Security and Privacy in Machine Learning

Homework 5

Javad Hezareh
98101074

Spring 2023

Contents

1 Poisoned Samples Generation 2

2 Logistic Regression Extraction 2

3 JdBA 2

4 Another Private Algorithm 2

5 Private Histogram Publication 3

6 Randomized Response 4

7 Differential Private SGD 5

1 Poisoned Samples Generation

You can find the answer in the `poisoning_example_generation.ipynb` file.

---

2 Logistic Regression Extraction

(a) As we have access to the output of the neural net ($h^{(i)}$), we can easily build a set of linear equations and solve them to find our unknown variables which are W and b. If we have N queries, then we will have:

$$z^{(i)} = \begin{bmatrix} W & b \end{bmatrix} \begin{bmatrix} h^{(i)} \\ 1 \end{bmatrix}$$

$$\implies \begin{bmatrix} z^{(1)} & \dots & z^{(N)} \end{bmatrix} = \begin{bmatrix} W & b \end{bmatrix} \begin{bmatrix} h^{(i)} & \dots & h^{(N)} \\ 1 & \dots & 1 \end{bmatrix}$$

$$\implies Z = \hat{W} H$$

Therefore we only need to solve the last system of linear equations.

(b) If we assume that independent inputs will lead to independent output in $\mathcal{H}$ and also $c > n + 1$, then at least we need $n + 1$ queries to be able to solve the above linear equation and find a unique answer. Using more inputs will lead to more equation than variables and hence we might find more than one answer for $\hat{W}$ that satisfies the above equation. Therefore we won't be sure about the result.

(c) In this case we can first solve equations to find the logits z_i for each input by solving system $u_i = p_i \sum_{j=1}^{c} u_j$ where $u_i = exp(z_i)$. We have c variables and c equations, therefore we can find a unique solution for z_is. After that we can use the method of part (a) and (b).

---

3 JdBA

(a) In the formula (1) we will move in the direction of the Jacobian of the models output for the class labeled by the oracle. In other words in each point first we calculate the oracle label and then move in a direction that changes the probability of this class the most. This will help our model to adjust its changes to the oracle model change. Given a set of points, those directions that we have the most change in our model output for the oracle's label are candidate to investigate. Those direction that do not change our model output for oracle's label behave almost the same as oracle's model. Therefore first and third method to augment our dataset our effective and we can get most out of this methods by changing periodically the sign of λ. But second method is not that effective cause we move in the direction that increases the chance of predicting what we have been predicting. Therefore we don't use any information about the oracle and thus will not be effective.

(b) You can find the answer in the `JdDA.ipynb` file.

---

4 Another Private Algorithm

Let's consider two dataset X and Y that are only different in one element. Without loss of generality, let's say $Y_i = 1 \neq X_i = 0$. Therefore by the definition of f we have:

$$f(Y) = f(X) + 1$$

Now we use the definition of ϵ-differential privacy, we need to find the following ratio when we have observed output z:

$$\frac{\mathbb{P}(\tilde{f}(X) = z)}{\mathbb{P}(\tilde{f}(Y) = z)}$$

We know $\tilde{f}(X) = f(X) + z_1$ and using the relation between $f(X)$ and $f(Y)$ we have $\tilde{f}(Y) = f(Y) + z_2 = f(X) + 1 + z_2$. Therefore we have:

$$\frac{\mathbb{P}(\tilde{f}(X) = z)}{\mathbb{P}(\tilde{f}(Y) = z)} = \frac{\mathbb{P}(z_1 = f(X) - z)}{\mathbb{P}(z_2 = f(X) - z - 1)}, \qquad z_1, z_2 \sim U\left(-\frac{3}{\epsilon}, +\frac{3}{\epsilon}\right)$$

Now if we assume $\epsilon > 3$, then this ratio will be infinity and we don't have any privacy. The intuition is that when we have $\epsilon > 3$, the added noise will be less than 1 and hence the output of running f on two neighbor dataset will never be the same. If we assume $\epsilon < 3$, there are some observations z that above ratio is finite, but also exists some z that again the above ratio is infinity. For example when $f(X) = 5$, and $\epsilon = 0.1$ then this is possible to $z_1 = 0$ but impossible to see $z_2 = -1$. Therefore for $\epsilon < 3$ also we don't have differential privacy. So this mechanism by using uniform distribution is not ϵ-differential private for any value of ϵ.

---

5 Private Histogram Publication

First of all if we have $d_{ham}(x^n, y^n) = 0$, then all observation in x^n and y^n are the same and therefore $\hat{p}(x^n) = \hat{p}(y^n)$ which leads to $Lip_{1,d_{ham}}(\hat{p}) = 0$ that satisfies our inequality. If we have $d_{ham}(x^n, y^n) = 1$, then these two datasets only differ in one element. Let this element be the jth element of both, in other words we have $x_i = y_i$ for all $i \neq j$, and $x_j \neq y_j$. Let $x_j = v$ and $y_j = u$. Using the definition of $\hat{p}$ it is clear that:

$$\begin{cases} \hat{p}(x^n)_{n,j} = \hat{p}(y^n)_{n,j} & \forall j \neq v, u \\ \hat{p}(x^n)_{n,v} = \dfrac{1}{n} + \hat{p}(y^n)_{n,v} & \\ \hat{p}(x^n)_{n,u} + \dfrac{1}{n} = \hat{p}(y^n)_{n,u} & \end{cases}$$

Therefore we have $\|\hat{p}(x^n) - \hat{p}(y^n)\|_1 = \frac{2}{n}$. So for all x^n and y^n that only differ in one element the L-1 distance of the result is $2/n$. So we can say for all x^n and y^n that differ at most in one element, the maximum value for the L-1 distance of the result will be $2/n$ or in other words we have:

$$Lip_{1,d_{ham}}(\hat{p}) \leq \frac{2}{n}$$

Now for the second part of this question we replace Z with $\hat{p} + W$, therefore we have:

$$\begin{aligned}
\mathbb{E}\left[\|Z - p\|_2^2\right] &= \mathbb{E}\left[\|\hat{p} - p + W\|_2^2\right] \\
&= \mathbb{E}\left[(\hat{p} - p + W)^T(\hat{p} - p + W)\right] \\
&= \mathbb{E}\left[(\hat{p} - p)^T(\hat{p} - p) + (\hat{p} - p)^T W + W^T W\right] \\
&= \mathbb{E}\left[\|\hat{p} - p\|_2^2\right] + \mathbb{E}\left[(\hat{p} - p)^T W\right] + \mathbb{E}\left[\|W\|_2^2\right]
\end{aligned}$$

The second term will be 0 because $(\hat{p}-p)$ and W are independent and as we have $\mathbb{E}[W] = \mathbf{0}$ we will have $\mathbb{E}[(\hat{p} - p)]\mathbb{E}[W] = \mathbf{0}$. For the last term we have:

$$\mathbb{E}\left[\|W\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^{k} W_i^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[W_i^2\right] = \frac{8k}{n^2\epsilon^2}$$

The last equality comes from this fact that $\mathbb{E}[W_i^2] = \text{Var}(W_i) + \mathbb{E}[W_i]^2$ and as W_i comes from Laplace with mean 0 and parameter $2/n\epsilon$ we have $\mathbb{E}[W_i^2] = \text{Var}(W_i) = 8/n^2\epsilon^2$. For the first part of decomposition we have:

$$\mathbb{E}\left(\|\hat{p} - p\|_2^2\right) = \mathbb{E}\left[\sum_{i=1}^{k} (\hat{p}_i - p_i)^2\right] \quad = \sum_{i=1}^{k} \mathbb{E}\left[(\hat{p}_i - p_i)^2\right] = \sum_{i=1}^{k} \text{Var}(\hat{p}_i) + \mathbb{E}[\hat{p}_i - p_i]^2$$

The last term is again 0 because $\mathbb{E}[\hat{p}_i] = p_i$. To calculate the variance of each $\hat{p}_i$ we have:

$$\text{Var}(\hat{p}_i) = \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{(X_j=i)}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(\mathbb{1}_{(X_j=i)}) = \frac{p_i(1-p_i)}{n}$$

Where the last equality comes from this fact that each $\mathbb{1}_{(X_j=i)}$ is a binary random variable with probability p_i. Therefore we have:

$$\mathbb{E}\left(\|\hat{p} - p\|_2^2\right) = \frac{1}{n}\sum_{i=1} p_i(1-p_i)$$

Now combining our results we have:

$$\begin{aligned}
\mathbb{E}\left[\|Z - p\|_2^2\right] &= \mathbb{E}\left[\|\hat{p} - p\|_2^2\right] + \mathbb{E}\left[(\hat{p} - p)^T W\right] + \mathbb{E}\left[\|W\|_2^2\right] \\
&= \frac{1}{n}\sum_{i=1} p_i(1-p_i) + 0 + \frac{8k}{n^2\epsilon^2} \\
&= \frac{1}{n}\left[\sum_{i=1}^{k} p_i - \sum_{i=1}^{k} p_i^2\right] + \frac{8k}{n^2\epsilon^2} \\
&= \frac{1}{n} - \frac{1}{n}\sum_{i=1}^{k} p_i^2 + \frac{8k}{n^2\epsilon^2} \\
&\leq \frac{1}{n} + \frac{8k}{n^2\epsilon^2}
\end{aligned}$$

---

6 Randomized Response

(a) We can use the expectation of Y_is. For each of this variables we have:

$$\mathbb{E}[Y_i] = \left(\frac{1}{2} + \alpha\right) X_i + \left(\frac{1}{2} - \alpha\right)(1 - X_i) = \alpha(2X_i - 1) + \frac{1}{2}$$

Therefore for the expectation of the Y_is resulting from the students we have:

$$\mathbb{E}[Y] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = 2\alpha \sum_{i=1}^{n} X_i + n\left(\frac{1}{2} - \alpha\right)$$

$$\implies \sum_{i=1}^{n} X_i = \frac{\mathbb{E}[Y] - n\left(\frac{1}{2} - \alpha\right)}{2\alpha}$$

Now if we want to estimate the value $p = (\sum_{i=1}^{n} X_i)/n$ which is the percentage of students who have cheated, we can use the following unbiased estimator:

$$\hat{p} = \frac{1}{2n\alpha} \sum_{i=1}^{n} \left[Y_i + \alpha - \frac{1}{2}\right]$$

One using the $\mathbb{E}[Y_i]$ can easily verify that $\mathbb{E}[\hat{p}]$ over the randomness of Y_is is equal to the p that means $\hat{p}$ is an unbiased estimator.

(b) Let $\alpha = 0$, in this case we will lose all the information about the real answer of the students because $\mathbb{E}[Y_i] = 1/2$ and its estimator will result in $n/2$ for the total number of cheats. Therefore by using $\alpha = 0$ we will have perfect privacy and no information will leak from the actual value of the students and yet lose the accuracy and predict $n/2$ for the number of cheats.

If we use $\alpha = 1/2$ then we will have $\mathbb{E}[Y_i] = X_i$ and our estimator will result in the actual value for the total number of cheats. We will have $\sum_i X_i = \sum_i Y_i$ and hence we have information leakage from students and no privacy. On the other hand we have perfect accuracy.

Therefore we can conclude that by increasing α from 0 to 1/2, we will be more accurate and have less privacy. If we want to have both accuracy and privacy, then we need to find a best value for α in the middle of interval $(0, 1/2)$.

(c) In order to use the Chebyshev's inequality first we need to find the variance of our estimator. For that we have:

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{1}{2n\alpha} \sum_{i=1}^{n} \left[Y_i + \alpha - \frac{1}{2}\right]\right) = \frac{1}{4n^2\alpha^2} \mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right)$$

As we now Y_is are i.i.d. then we can use this rule that the variance of the sum is equal to the sum of the variance. Therefore we have:

$$\mathrm{Var}(\hat{p}) = \frac{1}{4n^2\alpha^2} \sum_{i=1}^{n} \mathrm{Var}(Y_i)$$

And because each Y_i is a Bernoulli random variable, its variance will be at most 1/4. So:

$$\mathrm{Var}(\hat{p}) \leq \frac{1}{4n^2\alpha^2} \times \frac{n}{4} = \frac{1}{16n\alpha^2}$$

Now using the Chebyshev's inequality for the error of the estimator we have:

$$\mathbb{P}\left(|\hat{p} - p| < k\right) \geq 1 - \frac{\mathrm{Var}(\hat{p})}{k^2}$$

In order to bound the error we can set $k = \sqrt{\mathrm{Var}(\hat{p})}$ to have

$$\mathbb{P}\left(|\hat{p} - p| < \sqrt{\mathrm{Var}(\hat{p})}\right) \geq 1$$

Therefore we have:

$$|\hat{p} - p| < \frac{1}{4\alpha\sqrt{n}}$$

Now if we want to have error γ, then we need to have:

$$\gamma = \frac{1}{4\alpha\sqrt{n}} \implies n = \frac{1}{16\alpha^2\gamma^2}$$

Therefore we need $O\left(\frac{1}{\alpha^2\gamma^2}\right)$ students if we want to have error γ.

---

7 Differential Private SGD

You can find the answer in the `differentially_private_SGD.ipynb`