is around $(1/b)^{1/r}$. For larger b, the curve becomes steeper, and the inflection point shifts to smaller x-values.