Privacy Preserving Data Mining

Qingyang Zhao
Advisor: Shao-Lun Huang, Lin Zhang
Tsinghua University, University of California, Berkeley

TBSI

Abstract

The central problem of privacy preserving data mining is to solve privacy-utility tradeoff. In this research, we first formulate privacy-utility tradeoff into a non-convex optimization problem by introducing maximal correlation as metric. To efficiently solve this, we further propose a DNN model, called 'Privacy-Net', with a loss function named H-score. Along this optimization, we are able to perform privacy preserving disturbance on real dataset. The experiments on MNIST provides with interpretable and backup results for the superiority of our algorithm.

Introduction

Consider binary classification with RGB images of apples and bananas. APPLEs are labeled as (apple, red) and BANANAs are (banana, yellow). Define color information as a sensitive attribute, our proposed Privacy-Net is optimized to make targeted disturbance to raw dataset. The disturbed data can reveal little information about the private attribute, namely, in this problem, the color. In this way, we call our problem as privacy preserving data mining.

Theoretical Framework

$$f(X)$$
$$/$$
$$Y \;\; -> \;\; X \;\; -> \;\; Z$$

Y: color labels X: apples and bananas Z: apples and bananas,
f(X) a feature obtained by a DNN model.

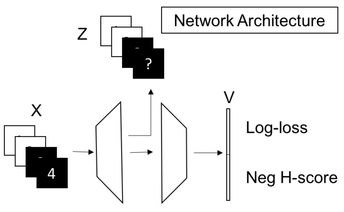
$$\max_{P_{Z|X}} \max_{h_1} E[f(X)h_1(Z)] - \gamma \max_{g,h_2} E[g(Y)h_2(Z)]$$

s.t. h_1, g, h_2 are all normalized

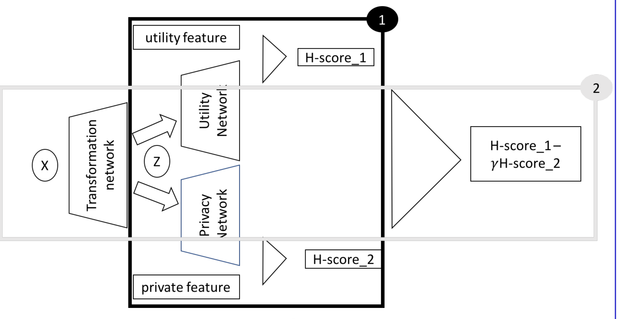
H-score:

$$H(s,v) \triangleq \frac{1}{2}\|\mathbf{B}\|_{\mathrm{F}}^2 - \frac{1}{2}\|\mathbf{B} - \mathbf{\Xi}^Y(\mathbf{\Xi}^X)^{\mathrm{T}}\|_{\mathrm{F}}^2$$
$$= \mathbb{E}_{P_{XY}}[\tilde{s}(X)\,\tilde{v}(Y)] - \frac{1}{2}\operatorname{tr}\left(\mathbf{\Lambda}_{\tilde{s}(X)}\mathbf{\Lambda}_{\tilde{v}(Y)}\right).$$

DNN Architectures

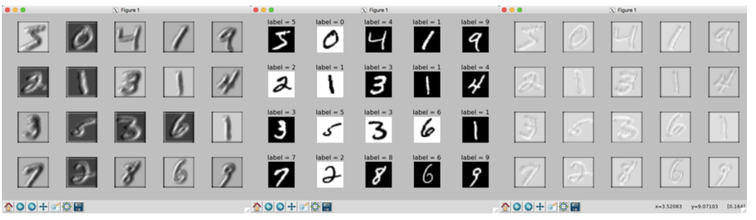


Log-loss & H-score Information Subtraction Comparison

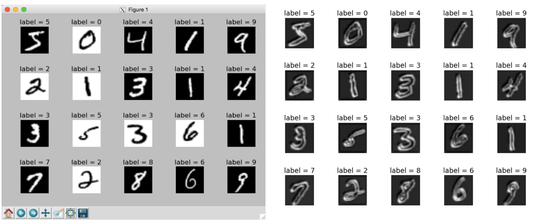


Adversarial Training Privacy Net

Experimental Results



Information subtraction on reversed MNIST dataset



Privacy-Net results on reversed MNIST dataset

Future Work

- **Implement Privacy-Net on Multi-Label classification tasks (MSCOCO):** Split multi-labels into public part and private part. Optimize the network, ideally, for each image, we generate the transformed feature, representing for the public information only.
- **Project onto original image space.** The transformed features are irrelevant to private labels. If considered as latent variables, a decoder/generator can map features back into real images. Candidate approaches would be Autoencoder or generative model, such as VAE or GANs.
- **Other Applications**, such as auto PS, or Natural Language privacy preserving, features simply degenerated to word embeddings.

References

[1] Wang, Hao, and Flavio P. Calmon. "An estimation-theoretic view of privacy." *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*. IEEE, 2017.
[2] Asoodeh, Shahab. Information and Estimation Theoretic Approaches to Data Privacy. Diss. 2017.
Makhdoumi, Ali, et al. "From the information bottleneck to the privacy funnel." *Information Theory Workshop (ITW), 2014 IEEE*. IEEE, 2014.
[3] Belghazi, Mohamed Ishmael, et al. "Mutual Information Neural Estimation." *International Conference on Machine Learning*. 2018.

Acknowledgement

The author would like to acknowledge my advisor Shao-Lun Huang and Lin Zhang.

Contact: Qingyang Zhao qingyang_zhao@berkeley.edu