Asmt 3: Distances and LSH

Turn in through Canvas by 2:45pm, then come to class:
Wednesday, February 24th
100 points

Student Name: - - - - - - -
Student UID: - - - - - - -

Overview

In this assignment you will explore LSH and Euclidean distances. You will use a data set for this assignment:

Canvas –> Files –> Assignments –> LSH –> R.csv

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. Find a sample form in: Canvas –> Files –> Assignments –> Assignment_Latex_Template.zip.

1 Choosing r, b (35 points)

Consider computing an LSH using $t = 160$ hash functions. We want to find all object pairs which have Jaccard similarity above $\tau = 0.65$.

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.

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

	A	B	C	D	E
A	1	0.72	0.35	0.15	0.55
B	0.72	1	0.42	0.85	0.44
C	0.35	0.42	1	0.25	0.50
D	0.15	0.85	0.25	1	0.66
E	0.55	0.44	0.50	0.66	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 $\tau = 0.65$)? Report 10 numbers. *(Show your work.)*

2 Generating Random Directions (30 points)

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 $d = 12$ 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.)*

B: (20 points) Generate $t = 200$ unit vectors in $\mathbb{R}^d$ for $d = 120$. Plot cdf of their pairwise dot products (yes, you need to calculate $\binom{t}{2}$ dot products).

3 Angular Hashed Approximation (35 points)

Consider the $n = 450$ data points in $\mathbb{R}^d$ for $d = 100$ in data set R, given at the top. 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 $\tau = 0.8$.

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 $\tau = 0.75$.

4 Bonus (3 points)

Implement the banding scheme with your choice of r, b, using your $t = 160$ random vectors, to estimate the pairs with similarity above $\tau = 0.75$ in the data set R. Report the fraction found above $\tau = 0.75$.