In the Name of God

Data Privacy (25100)

Problem Set 02

Fall 2023

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. M.H. Yassaee

Soft Dedline: 4 Bahman 1402 - 23:55
Hard Dedline: 8 Bahman 1402 - 23:55

- Delivering Assignment with LaTeX has 15% bonus mark.

1 Differentially Private Logistic Regression

Your goal in this question is to give an $(\varepsilon, 0)$-differentially private algorithm for logistic regression, by implementing the private gradient descent algorithm using Laplace noise rather than Gaussian noise, and applying the standard (rather than advanced) composition theorem. In particular, consider linear classifiers $f_\theta : [0,1]^d \to [-1,1]$ of the form $f_\theta(x) = \sum_{i=1}^d \theta_i x_i + \theta_0$ for $x \in [0,1]^d$, and let the loss of predicting label $f_\theta(x)$ instead of the true label y be given by

$$l\left(f_\theta(x), y\right) = \log\left(1 + e^{-y f_\theta(x)}\right).$$

For a dataset X consisting of labelled samples $(x_1, y_1), \ldots, (x_n, y_n) \in [0,1]^d \times \{-1, +1\}$, this gives the empirical loss function

$$L(\theta, X) = \frac{1}{n} \sum_{i=1}^{n} l\left(f_\theta\left(x_i\right), y_i\right)$$

(a) Let $\nabla L(\theta, X)$ be the gradient of L with respect to θ. Given an upper bound on the ℓ_1 sensitivity of $\nabla L(\theta, X)$, i.e., on

$$\max_{X \sim X'} \|\nabla L(\theta, X) - \nabla L(\theta, X)\|_1,$$

with the maximum taken over neighbouring datasets X, X' differing in a single data point.

(b) Show that, for an appropriate value of T, the ε-differentially private algorithm from the previous subquestion outputs $\theta^{\text{priv}} = \frac{1}{T}\sum_{t=0}^{T-1} \theta^t$ such that

$$\mathbb{E}L\left(\theta^{\text{priv}}, X\right) - \min_{\theta \in B_2^{d+1}(R)} L(\theta, X) \leq \alpha.$$

as long as

$$n \geq \frac{KR^2(d+1)^2}{\varepsilon\alpha^2}$$

for a sufficiently large constant K.Be specific about how you choose the value of T.

2 Graph Privacy and Different Types of Sensitivity

For $n \geq 2$, let $\mathcal{G}$ = the set of undirected graphs (without self-loops) on vertex set $V = \{1, \ldots, n\}$, and for $G, G' \in \mathcal{G}$, define $G \sim G'$ if there is a vertex $v \in V$ such that the only differences between G and G' involve edges incident to the vertex v. (That is, we are considering node-level privacy.) For an integer $d \in [2, n-1]$, let $\mathcal{H} \subseteq \mathcal{G}$ denote the set of graphs of degree at most d. Define $q : \mathcal{G} \to \mathbb{N}$ by taking $q(G)$ to be the number of isolated (i.e., degree 0) nodes in G. Calculate the following measures of sensitivity of q :

(a) The global sensitivity: GS_q.

(b) The minimum local sensitivity: $\min_{G \in \mathcal{G}} \mathrm{LS}_q(G)$. (For example some of the approaches are Privately Bounding Local Sensitivity, Propose-Test-Release, and Smooth Sensitivity aim to add noise that's not too much larger than the local sensitivity, which can sometimes be much smaller than the global sensitivity. It's not always possible to do this while preserving DP, but local sensitivity calculations like here and below help give a sense of how much we can gain from such methods.)

(c) The maximum local sensitivity on $\mathcal{H}$: $\max_{G \in \mathcal{H}} \mathrm{LS}_q(G)$.[1]

(d) The restricted sensitivity on $\mathcal{H}$: $\mathrm{RS}_q^{\mathcal{H}} = \max_{G, G' \in \mathcal{H}, G \sim G'} |q(G) - q(G')|$[2] (The material we surveyed on graph privacy and restricted sensitivity tells us that there is a mechanism that is ϵ-DP on all of $\mathcal{G}$, but only adds noise proportional to $\mathrm{RS}_q^{\mathcal{H}}$ for graphs in $\mathcal{H}$.)

3 Lipschitz Extensions

For each of the following sets $\mathcal{G}$ of datasets and neighbor relations $\sim$, hypotheses $\mathcal{H} \subseteq \mathcal{G}$, and functions $f : \mathcal{G} \to \mathbb{R}$, calculate (i) the global sensitivity of f (denoted GS_f or ∂f), (ii) the minimum local sensitivity of f, i.e. $\min_{x \in \mathcal{G}} \mathrm{LS}_f(x)$, and (iii) the restricted sensitivity of f (denoted $\partial_{\mathcal{H}} f$ or $\mathrm{RS}_f^{\mathcal{H}}$). For Part 1a, also describe an explicit Lipschitz extension of f from $\mathcal{H}$ to all of $\mathcal{G}$.

(a) $\mathcal{G} = \mathbb{R}^n$ where $x \sim x'$ if x and x' differ on one row, $\mathcal{H} = [a, b]^n$ for real numbers $a \leq b$, and $f(x) = (1/n) \sum_{i=1}^n x_i$.

(b) $\mathcal{G} = \mathbb{R}^n$ where $x \sim x'$ if x and x' differ on one row, $\mathcal{H} = [a, b]^n$ for real numbers $a \leq b$, and $f(x) = \mathrm{median}\,(x_1, \ldots, x_n)$.

(c) $\mathcal{G}$ = the set of undirected graphs (without self-loops) on vertex set $\{1, \ldots, n\}$ where $x \sim x'$ if x and x' are identical except for the neighborhood of a single vertex (i.e. node privacy), $\mathcal{H}$ = the set of graphs in $\mathcal{G}$ in which every vertex has degree at most d for a parameter $2 \leq d \leq n-1$, and $f(x)$ = the number of isolated (i.e. degree 0) vertices in x.

[1] The answer differs from and motivates why we use restricted sensitivity.

[2] The general definition of restricted sensitivity is a bit more involved, and also considers datasets G and G' that are not neighbors, but this simplified version is equivalent in the special case of $\mathcal{H}$ and considered here.