Machine Learning Exercise Sheet 13
Privacy

In-class Exercises

Differential Privacy

Problem 1: Prove that the Laplace mechanism is ϵ-Differentially Private.
Note: The Laplace mechanism is defined as follows: $\mathcal{M}_f(X) = f(X) + Z$ where $Z \sim \mathrm{Lap}(0, \frac{\Delta_1}{\epsilon})^d$ and the global l_1 sensitivity of a function $f : \mathcal{X} \to \mathbb{R}^d$ is $\Delta_1 = \sup_{X \simeq X'} ||f(X) - f(X')||_1$.

From the definition we have that a randomized mechanism $\mathcal{M}_f : \mathcal{X} \to \mathcal{Y}$ is ϵ-differentially private if **for all** neighboring inputs $X \simeq X'$ and **for all** sets of outputs $Y \subseteq \mathcal{Y}$ we have:

$$\exp^{-\epsilon} \leq \frac{\mathbb{P}[\mathcal{M}_f(X) \in Y]}{\mathbb{P}[\mathcal{M}_f(X') \in Y]} \leq \exp^{\epsilon}.$$

The Laplace mechanism is defined as follows: $\mathcal{M}_f(X) = f(X) + Z$ where $Z \sim \mathrm{Lap}(0, \frac{\Delta_1}{\epsilon})^d$ and the global l_1 sensitivity of a function $f : \mathcal{X} \to \mathbb{R}^d$ is $\Delta_1 = \sup_{X \simeq X'} ||f(X) - f(X')||_1$.

The Laplace mechanism itself follows an isotropic, multivariate Laplace distribution with mean $f(X)$, i.e. $\mathcal{M}_f(X) \sim \mathrm{Lap}(f(X), \frac{\Delta_1}{\epsilon})^d$. In the following, we use $p_X(y)$ as a shorthand for its probability density function $\mathrm{Lap}\left(y \mid f(X), \frac{\Delta_1}{\epsilon}\right)^d$ and $p_{X'}(y)$ as a shorthand for the probability density function of $\mathcal{M}_f(X')$

To prove ϵ-differential privacy, we first bound the ratio of the density functions of $\mathcal{M}_f(X)$ and $\mathcal{M}_f(X')$. We start by plugging in the definition $\mathrm{Lap}(y \mid \mu, b) = \frac{1}{2b} \exp^{-\frac{|Z-\mu|}{b}}$ and using the fact that the noise is i.i.d. per dimension. We have:

$$\begin{aligned}
\frac{p_X(y)}{p_{X'}(y)} &= \prod_{i=1}^{d} \frac{\exp^{-\frac{\epsilon}{\Delta_1}|f(X)_i - y_i|}}{\exp^{-\frac{\epsilon}{\Delta_1}|f(X')_i - y_i|}} \\
&= \prod_{i=1}^{d} \exp^{\frac{\epsilon}{\Delta_1}\left[|f(X')_i - y_i| - |f(X)_i - y_i|\right]} \\
&\leq \prod_{i=1}^{d} \exp^{\frac{\epsilon}{\Delta_1}\left[|f(X')_i - f(X)_i|\right]} \\
&= \exp^{\frac{\epsilon}{\Delta_1}\sum_{i=1}^{d}\left[|f(X')_i - f(X)_i|\right]} \\
&= \exp^{\frac{\epsilon}{\Delta_1}||f(X') - f(X)||_1} \\
&\leq \exp^{\frac{\epsilon}{\Delta_1}\Delta_1} = \exp^{\epsilon}
\end{aligned}$$

where the first inequality comes from the (reverse) triangle inequality and the second inequality is from the definition of global sensitivity.

We can now use the derived bound to find an upper bound on $\mathbb{P}[\mathcal{M}_f(X) \in Y]$:

$$\begin{aligned}\mathbb{P}[\mathcal{M}_f(X) \in Y] &= \int_Y p_X(y)\, dy \\ &\leq \int_Y \exp^{\epsilon} p_{X'}(y)\, dy \\ &= \exp^{\epsilon} \int_Y p_{X'}(y)\, dy \\ &= \exp^{\epsilon} \mathbb{P}[\mathcal{M}_f(X') \in Y].\end{aligned}$$

Thus

$$\frac{\mathbb{P}[\mathcal{M}_f(X) \in Y]}{\mathbb{P}[\mathcal{M}_f(X') \in Y]} \leq \exp^{\epsilon}.$$

Since the neighboring relation $\simeq$ is symmetric we can repeat the above derivation to obtain

$$\frac{\mathbb{P}[\mathcal{M}_f(X') \in Y]}{\mathbb{P}[\mathcal{M}_f(X) \in Y]} \leq \exp^{\epsilon}$$

where now $f(X')$ is in the numerator and $f(X)$ is in the denominator, which then gives us

$$\exp^{-\epsilon} \leq \frac{\mathbb{P}[\mathcal{M}_f(X) \in Y]}{\mathbb{P}[\mathcal{M}_f(X') \in Y]}.$$

Homework

Differential privacy

Problem 2: Assume that you have trained an univariate linear regression model $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ with $\boldsymbol{w} \in \mathbb{R}^D$ for D-dimensional binary data from input space $\mathcal{X} = \{0,1\}^D$. You want to make its prediction ϵ-differentially private with respect to changes in a single input dimension, i.e. $\boldsymbol{x} \simeq \boldsymbol{x}' \iff ||\boldsymbol{x} - \boldsymbol{x}'||_0 = 1$ for all points from input space $\mathcal{X}$.

a) Compute the global Δ_1 sensitivity of f w.r.t. "$\simeq$".

The global Δ_1 sensitivity is defined as $\sup_{\boldsymbol{x}\simeq\boldsymbol{x}'} ||f(\boldsymbol{x}) - f(\boldsymbol{x}')||_1$. Inserting the definition of f

results in

$$\begin{aligned}\Delta_1 &= \sup_{\boldsymbol{x}\simeq\boldsymbol{x}'} ||\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{x}'||_1\\ &= \sup_{\boldsymbol{x}\simeq\boldsymbol{x}'} ||\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}')||_1\\ &= \sup_{\boldsymbol{x}\simeq\boldsymbol{x}'} |\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}')|,\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}')$ is scalar.

By the definition of the neighboring relation, for any two $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}$ with $\boldsymbol{x} \simeq \boldsymbol{x}'$ we have $\boldsymbol{x} - \boldsymbol{x}' = \pm\boldsymbol{e}^{(d)}$ for some d, where $\boldsymbol{e}^{(d)}$ is the canonical unit vector with non-zero entry in dimension d. Therefore, $|\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}')| = |w_d|$ and thus

$$\Delta_1 = \max_{d\in\{1,\dots,D\}} |w_d| \tag{1}$$

That is, the Δ_1 sensitivity of f w.r.t. "$\simeq$" is determined by the largest weight.

b) To ensure differential privacy, you want to use the Laplace mechanism $\mathcal{M}_{f,\text{Lap}}(\boldsymbol{x}) = f(\boldsymbol{x}) + z$ with $z \sim \text{Lap}(\mu, b)$. Based on your result from a), which values do you have to use for μ and b to ensure ϵ-differential privacy w.r.t. to neighboring relation "$\simeq$"?

To ensure ϵ-differential privacy, one has to use $\text{Lap}\left(0, \frac{\Delta_1}{\epsilon}\right)$, which in our case is $\text{Lap}\left(0, \frac{\max_d |w_d|}{\epsilon}\right)$.

c) Instead of randomizing the output of our model, we can also guarantee differential privacy by randomizing its inputs. Prove that the randomized mechanism $\mathcal{M}' = f(\boldsymbol{x}+\boldsymbol{z})$ with $\boldsymbol{z} \sim \text{Lap}\left(0, \frac{1}{\epsilon}\right)^D$ is ϵ-differentially private w.r.t to neighboring relation "$\simeq$".

Using the identity matrix $\boldsymbol{I}$, the mechanism $\mathcal{M}' = f(\boldsymbol{x} + \boldsymbol{z})$ can be rewritten as $f(\boldsymbol{I}\boldsymbol{x} + \boldsymbol{z})$. Based on this formulation, it can be seen that $\mathcal{M}' = f \circ \mathcal{M}_{h,\text{Lap}}$ with $h(\boldsymbol{x}) = \boldsymbol{I}\boldsymbol{x}$.

The global Δ_1 sensitivity of h is 1, since changing one bit in its input will change exactly one bit in its output. Or more formally: For any $\boldsymbol{x} \simeq \boldsymbol{x}'$ that differ in dimension d, we have

$$||\boldsymbol{I}\boldsymbol{x} - \boldsymbol{I}\boldsymbol{x}'|| = ||\boldsymbol{I}(\boldsymbol{x} - \boldsymbol{x}')|| = ||\boldsymbol{I}(\pm\boldsymbol{e}^d)|| = 1.$$

This shows that $\text{Lap}\left(0, \frac{1}{\epsilon}\right)^D = \text{Lap}\left(0, \frac{\Delta_1}{\epsilon}\right)^D$, i.e. the noise is correctly calibrated to the global Δ_1 sensitivity of h and thus $\mathcal{M}_{h,\text{Lap}}$ is ϵ-differentially private.

Due to the robustness of differential privacy to post-processing (see p.22), $\mathcal{M}' = f \circ \mathcal{M}_{h,\text{Lap}}$ is also ϵ-differentially private.

Problem 3: You are given a dataset with n instances $\{x_1, \dots, x_n\}$, with $x_i \in \mathcal{X}$. The instances are randomly split into disjoint groups $G_1, G_2, \dots G_m$, each of size $\frac{n}{m}$ (assume that m divides n, i.e. $\frac{n}{m}$ is an

integer).
First you apply an *arbitrary* function $f : \mathcal{X}^{\frac{n}{m}} \to [a, b]$ (where a and b are given constants) to each of the groups, i.e. you compute $g_1 = f(G_1), g_2 = f(G_2), \dots, g_m = f(G_m)$. Then you compute the final output by aggregating the per-group outputs by computing either their mean or their median.

a) Derive the global Δ[illegible]unction $f' := \text{mean}(f(G_1), \dots, f(G_m))$.

> Changing any sin[illegible]lifies one of the groups G_i so it is sufficient to reason *only* about the sensiti[illegible]n function operating on the groups.
> Since f is bounde[illegible]ction takes as input m numbers, $g_1, \dots, g_m$ in the interval $[a, b]$. Changing [illegible]ge at most one g_i, and in the worst case the change can be anywhere in t[illegible]
> In the worst-case the output of one g_i changes from b to a. Thus, the global Δ_1 sensitivity of f' is $\frac{b-a}{m}$.

b) Derive the global Δ_1 sensitivity of the function $f'' := \text{median}(f(G_1), \dots, f(G_m))$.

> As in the previous subtask, we only have to reason about the worst-case scenario in which a single g_i changes. For the median, the worst-case scenario is that the median element changes its value from a to b or from b to a. For example:
>
> $$\text{Before: } g_1 = a, g_2 = a, \dots, g_{m/2} = a, g_{m/2+1} = b, \dots, \mathcal{M}' = f \circ \mathcal{M}_{h,\text{Lap}} g_{m-1} = b, g_m = b$$
>
> $$\text{After: } g_1 = a, g_2 = a, \dots, g_{m/2} = b, g_{m/2+1} = b, \dots, g_{m-1} = b, g_m = b$$
>
> In the above scenario, the median is $g_{m/2}$ and it has changed from a to b. Therefore, the global Δ_1 sensitivity of f'' is $b - a$.

c) Can you make the function f' and/or f'' differentially private for any function $f : \mathcal{X}^{\frac{n}{m}} \to [a, b]$? If yes, specify the noise distribution from which we have to sample to obtain an ϵ-DP private mechanism. If no, why not?

> We can obtain ϵ-DP by adding noise from the Laplace distribution with zero mean and variance $\frac{\Delta_1}{\epsilon}$ where $\Delta_1 = \frac{b-a}{m}$ for f' and $\Delta_1 = b - a$ for f''.

Problem 4: One of the fundamental properties of differential privacy is "group privacy" (see p.22): If mechanism $\mathcal{M}$ is ϵ-DP w.r.t $X \simeq X'$, then $\mathcal{M}$ is $(t\epsilon)$-DP w.r.t. changes of t instances/individuals.

Prove that group privacy holds when using the l_0 norm as the neighboring relation for vector data. That is: If mechanism $\mathcal{M}$ is ϵ-DP w.r.t. "$\simeq$", where $\boldsymbol{x} \simeq \boldsymbol{x}' \iff ||\boldsymbol{x} - \boldsymbol{x}'||_0 = 1$, then $\mathcal{M}$ is $(t\epsilon)$-DP w.r.t. "$\simeq_t$", where $\boldsymbol{x} \simeq_t \boldsymbol{x}' \iff ||\boldsymbol{x} - \boldsymbol{x}'||_0 = t$.

> Consider any pair of vectors $\boldsymbol{x} \simeq_t \boldsymbol{x}'$. The vector $\boldsymbol{x}'$ can be constructed from $\boldsymbol{x}$ by changing t of its dimensions, meaning that there must be a sequence of vectors $\left(\boldsymbol{x}^{(0)}, \dots, \boldsymbol{x}^{(t)}\right)$ with $\boldsymbol{x}^{(1)} = \boldsymbol{x}$, $\boldsymbol{x}^{(t)} = \boldsymbol{x}'$ and $\forall n : \boldsymbol{x}^{(n)} \simeq \boldsymbol{x}^{(n+1)}$.

Upload a single PDF file with your homework solution to Moodle by 02.02.2022, 11:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

We know that $\mathcal{M}$ is ϵ-DP w.r.t. $\simeq$, meaning that for any set of output values $\mathbb{Y}$

$$\mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}^{(n)}\right) \in Y\right] \leq e^{\epsilon} \cdot \mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}'^{(n+1)}\right) \in Y\right] \tag{2}$$

for all $n \in \{0, \ldots, t-1\}$.

Applying Equation 2 t times, starting from $\boldsymbol{x}^{(0)}$, shows that for any set of output values $\mathbb{Y}$

$$\begin{aligned}
\mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}\right) \in Y\right] &= \mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}^{(0)}\right) \in Y\right] \\
&\leq e^{\epsilon} \cdot \mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}^{(1)}\right) \in Y\right] \\
&\leq e^{2\epsilon} \cdot \mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}^{(2)}\right) \in Y\right] \\
&\cdots \\
&\leq e^{t\epsilon} \cdot \mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}^{(t)}\right) \in Y\right] = e^{t\epsilon} \cdot \mathbb{P}\left[\mathcal{M}_f\left(\boldsymbol{x}'\right) \in Y\right],
\end{aligned}$$

proving that $\mathcal{M}$ is $(t\epsilon)$-DP w.r.t. $\simeq_t$.