Mathematical Foundations of Machine Learning (CS 4783/5783)

Lecture 21: Differential Privacy

1 Differential Privacy

Differential Privacy is a strong notion of privacy for an algorithm that ensures that we cannot detect if one entry of a dataset is replaced. Specifically, let A be a randomized algorithm that takes as input a sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and outputs $A(S)$ in some arbitrary outcome space.

Definition 1. *We say that A is (ϵ, δ) differentially private if for any sample S and sample S' that differ on at most one data point, and for any set C over the space of outcomes,*

$$P(A(S) \in C) \leq e^{\epsilon} P(A(S') \in C) + \delta$$

Note that since S and S' differ on at most one data point, the above definition tells us that both

$$P(A(S) \in C) \leq e^{\epsilon} P(A(S') \in C) + \delta$$

and that

$$P(A(S') \in C) \leq e^{\epsilon} P(A(S) \in C) + \delta$$

Specifically, as ϵ and δ are taken to be very small this says $P(A(S) \in C)$ and $P(A(S') \in C)$ are very close and so we cant distinguish if we have run our method on S or S'.

2 The Laplace Mechanism

Say we want a differentially private version of a real valued function f on a given sample S. One way to obtain such a version is to first evaluate f on a given sample S then add noise to it to guarantee differential privacy. Specifically, say we want a differentially private version of function f. In this case, let

$$M = \max_{S,S' \text{ s.t. } S',S \text{ vary on one point}} f(S) - f(S')$$

Now we could set

$$A(S) = f(S) + \frac{M}{\epsilon} X$$

where X is drawn from the Laplace distribution $\text{Laplace}(0, 1)$. That is, distribution with density function

$$p(X) = \frac{1}{2} e^{-|X|}$$

Lemma 1. *Let*

$$A(S) = f(S) + \frac{M}{\epsilon} X$$

where $X \sim \text{Laplace}(0, 1)$ and $M = \max_{S,S' \text{ s.t. } S',S \text{ vary on one point}} f(S) - f(S')$. The algorithm A is $(\epsilon, 0)$ differentially private.

Proof. Since $A(S) = f(S) + \frac{M}{\epsilon}\ X$, we have that $A(S) \sim \text{Laplace}(f(S), \frac{M}{\epsilon})$. Hence, we have that the probability density function of $A(S)$ is given by

$$p_{A(S)}(x) = \frac{\epsilon}{2M} e^{-\frac{\epsilon|x-f(S)|}{M}}$$

Similarly, the density function for $A(S')$ for any S' that differs from S on at most one point if given by

$$p_{A(S')}(x) = \frac{\epsilon}{2M} e^{-\frac{\epsilon|x-f(S')|}{M}}$$

Hence,

$$\frac{p_{A(S)}(x)}{p_{A(S')}(x)} = \frac{e^{-\frac{\epsilon|x-f(S)|}{2M}}}{e^{-\frac{\epsilon|x-f(S')|}{2M}}} = e^{\frac{\epsilon}{2M}(|x-f(S')|-|x-f(S)|)} \leq e^{\frac{\epsilon}{M}|f(S)-f(S')|} \leq e^{\epsilon}$$

Next note that for any set C, using the above,

$$P(A(S) \in C) = \int_C p_{A(S)}(x)dx \leq e^{\epsilon} \int_C p_{A(S')}(x)dx = e^{\epsilon} P(A(S') \in C)$$

Thus we have proved that the algorithm is $(\epsilon, 0)$ differentially private. $\square$

An example application is when $S = \{x_1, \ldots, x_n\}$ where each $x_t \in [-1, 1]$ and $f(S) = \frac{1}{n}\sum_{t=1}^n x_t$. In this case note that if $S' = \{x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n\}$, then,

$$f(S) - f(S') = \frac{1}{n}(x_i - x'_i) \leq \frac{2}{n}$$

Hence $M \leq \frac{2}{n}$ and so in this case, to make mean $\epsilon, 0$ differentially private, we need to add Laplace noise of $\text{Laplace}(0, \frac{2}{\epsilon n})$

3 Some Properties

The first important property of differential privacy is that post processing preserves privacy. Say algorithm A is (ϵ, δ) differentially private and say we apply a function g on outcome of algorithm A and output $g(A(S))$. Such post processing preserves privacy.

Lemma 2. *Let A be an (ϵ, δ) differentially private algorithm. Let g be any function on the space of outcomes of the algorithm A. Then, the algorithm B that computes $B(S) = g(A(S))$ is also (ϵ, δ) differentially private.*

Proof. Consider any set C on the space of outcomes of algorithm B. Define the set

$$D = \{d : g(d) \in C\}$$

that is D is the set of entries such that g applied to an element in D returns an outcome in set C. Note that,

$$P(B(S) \in C) = P(g(A(S)) \in C) = P(A(S) \in D)$$

Now using the differential privacy of A, we have

$$P(B(S) \in C) = P(A(S) \in D) \leq e^{\epsilon} P(A(S') \in D) + \delta$$

But if $A(S') \in D$, then $g(A(S')) \in C$ by definition of set D and so

$$P(B(S) \in C) \leq e^{\epsilon} P(A(S') \in D) + \delta = e^{\epsilon} P(g(A(S')) \in C) + \delta = e^{\epsilon} P(B(S') \in C) + \delta$$

Thus we can conclude that B is ϵ, δ differentially private.

$\square$