CS6781 - Theoretical Foundations of Machine Learning

Lecture 25: Differential Privacy

05/07/2020

Lecturer: Nika Haghtalab *Readings: N/A*
Scribe: Rishi Bommasani and Jonathan Chang

1 Introduction

In this lecture, we study theoretical definitions of privacy.

1.1 Motivating Example

A motivating scenario where we would want to have some notion of privacy is when an instructor would like to release a population statistic like the average grade for an assignment. A student taking this class would not want their individual grades revealed or have an adversarial algorithm exploit the population statistics to extract this information. Here are some ways that a student's information is not private with the average

- If there are n students in the class, $n-1$ students could collude to figure out a single student's grade.
- If a student(s) drop the course and the mean is updated, the differential in the average could be exploited to reveal the grade(s) of the student(s) who dropped the course.

From this example we can see that the *reporting of statistics at population-scale is not sufficient to ensure the privacy of individuals*.

2 ϵ-Differential Privacy

Given the insufficiency of reporting statistics at population-scale, we consider *differential privacy* as a formalism for providing rigorous privacy guarantees.

Definition 2.1. (Adjacent Sets) Sets S_1 and S_2 are *adjacent sets* if they differ by a single element.

Definition 2.2. (ϵ-Differential Privacy) A randomized algorithm $\mathcal{A} : \mathcal{X} \to \mathcal{R}$ satisfies *ϵ-differential privacy* (ϵ-DP) if for all adjacent sets $S_1, S_2 \in \mathcal{X}$, range of outcomes $\mathcal{R}$, and any subset of outcomes $R \subseteq \mathcal{R}$ we have

$$\Pr\left[\mathcal{A}(S_1) \in R\right] \leq \exp(\epsilon) \Pr\left[\mathcal{A}(S_2) \in R\right]$$

Approximating $\exp(\pm\epsilon)$ by $1 \pm \epsilon$, for small ϵ this approximately states

$$1 - \epsilon \leq \frac{\Pr\left[\mathcal{A}(S_1) \in R\right]}{\Pr\left[\mathcal{A}(S_2) \in R\right]} \leq 1 + \epsilon$$

In general, we use R that is a single element of a discrete set or infinitesimal set around a single element r of a continuous set. The latter part is the same as requiring that the ratio of the densities of each outcome $r \in \mathcal{R}$ is bounded between $\exp(-\epsilon)$ and $\exp(\epsilon)$.

Intuitively, ϵ-DP is a stability guarantee with respect to the information provided by any one individual. That is, the behavior of the mechanism does not change substantially when an individual's information is included/excluded.

One simple way to achieve this stability is for the algorithm to completely ignore the input set S and output a fix value. This, of course, is differentially private even for $\epsilon = 0$. But, such a constant function also reveals no population level information about S as a whole either. In this lecture, we see how we can introduce privacy while still computing useful quantities. A basic technique for doing this is to introduce some noise to the computation. We will see an example of this in next section.

2.1 Laplace Mechanism

We first explore how to make the computation of a general function f by adding noise to its outcome. This mechanism is for a single real number in a differentially-private manner, e.g., the average grade in a course. We do this by adding to $f(S)$ a noise that is drawn from a Laplace distribution.

Definition 2.3. (Laplace distribution) The *Laplace distribution* is a continuous probability distribution centered at 0 and with parameter b has the density function

$$\mathrm{Lap}(x \mid b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right). \tag{1}$$

The distribution is like gluing together two exponential distributions.

Let's see how adding Laplace noise to the average of m numbers makes it differentially private. Suppose we have a set S of m values in $[0, b]$ and we want to output an estimate of their average while preserving differential privacy. That is, $f(S) = \frac{1}{|S|} \sum_{z \in S} z$. And we release

$$\mathcal{M}(S) = f(S) + \eta \quad \text{For } \eta \sim Lap\left(\frac{b}{m\epsilon}\right).$$

First, we show that $\mathcal{M}(S) \approx f(S)$ with high probability. That is, the outcome of this mechanism is very accurate.

$$\Pr\left[\left|\mathcal{M}(S) - \frac{1}{|S|}\sum_{z \in S} z\right| \geq \alpha\right] = \Pr_{x \sim Lap\left(\frac{b}{m\epsilon}\right)}\left[|x| \geq \alpha\right] = \exp\left(\frac{-\alpha\epsilon m}{b}\right) \leq \beta$$

for $\alpha = \frac{b}{m\epsilon}\ln(1/\beta)$.

Next, we show that $\mathcal{M}(S)$ is ϵ-differentially private:

$$\Pr[\mathcal{M}(S_1) = r] = \exp\left(-\frac{|r|\epsilon m}{b}\right).$$

Now the worst case S_2 is such that the average has moved by $\pm b/m$ value. In that case we have

$$\Pr[\mathcal{M}(S_2) = r] = \exp\left(-\frac{|r - b/m|\epsilon m}{b}\right).$$

Then, we have

$$\frac{\Pr[\mathcal{M}(S_1) = r]}{\Pr[\mathcal{M}(S_2) = r]} \leq \exp\left(\frac{-|r|\epsilon m + |r - b/m|\epsilon m}{b}\right) \leq \exp\left(\epsilon\right).$$

Adding Laplace noise to make computations private goes beyond computing the average. But it can make any function $f : \mathcal{X}^m \to \mathbb{R}$ private. To discuss the notion of stability, we need to understand the effect one single entry can have on $f(S)$. This is defined below.

Definition 2.4. (Sensitivity) The sensitivity of a real-valued function f is

$$\Delta f \triangleq \max_{\text{adjacent } S_1, S_2} |f(S_1) - f(S_2)|.$$

Intuitively, the larger that the sensitivity Δf is, the more difficult it is to preserve ϵ-differential privacy. Therefore, we take Δf into account when deciding how much noise to add to the computation.

Definition 2.5. (Laplace mechanism) Given a function $f : \mathcal{X}^m \to \mathbb{R}$ the *Laplace mechanism* is defined by

$$\mathcal{M}_{\text{Lap}}\left(S, f, \epsilon\right) \triangleq f(S) + \eta. \tag{2}$$

where η is a random variable drawn from $\text{Lap}\left(\frac{\Delta f}{\epsilon}\right)$.

Just as we showed that adding Laplace noise to average computation, where sensitivity is b/m, makes the computation private and has high accuracy, we can prove the following guarantees for the general Laplace Mechanism.

Theorem 2.6. *$\mathcal{M}_{Lap}\left(S, f, \epsilon\right)$ is ϵ-differentially private. Furthermore, with probability* $1 - \beta$

$$|\mathcal{M}_{Lap}\left(S, f, \epsilon\right) - f(S)| \leq \frac{\Delta f}{\epsilon}\ln(1/\beta).$$

3 Privacy-preserving Machine Learning

As we saw, the Laplace mechanism is not only a valid method for satisfying ϵ-differential privacy but also is accurate. Given that we are interested in accurate private mechanisms, a natural next question is to ask what can we learn under the constraints of privacy.

3.1 Privacy-preserving PAC Learning

We define a ϵ-differentially private (α, β)-PAC learner on the realizable PAC learning setting.

Definition 3.1. A learning algorithm $\mathcal{A}$ is an ϵ-differentially private (α, β)-PAC learner for hypothesis class $\mathcal{H}$ if, for any distribution $\mathcal{D}$ that is realizable with respect to an unknown $h^* \in \mathcal{H}$, for privacy parameter $\epsilon > 0$, for error parameter $\alpha > 0$, and confidence parameter $\beta > 0$, with probability at least $1 - \beta$, $\mathcal{A}$ outputs a hypothesis $h \in \mathcal{H}$ using at most $\mathrm{poly}(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\epsilon}, \text{rep. of } \mathcal{H})$ samples such that

$$\mathrm{err}_{\mathcal{D}}(h) \leq \alpha.$$

We say that $\mathcal{H}$ is ϵ-differentially private PAC learnable, if there is an ϵ-differentially private (α, β)-PAC learner for all α and β.

In the last lecture, we worked with statistical query model and alluded to the fact that using population level statistics provides a way to learn while preserving privacy. We formalize that now.

Theorem 3.2. *Any $\mathcal{H}$ that is efficiently learnable in the statistical query model is also efficiently PAC learnable with ϵ-differentially privacy. In particular, if the statistical query model uses K queries of tolerance at least τ, then m samples suffice to ϵ-privately PAC learn $\mathcal{H}$, where*

$$m \in \mathcal{O}\left(\frac{K}{\epsilon\tau} \ln\left(\frac{K}{\beta}\right) + \frac{1}{\tau^2} \ln\left(\frac{K}{\beta}\right)\right).$$

Before we prove this, let us discuss how we can combine privately computed values to retain their privacy guarantees.

Theorem 3.3 (Privacy under composition)**.** *Given mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$ that are each ϵ-differentially private, any $\mathcal{M}$ that is defined as a combination of the $\mathcal{M}_i$ is $k\epsilon$-differentially private.*

We now prove Theorem 3.2.

Proof. The main idea is that sensitivity of statistical queries is small. So we can use Laplace mechanism to effectively compute statistical queries for the required tolerance. Assume that we used statistical queries $\psi_1, \ldots, \psi_K$ of tolerance τ for learning $\mathcal{H}$.

$$\forall i \in [K], \text{let } f_i(S) = \frac{1}{m} \sum_{x \in S} \psi_i\left(x, h^*(x)\right) \approx \mathbb{E}_{x \in \mathcal{D}}\left[\psi_i\left(x, h^*(x)\right)\right] \tag{3}$$

$$\forall i \in [K], \Delta f_i = \max_{\text{adjacent } S_1, S_2} \frac{1}{m} \left| \left(\sum_{x \in S_1} \psi_i\left(x, h^*(x)\right) - \sum_{x \in S_2} \psi_i\left(x, h^*(x)\right) \right) \right| \leq \frac{1}{m} \tag{4}$$

We will compute each f_i and given set S, we will release $\mathrm{Lap}\left(S, f_i(\cdot), \frac{K}{\epsilon m}\right)$.

By the privacy guarantees on the Laplace mechanism, we know this is $\frac{\epsilon}{K}$-differentially private. By the composition theorem for differential privacy, we know that any computation that depends on these k mechanisms will yield a ϵ-differentially private mechanism overall. This means we have shown the privacy guarantees that are desired.

By the accuracy guarantees on the Laplace mechanism, for every f_i and every S, we know that with probability at least $1-\beta$, the error in approximating $f_i(S)$ is at most $\frac{K}{\epsilon m} \ln\left(\frac{1}{\beta}\right)$. If the number of samples m is at least $O\left(\frac{2K}{\epsilon\tau} \ln\left(\frac{1}{\beta}\right)\right)$, then with probability at least $1-\beta$, the error is at most $\frac{\tau}{2}$. Further, as we have seen last lecture, if the number of samples m is at least $O\left(\frac{1}{\tau^2} \ln\left(\frac{K}{\beta}\right)\right)$, then $|f_i(S) - \mathbb{E}_{x\in\mathcal{D}}\left[\psi_i\left(x, h^*(x)\right)\right]| \leq \frac{\tau}{2}$. If m is at least both of these amounts, then by triangle inequality we have that with probability at least $1-\beta$, $\left|\text{Lap}\left(S, f_i(\cdot), \frac{K}{\epsilon m}\right) - \mathbb{E}_{x\in\mathcal{D}}\left[\psi_i\left(x, h^*(x)\right)\right]\right| \leq \frac{\tau}{2}$.

□