Dimensionality Reduction

Steve Avsec

Illinois Institute of Technology

April 22, 2024

Overview

1. Locality-Sensitive Hashing
2. Differential Privacy

Distances

A "distance" function has to have the following properties:

- $d(x, y) \geq 0$.

Distances

A "distance" function has to have the following properties:

- $d(x, y) \geq 0$.
- $d(x, y) = 0 \Leftrightarrow x = y$.

Distances

A "distance" function has to have the following properties:

- $d(x, y) \geq 0$.
- $d(x, y) = 0 \Leftrightarrow x = y$.
- $d(x, y) = d(y, x)$.

Distances

A "distance" function has to have the following properties:

- $d(x, y) \geq 0$.
- $d(x, y) = 0 \Leftrightarrow x = y$.
- $d(x, y) = d(y, x)$.
- $d(x, z) \leq d(x, y) + d(y, z)$.

Some Examples

- Jaccard distance on sets:

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Some Examples

- Jaccard distance on sets:

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

- Hamming distance on vectors:

$$d(\mathbf{x}, \mathbf{y}) = |\{j : x_j \neq y_j\}|$$

Some Examples

- Jaccard distance on sets:

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

- Hamming distance on vectors:

$$d(\mathbf{x}, \mathbf{y}) = |\{j : x_j \neq y_j\}|$$

- Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{N} |x_j - y_j|^2$$

Some Examples

- Jaccard distance on sets:

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

- Hamming distance on vectors:

$$d(\mathbf{x}, \mathbf{y}) = |\{j : x_j \neq y_j\}|$$

- Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{N} |x_j - y_j|^2$$

- Edit distance on strings (also equal to the longest common subsequence).

Locality-Sensitive Functions

A family of functions **F** is said to be (d_1, d_2, p_1, p_2)-sensitive with respect to some distance if for every function $f \in \mathbf{F}$

- $d(x, y) \leq d_1$ implies that $f(x) = f(y)$ with probability at least p_1.
- $d(x, y) \geq d_2$ implies that $f(x) = f(y)$ with probability at most p_2.

Locality-Sensitive Functions

A family of functions **F** is said to be (d_1, d_2, p_1, p_2)-sensitive with respect to some distance if for every function $f \in$ **F**

- $d(x, y) \leq d_1$ implies that $f(x) = f(y)$ with probability at least p_1.
- $d(x, y) \geq d_2$ implies that $f(x) = f(y)$ with probability at most p_2.

Example: Minhash and Jaccard distance is $(d_1, d_2, 1 - d_1, 1 - d_2)$-sensitive for $0 \leq d_1 < d_2 \leq 1$.

Locality-Sensitive Functions

A family of functions **F** is said to be (d_1, d_2, p_1, p_2)-sensitive with respect to some distance if for every function $f \in \mathbf{F}$

- $d(x, y) \leq d_1$ implies that $f(x) = f(y)$ with probability at least p_1.
- $d(x, y) \geq d_2$ implies that $f(x) = f(y)$ with probability at most p_2.

Example: Minhash and Jaccard distance is $(d_1, d_2, 1 - d_1, 1 - d_2)$-sensitive for $0 \leq d_1 < d_2 \leq 1$.

For Hamming distance, the coordinate functions $f_i(\mathbf{x}) = x_i$ are $(d_1, d_2, 1 - \frac{d_1}{N}, 1 - \frac{d_2}{N})$-sensitive since the probability of agreement for a single f_i is exactly $1 - \frac{d(x,y)}{N}$.

Definition

A randomized algorithm $\mathcal{A}$ which takes a database as input is said to provide (ε, δ)-differential privacy if for datasets D_1 and D_2 that differ on a single element and all subsets $S \subseteq range(\mathcal{A})$:

$$P(\mathcal{A}(D_1) \in S) \leq e^{\varepsilon} P(\mathcal{A}(D_2) \in S) + \delta$$

Big Idea

Using distances like the Hamming distance and other distances that are similar plus some ideas from locality-sensitive hashing, one can create queries that compute

- Sums
- Counts
- Averages
- Mins and maxes

that provide differential privacy.