Asmt 3: Distances and LSH

Turn in through Gradescope by 1:00pm, then come to class:
Wednesday, February 12
100 points

Overview

In this assignment you will explore LSH and Euclidean distances.

You will use two data sets for this assignment:

- http://www.cs.utah.edu/~jeffp/teaching/DM/A3/R1-updated.csv [updated!, Feb 4]
- http://www.cs.utah.edu/~jeffp/teaching/DM/A3/R2.csv

It is recommended that you use LaTeX for this assignment (or other option that can properly digitally render math). If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample in this place.

You could also utilize an LaTeX template specifically created for this assignment. Click here.

Below is an example of how to read .csv files from the local folder

```
import pandas as pd

# Path to your CSV file
file_path = "path_to_your_file/filename.csv"

# Read the CSV file into a DataFrame
df = pd.read_csv(file_path)

# Display the first few rows of the DataFrame
print(df.head())
```

Explanations of R1.csv and R2.csv

The two datasets come from a dataset of 500 most common words from a GloVe embedding in d=100 (each row unnormalize) and a 100-dimensional dataset with 500 points generated randomly (each row unnormalize). In Q2, you will figure out which is which.

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 $\tau = .70$.

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, with the following pair-wise similarities:

	A	B	C	D
A	1	0.72	0.32	0.35
B	0.72	1	0.20	0.55
C	0.32	0.20	1	0.89
D	0.35	0.55	0.89	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 $\tau = 0.70$)? Report 6 numbers. *(Show your work.)*

2 Angular Hashed Approximation and Random Vectors (65 points)

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 this converts them to $\bar{a} = a/\|a\|_2$ and $\bar{b} = b/\|b\|_2$ as part of the computation. The similarity $\mathsf{s}_{\text{ang}}(a, b)$ reports a value between 0 and 1, with, as usual, 1 meaning most similar. **Hint:** Compute L2 norm via np.linalg.norm( ).

A: (20 points) Now consider the $n = 500$ data points in $\mathbb{R}^d$ for $d = 100$ in a data set; there are two data sets: $R1$ and $R2$, given at the top. 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 ~~$\tau = 0.70$~~ $\tau = 0.6$.

B: (25 points) Now generate $t = 250$ unit vectors in $\mathbb{R}^d$ uniformly from $\mathbb{S}^{d-1}$ for $d = 100$. Plot the cdf of their pairwise angular similarities among $\binom{t}{2}$ pairs. Report the number with angular similarity above ~~$\tau = 0.70$~~ $\tau = 0.6$.

C: (20 points) Repeat the following process for the two data sets $R1$ and $R2$. Compute the Kolmogorov-Smirnov distance between the distributions (the largest difference measured across all cdf values) against the random distribution from part **B**. Report these values. Predict which distribution was uniformly random, and which was not.

3 Bonus (5 points)

Derive a model where the Kolmogorov-Smirnov distance (maximum difference between cdfs) corresponds to finding an anomaly. Write down the log-likelihood ratio for each $R1$ and $R2$ using the random vectors as a baseline.