Foundations of machine learning

Differential privacy

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

Outline

- Precedents of differential privacy in the design of sensitive surveys.
- The definition of differential privacy:
  It should make (almost) no observable difference
  whether an individual is in the data or not.
- Properties:
  - Immunity to post-processing.
  - Composition and the “privacy budget.”
- Simple constructions of differentially private mechanisms:
  Add random noise to queries.

Takeaways for this part of class

- Naive notions of privacy (“removing identifying information” or “aggregation”) are not immune to the availability of auxiliary information.
- “Differential privacy” provides a coherent and robust definition.
- Random noise is necessary for privacy.
- Responding to additional queries spends a “privacy budget.”

Naive notions of privacy

- Removing “identifying information” does not preserve privacy:
    - A small number of “non-sensitive” variables
      (e.g., what movies you recently watched, what you had for breakfast the last few days, ...)
    - typically identifies you uniquely!
- Aggregation does not preserve privacy:
    - A study reports, for a sample of patients with a certain disease, the share of patients with a certain genetic variant (SNP), for a large number of genes.
    - It turns out that from such aggregates, we can identify whether any given individual was in the sample (and thus has the disease).

An example and historical precedent

- Suppose you are running a sensitive survey.
  E.g., you might want to learn what share of students consume illegal drugs.
- How can you do so such that
  1. no respondent runs a legal risk by responding truthfully, and
  2. you still learn the aggregate share θ accurately?
- Possible solution: Instruct each respondent to do the following.
  1. Flip a coin.
     If the coin comes up heads, respond truthfully.
  2. If the coin comes up tails, flip again.
     If the second flip is heads, respond truthfully, else lie.

Example continued

Properties of this scheme:

1. Every participant has plausible deniability.
2. The share p responding "yes" equals

$$p = \frac{3}{4}\theta + \frac{1}{4}(1 - \theta) = \frac{1}{4} + \frac{1}{2}\theta,$$

from which we can easily recover the true share θ.

Definitions

Construction of differentially private mechanisms

References

Definitions

- Throughout, we focus on discrete data, represented by vectors $x \in \mathbb{N}^{\mathcal{X}}$.
  x_i is the count of individuals of type $i \in \mathcal{X}$ in the data.
- *Randomized Algorithms (Def 2.2)*:
  Random mappings $\mathcal{M}$ from $\mathbb{N}^{\mathcal{X}}$ to some discrete range B.
  $M(x) \in \Delta(B)$ is the probability distribution over B.
- *Distance between databases (Def 2.3)* x and y:
  $\|x - y\|_1 = \sum_{i \in \mathscr{X}} |x_i - y_i|$.
  In particular, if y adds or drops one individual relative to x, then $\|x - y\|_1 = 1$.

Definitions continued

- *Differential privacy (Def 2.4)*:
  A randomized algorithm $\mathcal{M}$ is ε-differentially private if
  For all $\mathcal{S} \subset B$, and for all x, y with $\|x - y\|_1 = 1$,

$$\frac{P(\mathcal{M}(x) \in \mathcal{S})}{P(\mathcal{M}(y) \in \mathcal{S})} \leq \exp(\varepsilon).$$

- *Privacy loss* from observing ξ:

$$\log\left(\frac{P(\mathcal{M}(x) = \xi)}{P(\mathcal{M}(y) = \xi)}\right).$$

  This is bounded by ε for ε-differentially private $\mathcal{M}$.

Practice problem

Discuss: Does differential privacy capture the socially relevant notion of privacy?

Some properties

- *Post-processing (Prop 2.1)*:
  If $\mathcal{M}$ is ε-differentially private
  then the same holds true for $f \circ \mathcal{M}$ for any function f.

- *Composition (Theo 3.14)*:
  If $\mathcal{M}_j$ is ε_j-differentially private for $j = 1, 2$, and the $\mathcal{M}_j$ are statistically independent,
  then $(\mathcal{M}_1, \mathcal{M}_2)$ is $(\varepsilon_1 + \varepsilon_2)$- differentially private.

  This compositional property is often described in terms of a “privacy budget” that we can spend.

Practice problem

Prove these properties.

What differential privacy does and does not deliver

- It makes (almost) no difference to an individual whether they are represented in the data or not.
- This holds no matter who gets to see the queries, what other information they possess, or what actions they might take based on the queries.
- This does *not* mean that no harm can result to an individual from the data – just that their individual participation makes no difference.
- Example:
  - A study based on medical records, released in a differentially private manner, documents the relation between smoking and cancer.
  - As a consequence, the insurance premiums for a smoker go up.
  - But: This would have happened whether the individual's records were part of the study or not.

Definitions

Construction of differentially private mechanisms

References

Randomization is necessary for differential privacy

- Consider a deterministic mechanism $\mathcal{M}$.
- Unless $\mathcal{M}$ is trivial, there are values x, y of the data such that $\mathcal{M}(x) \neq \mathcal{M}(y)$.
- We can reach y from x by adding or removing entries to the data one at a time.
- At one of these steps from u to v, we must have $\mathcal{M}(u) \neq \mathcal{M}(v)$, while $\|u - v\|_1 = 1$.
- If some adversary has auxiliary information that the data are either u or v,
  they can identify which it is from query $\mathcal{M}$,
  and thus identify whether a particular individual is in the data or not.

The Laplace mechanism

- The *Laplace distribution* $Lap(b)$ has density

$$\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right).$$

- The $\mathcal{L}_1$ sensitivity of a function f from $\mathbb{N}^{\mathcal{X}}$ to $\mathbb{R}^k$ is defined as

$$\Delta f = \max_{x,y:\|x-y\|_1=1} \|f(x)-f(y)\|_1$$

- For such a function f, consider the randomized algorithm

$$\mathcal{M}(x,f,\varepsilon) = f(x) + (Y_1,\ldots,Y_k),$$

  where the Y_j are i.i.d. $Lap(\Delta f/\varepsilon)$.

Practice problem

Prove that this algorithm satisfies ε-differential privacy.

Examples

- *Counts*:
  Let $f(x)$ be the number of individuals in the data satisifying some property.
  Then $\Delta f = 1$, and $f(x) + Y$ with $Y \sim Lap(1/\varepsilon)$ is ε-differentially private.

- *Composition of counts*:
  We can report k such queries, each with $Y \sim Lap(k/\varepsilon)$,
  to get an ε-differentially private algorithm for their composition.

- *Histograms*:
  Let $f(x)$ be the vector of counts of individuals falling into each of a number of categories.
  Then $\Delta f = 1$ again, and $f(x) + (Y_1, \ldots, Y_k)$ with $Y_j \sim Lap(1/\varepsilon)$
  is again ε-differentially private.

  Note that we need much less noise
  relative to the case where the counts for each category are independent.

The exponential mechanism

- Suppose the query is to inform a decision a.
- The decision-maker's expected utility given the full data x is $u(x,a)$.
- Let

$$\Delta u = \max_{a} \max_{x,y:\|x-y\|_1=1} \|u(x,a) - u(y,a)\|_1.$$

- The *exponential mechanism* reports a with probability

$$\frac{\exp\left(\frac{\varepsilon u(x,a)}{2\Delta u}\right)}{\sum_{a'} \exp\left(\frac{\varepsilon u(x,a')}{2\Delta u}\right)}.$$

- This mechanism
  1. Satisfies ε-differential privacy.
  2. Delivers high expected utility.

References

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, *9(3–4):211–407, chapters 2 and 3.*