Asmt 3: Distances and LSH

Turn in through **Gradescope** by Wednesday, February 8 at 2:45pm, then come to class:
100 points

Overview

In this assignment, you will explore LSH and Euclidean distances. You will use a data set for this assignment available on Canvas.

Note: Homework assignments are intended to help you learn the course material, and successfully solve mid-term and final exams that will be done on paper in person.

As usual, it is recommended that you use LaTeX for this assignment. If you do not, you may lose points if your assignment is difficult to read or hard to follow. The latex source for this homework is available on Canvas. I recommend that you drop the two latex files (.tex and .sty) in a new Overleaf project, and compile/edit the .tex file there.

1 Choosing r, b (35 points)

Consider computing an LSH using $t = 200$ hash functions. We want to find all object pairs which have Jaccard similarity above $\tau = .75$.

A: (15 points) Use the trick mentioned in class and the notes to estimate the best values of hash functions b within each of r bands to provide the S-curve

$$f(s) = 1 - (1 - s^b)^r$$

with good separation at τ. Report these values b, r.

Answer: we will use $b \approx -\log \tau(t)$
plugging in the numbers: $b \approx -\log .75(200)$
we get b $\approx$ 18.417266399
so b = 20 and r = 10

B: (20 points) Consider the 4 objects A, B, C, D, with the following pair-wise similarities:

	A	B	C	D
A	1	0.77	0.25	0.33
B	0.77	1	0.20	0.55
C	0.25	0.20	1	0.91
D	0.33	0.55	0.91	1

Use your choice of r and b and $f(\cdot)$ designed to find pairs of objects with similarity greater than τ: what is the probability, for each pair of the four objects, of being estimated as similar (i.e., similarity greater than $\tau = 0.75$)? Report 6 numbers. *(Show your work.)*

Answer:
I used the following formula

$$f(s) = 1 - (1 - s^b)^r$$

for each of the pairs I plugged in the value in the table for s, I used 20 for b and 10 for r

$ab \approx .782$
$ac \approx 1.9 * 10^{-5}$
$ad \approx 3 * 10^{-4}$
$bc \approx 2 * 10^{-6}$
$bd \approx 4.9 * 10^{-2}$
$cd \approx .9999$

2 Generating Random Directions (30 points)

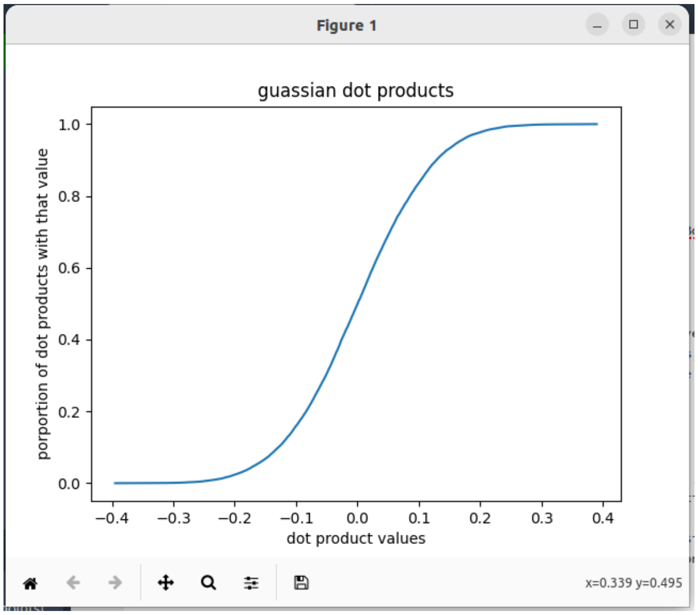
Read Section 5.3.1 in these notes `https://www.cs.utah.edu/~jeffp/DMBook/L5-LSH.pdf` and the part **"Generating random unit vectors"** in Section 4.6.4 in M4D (`https://mathfordata.github.io/versions/M4D-v0.6.pdf`).

A: (10 points) Describe how to generate a single random unit vector (chosen uniformly over from the space of all unit vectors $\mathbb{S}^{d-1}$) in $d = 10$ dimensions. To generate randomness, use only the operation $u \leftarrow \mathsf{unif}(0, 1)$, which generates a uniform random variable between 0 and 1 (then other linear algebraic and trigonometric, etc operations are allowed). *(This random uniform value can be called multiple times.)*

Answer:
we would generate 5 random scalar quantities using unif(0,1) [u1, u2, ..., u5]
we would use each ui from [u1, u2, ..., u5] in the box muller transformation to get two Gaussian random variables, resulting in 10 total Gaussian random variables
Once we have all 10 of the Gaussian random variables that we need, we assign coordinates in our space to those random variables and BOOM we have a randomized point in 10-dimensional space.

B: (20 points) Generate $t = 200$ unit vectors in $\mathbb{R}^d$ for $d = 100$. Plot of cdf of their pairwise dot products (yes, you need to calculate $\binom{t}{2}$ dot products).

Answer:



3 Angular Hashed Approximation (35 points)

Read Section 4.6.4 in M4D (`https://mathfordata.github.io/versions/M4D-v0.6.pdf`).

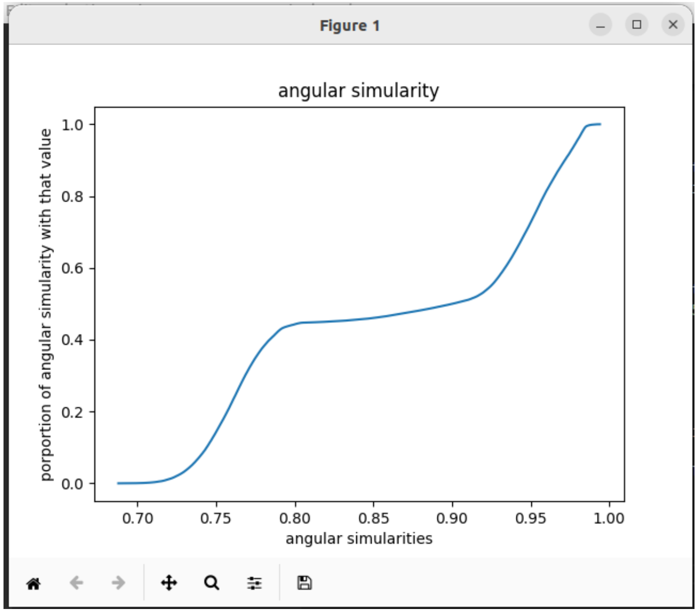
Consider the $n = 500$ data points in $\mathbb{R}^d$ for $d = 100$ in data set R, given in Canvas. We will use the angular similarity, between two vectors $a, b \in \mathbb{R}^d$:

$$\mathsf{s}_{\text{ang}}(a,b) = 1 - \frac{1}{\pi}\arccos(\langle \bar{a}, \bar{b} \rangle)$$

If a, b are not unit vectors (e.g., in $\mathbb{S}^{d-1}$), then we convert them to $\bar{a} = a/\|a\|_2$ and $\bar{b} = b/\|b\|_2$. The definition of $\mathsf{s}_{\text{ang}}(a,b)$ assumes that the input are unit vectors, and it reports a value between 0 and 1, with as usual 1 meaning most similar.

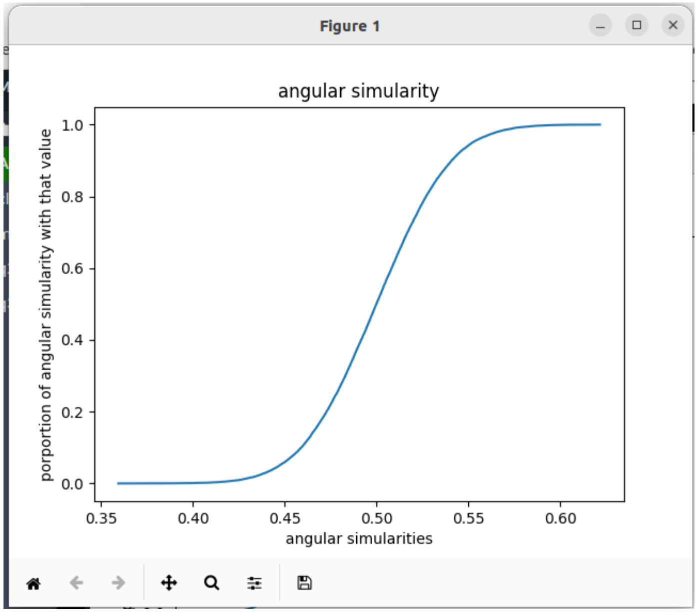
A: (15 points) Compute all pairs of dot products *(Yes, compute $\binom{n}{2}$ values)*, and plot a cdf of their angular similarities. Report the number with angular similarity more than $\tau = 0.75$.

Answer: 107004



B: (20 points) Now compute the angular similarities among $\binom{t}{2}$ pairs of the t random unit vectors from Q2.B. Again plot the cdf, and report the number with angular similarity above $\tau = 0.75$.

Answer: 0



4 Bonus (3 points)

Implement the banding scheme with your choice of r, b, using your $t = 200$ random vectors, to estimate the pairs with similarity above $\tau = 0.85$ in the data set R. Report the fraction found above $\tau = 0.75$. Compare the runtime of this approach versus a brute force search.

Answer: