Privacy-Preserving Natural Language Processing

Lecture 5 – Local DP and Randomized Response

Prof. Dr. Ivan Habernal

May 15, 2025

www.trusthlt.org
Chair of Trustworthy Human Language Technologies (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

RUHR UNIVERSITÄT BOCHUM RUB

TrustHLT 1001100110101

CENTER FOR TRUSTWORTHY DATA SCIENCE AND SECURITY

Recap

What we covered so far

- For provably private data analysis we need randomized algorithms
- Central (with a trusted curator) pure $(\varepsilon, 0)$ differential privacy
- Laplace mechanism: numeric queries, ℓ_1 sensitivity,
- Exponential mechanism: 'any-range' queries (arbitrary sets), utility function and its sensitivity

Today

- . . .
- Central (with a **trusted curator**) pure $(\varepsilon, 0)$ differential privacy
- . . .

What if we don't trust anyone?

Math refresher

Bernoulli trial (Bernoulli distribution, Bernoulli R.V., etc.)

Suppose that we run an experiment that succeeds with probability θ and fails with probability $1 - \theta$

A Bernoulli (or an indicator) random variable Y has the following probability distribution on $k \in \{0, 1\}$

$$\Pr(Y = 1) = \theta$$

$$\Pr(Y = 0) = 1 - \theta$$

which can be also expressed as

$$\Pr(Y = k) = p^k(1 - \theta)^{1-k}$$

Local differential privacy

Motivation

The notion of neighboring datasets, revisited

Database of size 1, replacing one person

Can we ask numeric queries on database of size 1?

Motivation

The notion of neighboring datasets, revisited

Database of size 1, replacing one person

Can we ask numeric queries on database of size 1?

Sure!

Example query: How many classes did you skip (and binge-watched Netflix instead)?

True value: between 0 and 15

You just need the ℓ_1 sensitivity for Laplace mechanism, decide on the ε and report

DP enables learning about population, not individuals!

What I'm really interested in is some statistics, e.g., the average number of lectures skipped across all students.

But 'privatizing' the 'raw' values shown previously removes the need for a trusted curator.

It comes with extra price (to be discussed later)

Simplifying queries

Now we know that we can privately (DP) answer statistical queries without a trusted curator

Let's simplify our example even more

Reporting single bit

Each person has a secret bit, e.g. "I took illegal drugs"

The query I might be interested in is 'What percentage of students take illegal drugs'

How to report a single bit

From true value to reported value

Two extremes

How to report a single bit

From true value to reported value

Two extremes

- Just report the true value
- Completely random answer (e.g., toss a coin)

Something between these two extremes

How to report a single bit: From true value to reported value

Idea: Report the true value with some probability, ‘flip’ and report otherwise

How to do the above ‘with some probability’ exactly?

Again: Use a randomized algorithm!

- Step 1: Draw a random value, either 0 or 1, from a Bernoulli distribution
- Step 2: If 1, report true value, if 0 report flipped value

(1 is usually associated with ‘true’ or ‘success’ and 0 with ‘false’ or ‘failure’)

How to parametrize this Bernoulli distribution?

Step 1: Draw 0 or 1 from $\mathrm{Ber}(\theta)$, Step 2: If 1, report true value, if 0 report flipped value

What θ would we need for modeling the extremes? (by the way: θ is publicly known! No 'security through obscurity'!)

Extreme 1: Just report the true value

How to parametrize this Bernoulli distribution?

Step 1: Draw 0 or 1 from $\mathrm{Ber}(\theta)$, Step 2: If 1, report true value, if 0 report flipped value

What θ would we need for modeling the extremes? (by the way: θ is publicly known! No 'security through obscurity'!)

Extreme 1: Just report the true value

$\mathrm{Ber}(\theta = 1)$ always true value (no privacy)

Extreme 2: Completely random answer

How to parametrize this Bernoulli distribution?

Step 1: Draw 0 or 1 from $\mathrm{Ber}(\theta)$, Step 2: If 1, report true value, if 0 report flipped value

What θ would we need for modeling the extremes? (by the way: θ is publicly known! No 'security through obscurity'!)

Extreme 1: Just report the true value

$\mathrm{Ber}(\theta = 1)$ always true value (no privacy)

Extreme 2: Completely random answer

$\mathrm{Ber}(\theta = 0.5)$ completely random $\rightarrow$ no matter what the raw value is, this is equivalent to reporting a random value

How to parametrize this Bernoulli distribution?

We saw that θ controls the privacy strength and want

$0.5 < \theta < 1.0$

where $0.5 \leftarrow \theta$ — more privacy

and $\theta \rightarrow 1.0$ — less privacy

How do we control privacy strength in DP?

$0 \leftarrow \varepsilon$ — more privacy

$\varepsilon \rightarrow \infty$ — less privacy

We need to find a function that maps ε to θ (and also satisfies DP)

First attempt

We want $0.5 < \theta < 1.0$ as a function of $0 \leq \varepsilon \leq \infty$

$\lim_{\varepsilon \to \infty} 1 = 1$ so does $\lim_{\varepsilon \to \infty} \frac{\varepsilon}{\varepsilon} = 1$ as well as $\lim_{\varepsilon \to \infty} \frac{\varepsilon}{\varepsilon+1} = 1$

But it doesn't work for $\lim_{\varepsilon \to 0}$ yet because $\lim_{\varepsilon \to 0} \frac{\varepsilon}{\varepsilon+1} \neq 0.5$

The easiest fix: Just add 1!

$\lim_{\varepsilon \to 0} \frac{\varepsilon+1}{\varepsilon+1+1} = 0.5$ and $\lim_{\varepsilon \to \infty} \frac{\varepsilon+1}{\varepsilon+1+1} = 1$

So let's try with $\theta = \frac{\varepsilon+1}{\varepsilon+2}$

Try to prove DP with our Bernoulli trial

For any 'neighboring' $D, D' \in \{0, 1\}$ and any $y \in \{0, 1\}$ the ε-DP definition must hold:

$$\frac{\Pr(\mathcal{M}(D) = y)}{\Pr(\mathcal{M}(D') = y)} \leq \exp(\varepsilon)$$

So let's try with Bernoulli trail and report the true with probability $\theta = \frac{\varepsilon+1}{\varepsilon+2}$ (and flip with probability $1 - \theta$)

Try to prove DP with our Bernoulli trial

For any 'neighboring' $D, D' \in \{0, 1\}$ and any $y \in \{0, 1\}$ the ε-DP definition must hold:

$$\frac{\Pr(\mathcal{M}(D) = y)}{\Pr(\mathcal{M}(D') = y)} \leq \exp(\varepsilon)$$

So let's try with Bernoulli trail and report the true with probability $\theta = \frac{\varepsilon+1}{\varepsilon+2}$ (and flip with probability $1 - \theta$)

Let $D = 0, D' = 1, y = 0$ (D' flipped from 1 to 0):

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\theta}{1-\theta} = \frac{\frac{\varepsilon+1}{\varepsilon+2}}{1-\frac{\varepsilon+1}{\varepsilon+2}} = \frac{\frac{\varepsilon+1}{\varepsilon+2}}{\frac{\varepsilon+2-(\varepsilon+1)}{\varepsilon+2}} = \frac{\varepsilon+1}{\varepsilon+2} \cdot \frac{\varepsilon+2}{1}$$
$$= \varepsilon + 1 \leq \exp(\varepsilon)$$

The same holds also for $D = 1, D' = 0, y = 1$ (D' flipped).
For not flipping, the ratio is 1, so the inequality also holds!

Great! We designed a DP algorithm for reporting a single bit (locally)

But can we do any better (with better utility?)

Improving the inequality

Our first attempt

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\theta}{1-\theta} = \frac{\frac{\varepsilon+1}{\varepsilon+2}}{1-\frac{\varepsilon+1}{\varepsilon+2}} = \varepsilon + 1 \leq \exp(\varepsilon)$$

We can approximate $\exp(\varepsilon)$ with a line around 0

$$\exp(\varepsilon) \geq 1 + \varepsilon$$

but also with a polynom of degree two

$$\exp(\varepsilon) \geq 1 + \varepsilon + \frac{\varepsilon^2}{2}$$

We get equality with the very definition of using power series $\exp(\varepsilon) := \sum_{k=0}^{\infty} \frac{x^k}{k!}$

Improving the inequality

Our first attempt

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\theta}{1-\theta} = \frac{\frac{\varepsilon+1}{\varepsilon+2}}{1-\frac{\varepsilon+1}{\varepsilon+2}} = \varepsilon + 1 \leq \exp(\varepsilon)$$

What if we replace $\varepsilon + 1$ with $\exp(\varepsilon)$ and worked it out backwards?

Our second attempt

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\theta}{1-\theta} = \ldots ? \ldots = \exp(\varepsilon) \leq \exp(\varepsilon)$$

Second attempt: Replace $\varepsilon + 1$ with $\exp(\varepsilon)$ backwards

First attempt recap (now rewriting with $2 = 1 + 1$)

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\frac{\varepsilon+1}{\varepsilon+1+1}}{1-\frac{\varepsilon+1}{\varepsilon+1+1}} = \frac{\frac{\varepsilon+1}{\varepsilon+1+1}}{\frac{\varepsilon+1+1-(\varepsilon+1)}{\varepsilon+1+1}}$$
$$= \frac{\varepsilon+1}{\varepsilon+1+1} \cdot \frac{\varepsilon+1+1}{1} = \varepsilon+1 \leq \exp(\varepsilon)$$

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\frac{\exp(\varepsilon)}{\exp(\varepsilon)+1}}{1-\frac{\exp(\varepsilon)}{\exp(\varepsilon)+1}} = \frac{\frac{\exp(\varepsilon)}{\exp(\varepsilon)+1}}{\frac{\exp(\varepsilon)+1-\exp(\varepsilon)}{\exp(\varepsilon)+1}}$$
$$= \frac{\exp(\varepsilon)}{\exp(\varepsilon)+1} \cdot \frac{\exp(\varepsilon)+1}{1} = \exp(\varepsilon) \leq \exp(\varepsilon)$$

Second attempt: Replace $\varepsilon + 1$ with $\exp(\varepsilon)$ backwards

We found the tightest θ

$$\frac{\Pr(y=0|D=0)}{\Pr(y=0|D'=1)} = \frac{\theta}{1-\theta} = \frac{\frac{\exp(\varepsilon)}{\exp(\varepsilon)+1}}{1-\frac{\exp(\varepsilon)}{\exp(\varepsilon)+1}} \underbrace{\leq}_{\text{(actually =)}} \exp(\varepsilon)$$

Recap our algorithm:

- Step 1: Draw a random value, either 0 or 1, from a Bernoulli distribution
- Step 2: If 1, report true value, if 0 report flipped value

We found $\theta = \frac{\exp(\varepsilon)}{\exp(\varepsilon)+1} = \frac{1}{1+\exp(-\varepsilon)}$ makes it $(\varepsilon, 0)$-DP

We discovered the "Randomized response" algorithm

- Step 1: Draw a random value, either 0 or 1, from a Bernoulli distribution
- Step 2: If 1, report true value, if 0 report flipped value

Parametrize Bernoulli with $\theta = \frac{\exp(\varepsilon)}{\exp(\varepsilon)+1} = \frac{1}{1+\exp(-\varepsilon)}$

Homework: Prove that Randomized response is $(\varepsilon, 0)$-DP (just do what we did for any pair of D, D', y)

We discovered the "Randomized response" algorithm

- Step 1: Draw a random value, either 0 or 1, from a Bernoulli distribution
- Step 2: If 1, report true value, if 0 report flipped value

Parametrize Bernoulli with $\theta = \frac{\exp(\varepsilon)}{\exp(\varepsilon)+1} = \frac{1}{1+\exp(-\varepsilon)}$

Homework: Prove that Randomized response is $(\varepsilon, 0)$-DP (just do what we did for any pair of D, D', y)

Local DP in NLP (discussion)

Summary

Local DP: No trusted curator needed

Enables privatize raw data and publish with plausible deniability

Randomized response: Known and used since the 1960's

Relevance to NLP: Publishing data for training (but comes with a cost)

License and credits

Credits

Ivan Habernal

`https://www.aclweb.org/anthology`

Partly inspired by lectures from Antti Honkela, Aurélien Bellet, Gautam Kamath