NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Mitigating the Inference of Sensitive Training Data with Differential Privacy

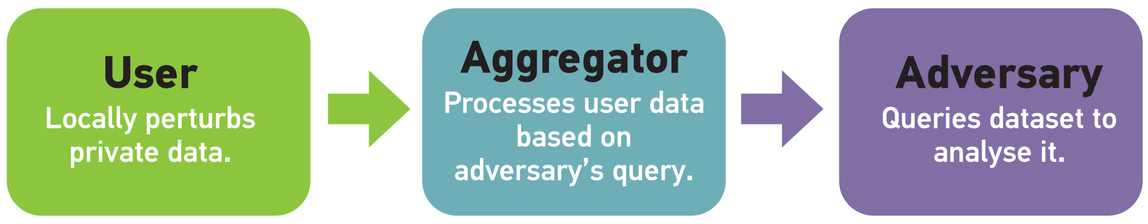
Local Differential Privacy in Machine Learning Algorithms

Student: Hans Albert Lianto

Supervisor: Asst Prof Zhao Jun

Background and Local Differential Privacy

Local differential privacy (LDP) is a rigorous quantitative standard for user data protection that protects individual tuples of data, protecting it from aggregators who process the data and return query results to adversaries.



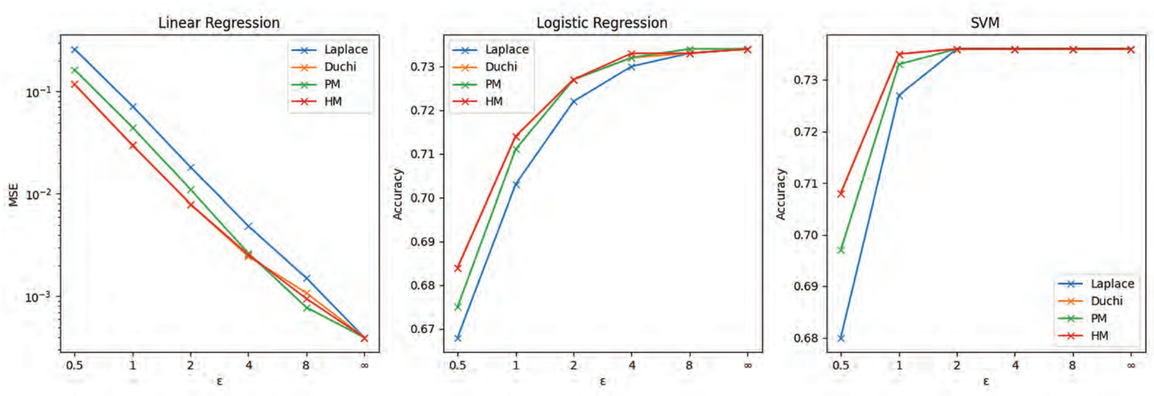
An algorithm f satisfies ε-LDP if for any two tuples t and t' in f's domain, and for any output t*:

$$Pr[f(t)=t^*] \leq e^{\varepsilon} \times Pr[f(t')=t^*]$$

where parameter ε controls the privacy-utility tradeoff. Lower ε corresponds to higher privacy.

Innovations and Experiments

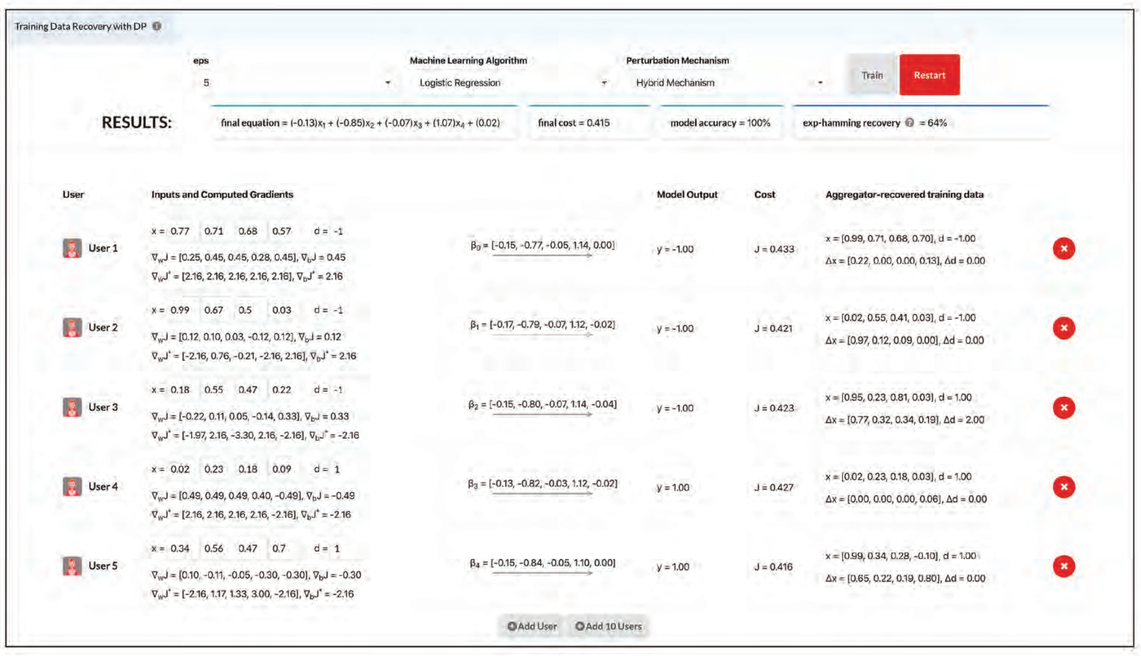
LDP algorithms PM (Piecewise Mechanism) and HM (Hybrid Mechanism) are proposed in this project, offering smaller worse case variance than current state-of-the-art methods (i.e. Duchi's method and Laplace). These algorithms are applied to user gradients during the stochastic gradient descent process of 3 popular machine learning algorithms: (1) linear regression, (2) logistic regression and (3) SVM on a 'mexico2000' dataset with 4,000,000 tuples.



The results that our LDP algorithms outperform or match current LDP methods, resulting in models with similar accuracy as those trained without LDP. Models trained with LDP are also less vulnerable to model inversion attacks (i.e. it is possible for untrusted aggregators to fully recover training data during SGD without LDP).

Demonstration Using React.js User Interface

A web user interface deployed on Heroku is created to demonstrate LDP in these machine learning algorithms, allowing learners for a more visual understanding of LDP. It allows the modification of several parameters, including ε, the LDP algorithm (i.e. PM, HM), and the machine learning algorithm used (i.e. SVM). The interface display is shown below:



It can be shown through a custom measure called exp-hamming recovery, that the higher the privacy, the less this recovery measure becomes- hence the demo shows that it is harder for an untrusted aggregator to recover user data with more privacy.

Improvements

PM and HM could be further enhanced to perturb gradients in more complex machine learning algorithms such as deep neural networks and CNN and RNN architectures.