Week 4

James Quacinella

09/27/2015

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 ?

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: h 1 (x) = 2x + 1 mod 6; h 2 (x) = 3x + 2 mod 6; h 3 (x) = 5x + 2 mod 6.

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

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

Exercise 3.5.5 : Compute the cosines of the angles between each of the fol- lowing pairs of vectors. (a) (3, -1, 2) and (-2, 3, 1). (b) (1, 2, 3) and (2, 4, 6). (c) (5, 0, -4) and (-1, -6, 2). (d) (0, 1, 1, 0, 1, 1) and (0, 0, 1, 0, 0, 0).

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?

Question (Discussion) Exercise 3.5.1

On the space of nonnegative integers, which of the following functions are distance measures? If so, prove it; if not, prove that it fails to satisfy one or more of the axioms.

(a) max(x, y) = the larger of x and y.
(b) diff(x, y) = |x - y| (the absolute magnitude of the difference between x and y).
(c) sum(x, y) = x + y.

Answer

For a distance metric, we need to meet these conditions:

1. $d(x, y) >= 0$ (no negative distances)
2. $d(x, y) = 0$ if and only if $x = y$ (distances are positive, except for the distance from a point to itself).
3. $d(x, y) = d(y, x)$ (distance is symmetric).
4. $d(x, y) <= d(x, z) + d(z, y)$ (the triangle inequality)

Lets look at $diff(x, y) = |x - y|$. The first condition is obviously true from the definition of the absolute value operator. All elements in the range of $|z|$ are greater than 0, so condition one is satisfied.

The second condition is also true since the only time the $abs(z)$ function is 0 is when $z = 0$, in this case $z = x - y$. This infers that $x = y$ for this condition, and property two is satisfied.

The third condition is based on symmetry. To prove this, we need to prove that $d(x, y) = d(y, x)$, or $|x - y| = |y - x|$. We do know that the asolute value has the property that $|z| = | - z|$. Therefore, the left side of the proof, |x - y|, can be rewritten as:

$$
\begin{aligned}
&|x - y| \\
&= | - (x - y)| \\
&= | - x + y| \\
&= |y - x|
\end{aligned}
$$

The third condition is therefore satsified. The last condition states that:

$$d(x, y) <= d(x, z) + d(z, y)$$

Therefore, the following must hold:

$$|x - y| <= |x - z| + |z - y|$$