CS 6140: Data Mining Assignment 3

Distances and LSH

Yash Gangrade (u1143811), MS in Computing

30th January 2019

Contents

List of Figures

1 Choosing r,b

1.1 Estimating the best values of hash functions

Ans: Given a budget of $t = 160$ hash functions and threshold value of $\tau = 0.85$ and then using the trick mentioned in the class to estimate the value of b, we get,

$$\tau = \left(\frac{b}{t}\right)^{1/b}$$

$$b \approx -\log_\tau t$$

$$b \approx -\log_0 .85160$$

$$b = 31.2283 \approx 32$$

Therefore, we have, number of rows of bands i.e. $r = \frac{t}{b} = \frac{160}{32} = 5$.

Now, for testing the pair of (r,b), we perform experiments by trying multiple close values of b and r and calculating the similarity on those values. The plot for S-curve is shown for each pair of (r,b) chosen. To get the best pair, we draw a line $s = 0.85$ and see which plot is the most steep at the point of intersection. After all these experimental pairs, we found that the $(r, b) = (\mathbf{10}, \mathbf{16})$ is the best of all. Please refer to the figure below for explanation.

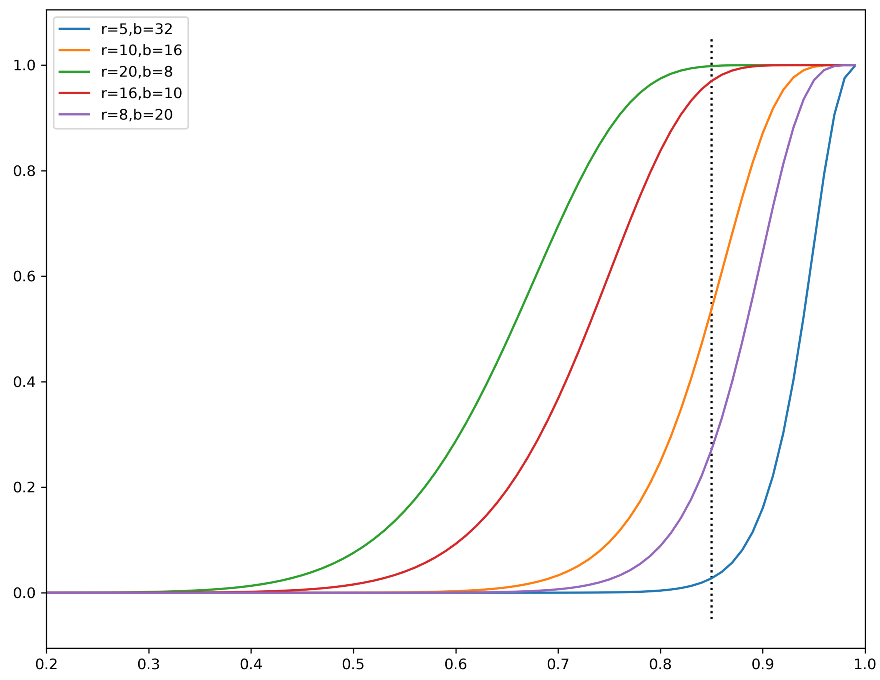

Figure 1: S-curve for different r,b values

1.2 Probability of each pair of the four objects for being estimated to having similarity greater than tau = 0.85

Ans: The probabilities are shown as follows (here $e\text{-}<x>$ refers to 10^{-x}):

Pair	**Probability**
(A, B)	0.0958
(A, C)	2.328e-09
(A, D)	5.071e-07
(B, C)	1.11e-15
(B, D)	2.827e-05
(C, D)	0.953

2 Generating Random Directions

2.1 Procedure

Ans: The algorithm for generating random directions is as follows:

Algorithm 1 *generateRandomDirection*(d) Generate Random Directions

function generateRandomDirection(d):
    for i in $\{1, \cdots, d\}$ do:
        $U_1 \leftarrow U(0,1)$

        $U_2 \leftarrow U(0,1)$

        $r_i \leftarrow \sqrt{-2\ln U1} \cdot cos(2\pi U_2)$

    return r/||r||^2

    end for loop

The algorithm stated above can be used to generate random directions. Here, d is the dimension of the unit vector U(0, 1) which will be used for generating the random directions. U(0, 1) generates a uniform random variable between 0 and 1. To get different entries filled up in the r vector, we use the box muller transform to get an independent random variable. After that, r is normalized to unit normal and then it's returned function.

2.2 CDF Plot of pairwise dot products

Ans: The plot of the CDF of dot products is shown below.

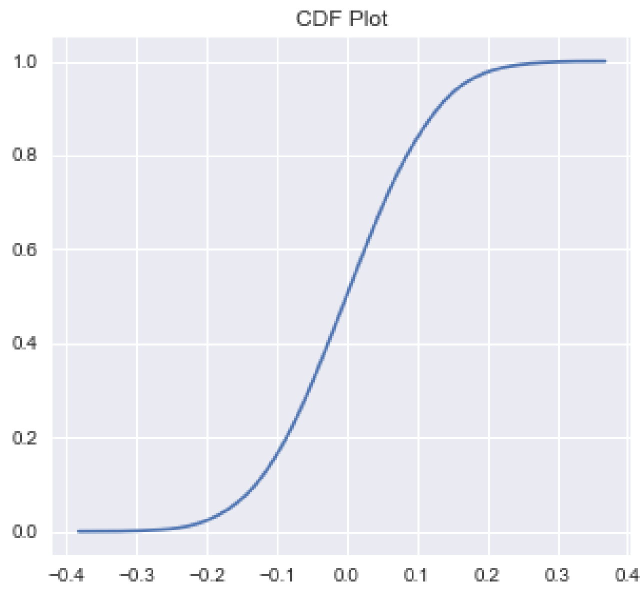

Figure 2: CDF Plot (part 2b)

3 Angular Hashed Approximation

3.1 CDF Plot of all pairs of dot products

Ans: The plot of the CDF of dot products is shown below. Also, the number of pairs with angular similarity greater than $\tau > 0.85$ is approximately 67299.

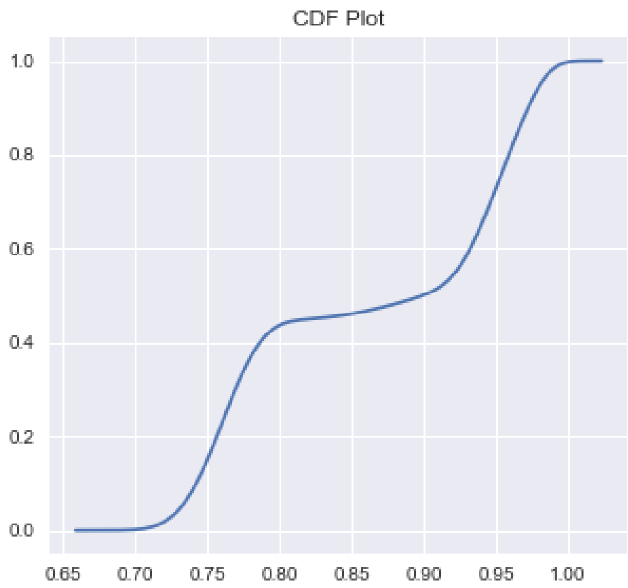

Figure 3: CDF Plot (part 3a)

3.2 CDF Plot of dot products of t random unit vectors

Ans: The plot of the CDF of dot products is shown below. Also, the number of pairs with angular similarity greater than $\tau > 0.85$ is nearly 0. Since the distribution is centered at 0.5 and it follows a Gaussian model, it's understandable that given the current variance almost all of the distribution vanishes before 0.85.

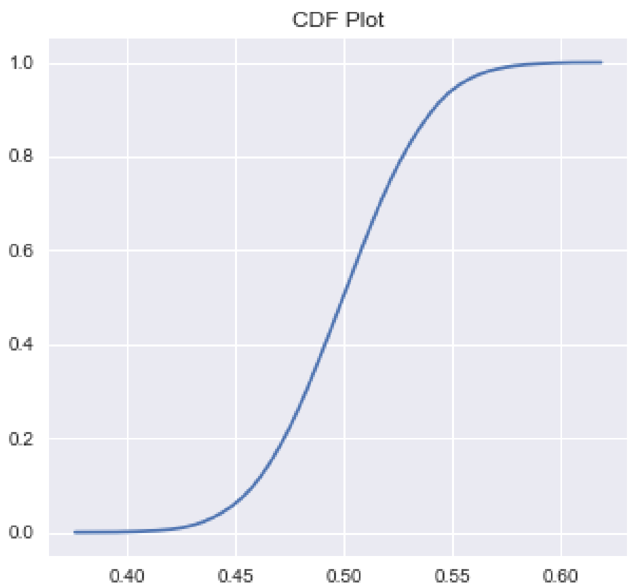

Figure 4: CDF Plot (part 3b)