Topics in Data Science, Data Protection — Spring 2020
CS 4364 — L. Longpré

Class notes 3/12/2020

Differential Privacy

Definition

A randomized function K gives ε-differential privacy if for all data sets D_1 and D_2 differing on at most one element,

$$\Pr[K(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[K(D_2) \in S]$$

Last class example

Statistical information about embarassing behavior:

1. Flip a coin
2. If tails, respond truthfully
3. If heads, flip a second coin and respond yes if heads and no if tails.

Example in new setting

Database where each entry is a respondent's truthful answer. The database is published with each entry modified by the coin flip randomized algorithm.

The published database is $(\ln 3)$-differentially private. (How does changing an entry of the database change the probabilities of the corresponding published data?)

Laplace distribution

Laplace distribution centered at 0 and scale b:

$$\text{Lap}(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}}$$

It has variance $\sigma^2 = 2b^2$ and standard deviation $\sigma = \sqrt{2}\, b$

ℓ_1-sensitivity of a function f

General definition, but we apply the definition where f is a query on a database, x and y are databases that differ in at most one entry.

$$\Delta f = \max ||f(x) - f(y)||_1$$

Adding noise to a database

Given any function f applied on databases, modify the output of f by adding noise y generated with a Laplace distribution with scale $\Delta f/\varepsilon$. This mechanism will preserve ε-differential privacy.

Exercise

Suppose the database contains salaries. Suppose the maximum potential salary is \$10,000,000 and the database has at least 1,000 entries. The possible queries are COUNT and TOTAL.

1. What is ΔCOUNT?
2. What is ΔTOTAL?
3. What is standard deviation of TOTAL as modified by a Laplace distribution that will preserve $\sqrt{2}$-differential privacy?
4. What about COUNT?