Summary of section 3.7:LSH Families for Other Distance Measures

Jixiang Hao

April 2020

1 1. introduction

there are many ways to evaluate the similarity between two vectors. this summary will introduce other distance such like Hamming distance, Cosin distance, and normal Euclidean distance, which have a ability to represent the locality-sensitive families.

2 LSH families for hamming distance

Hamming distance is way to indicate the similarity in two binary data strings Where two string have same length, the measuring method of hamming distance is to comparing how many different bit positions in two binary string. For example, if we want to compare 5 with 3 by using hamming distance, first we transfer 5 and 3 to binary data, 1010 and 1100 respectively then we obverse that there are 2 different position between 1010 and 1100. Thus, the hamming distance between 5 and 3 is 2. by the description of hamming distance, we find that if two binary data strings is completely same, the hamming distance is 0, so the hamming distance show the similarity of two data. Based on the finding, we will introduce vector similarity measured by hamming distance.
the vector is set of different numerical numbers, different number represent the different dimensions of vector. so we can randomly select positions for two vectors that we want to compared and measure the hamming distance for two vectors. we adopt a group of locality-sensitive functions F = f1, f2, f3,..fn. for each function fi, it can randomly selection a bit position of vector x, such as fi(x)=xi and fi(y)=yi. Thus, n hash function can select n pair of bit position of two vector. The probability that fi(x) = fi(y) for the randomly chosen i is exactly 1 - (the number of pair that x, y have same position) / x or y dimensions. the sensitive of group Locality-sensitive functions f1...fn is (d1, d2, 1-d1/d, 1-d2/d).There are a flaw for hamming distance LSH. Since the Jaccard distance runs in 0 to 1, we do the regularization by divided vector dimensions. When the measured vector have low dimensions, the result of similarity may not precision since no enough dimension are compared by locality-sensitive function.

3 Random Hyperplanes and the Cosine Distance

By the definition, the Hyperplane is subspace of ambient space, the dimension of Hyperplane is dimension of ambient space - 1. For example, if the ambient space is a 3d space, the Hyperplane of this ambient space is 2d plant, while if the space is 2d space, the Hyperplace is 1d line. Because of reduced dimension, the Hyperplane have infinite choice in a lower dimension. For example, in a 3d space can be split by infinite 2d plane. The figure below assume a 2d plane build by vector x, y.

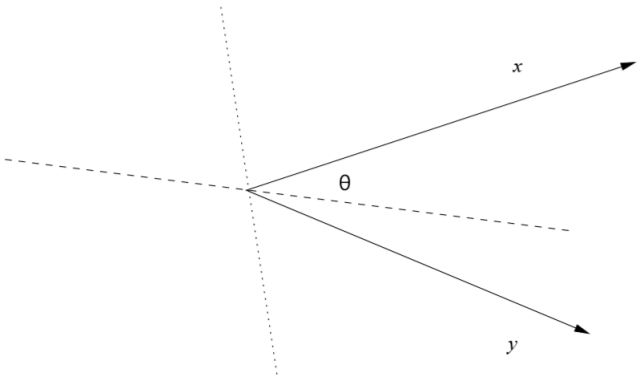

Figure 1: asclic 2d-Hyperplanes

The figure shows that the dash line and dot line represent normal vectors for each Hyperplane, which mean that the Hyperplane build by all points which dot product with v is 0. Thus, when we project vector x and y into Hyperplane with normal vector dashed line, the projections of x and v have different direction. In opposite, when we project vector x and y into Hyperplane with normal vector dotted line, the projections of x and y have same direction.
We hope to find the vector that normal to the Hyperplane with dashed normal vector rather than Hyperplane with dotted normal vector. Thus, we assumn a locality-sensitive family F = (f1,...fn) with n hash function f, each hash function fi can randomly select a vector vf , if and only if fi(x) and fi(y) have same sign, we think fi(x) = fi(y). Then F is a locality-sensitive family for the cosine distance.

4 Sketches

the Sketches is method that replace the random vectors to a Symbolized vectors. For example, assuming we have 3 random vector x, y, z and vectors v1, v2 that we want to know similarity between both, and all vector have same dimensions.

We replace all positive value in with +1 and replace all negative value with -1 in vector x, and do same method to y and z, this step is called Sketching vector. Then, we dot product v1 with other 3 Sketched vector3 and sketch it and do the same step to v2. Thus, we can calculate the cos similarity between sketched vectors, the math formula of computing vector cos value i:

$$cosin(v1_s, v2_s) = \frac{v1_s \cdot v2_s}{|v1_s| \cdot |v1_s|} \tag{1}$$