Exploring Differentially Private Data with PATE

Anya Khatri ak5379

What I want to do

For my final project, I want to investigate how differential privacy used in the PATE framework can be applied to the process of generating differentially private labels for training models and investigate whether a model trained on such data can maintain performance and fairness. In many machine learning applications, data privacy is a big concern. Even when data is anonymized, models can unintentionally leak private information through inference attacks or overfitting.

PATE is a technique for achieving differential privacy by training multiple teacher models on disjoint subsets of private data. These teachers independently label a set of public or unlabeled examples and their predictions are then aggregated using a noisy voting mechanism for privacy. The final output labels are used to train a student model which does not directly access any private training data. Each teacher only sees a fraction of the private data and the final labels are concealed by noise.

I am interested in this topic because I have noticed an increasing use of synthetic data in lots of industries with stricter compliance and regulations such as finance and healthcare and I am interested in how the PATE technique can be implemented with synthetic data to generate the best performance. I came across this paper entitled *Scalable Private Learning with PATE[1]* and I thought it would be great to dive deeper into a topic I am not as familiar with but am interested in learning more about.

Research Questions

This project will address the following key research questions:

RQ1: Can PATE-generated labels allow models to achieve high utility and fairness?

RQ2: How does the level of noise impact model performance and fairness?

RQ3: What are the tradeoffs between privacy and utility in the PATE framework?

Experimental Setup

I will conduct a small-scale experiment using the MNIST dataset and split the dataset into three distinct parts: a private training set for training teacher models, a public or unlabeled set for synthetic labeling using the PATE mechanism, and a held-out real test set that will be used to evaluate the final model.

Next, I will train a group of teacher models on the private training data. Each teacher will be trained on a disjoint subset of the private dataset. These models will be relatively simple classifiers and I plan to use around ten teacher models in total. Instead of directly using their raw predictions, I will apply noisy aggregation by adding randomized noise to the vote counts of each class label before selecting the majority label. This noisy voting process ensures differential privacy by obscuring the contribution of any individual teacher model, and by tuning the amount of added noise, I can simulate different levels of privacy protection, defined by the privacy budget ε.

After generating noisy synthetic labels, I will use them to train a student model. The student model is only exposed to the public inputs and the differentially private labels. It never accesses the original private training data and once trained, the student model will be evaluated using standard classification metrics such as accuracy, precision, recall, and F1 score. For comparison, I will replicate the same evaluation for real data without privacy constraints and another trained using non-private synthetic labels.

As an extension, I am interested in exploring synthetic data generation techniques like PATE-GAN [2] to create generative models to produce synthetic datasets with privacy guarantees and compare these metrics to the previous experiment run on the MNIST dataset. To build up on this research, I would also be interested in reverse engineering a Membership Inference Attack to see if the attacker can tell if a particular record was used in the teacher models. We could use PrivacyMeter [3] to automatically train shadow models and a binary attack classifier to estimate how well an adversary can determine membership. We can collect metrics like attack accuracy and AUC to see which models may be leaking.

Citation

[1] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., & Erlingsson, Ú. (2018). *Scalable Private Learning with PATE*. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)

[2] Jordon, J., Yoon, J., & van der Schaar, M. (2018). *PATE-GAN: Generating synthetic data with differential privacy guarantees*. arXiv preprint arXiv:1806.03384.

[3] Murakonda, S. K., & Shokri, R. (2020). *ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning*. Data Privacy and Trustworthy ML Research Lab, National University of Singapore https://arxiv.org/abs/2007.09339