

Asmt 3: Distances and LSH

Turn in through Canvas by 2:45pm, then come to class:
Wednesday, February 6
100 points

Overview

In this assignment you will explore LSH and Euclidean distances.

You will use a data set for this assignment:

- <http://www.cs.utah.edu/~jeffp/teaching/cs5140/A3/R.txt>

As usual, it is highly 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 this directory: <http://www.cs.utah.edu/~jeffp/teaching/latex/>

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 $\tau = .85$.

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

	A	B	C	D
A	1	0.75	0.25	0.35
B	0.75	1	0.1	0.45
C	0.25	0.1	1	0.92
D	0.35	0.45	0.92	1

Using your choice of r and b and $f(\cdot)$, what is the probability of each pair of the four objects for being estimated to having similarity greater than $\tau = 0.85$? Report 6 numbers. (*Show your work.*)

2 Generating Random Directions (30 points)

A: (10 points) Describe how to generate a single random unit vector in $d = 10$ dimensions using only the operation $u \leftarrow \text{unif}(0, 1)$ which generates a uniform random variable between 0 and 1. (*This can be called multiple times.*)

B: (20 points) Generate $t = 160$ 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).

3 Angular Hashed Approximation (35 points)

Consider the $n = 500$ 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 $a, b \in \mathbb{R}^d$:

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

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 $s_{\text{ang}}(a, b)$ assumes that the input are unit vectors, and it takes 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 $\tau = 0.85$.

B: (20 points) Now compute the dot products and 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 $\tau = 0.85$.

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 $\tau = 0.85$ in the data set R . Report the fraction found above $\tau = 0.85$. Compare the runtime of this approach versus a brute force search.