A Differentially Private GAN to Generate Demographics

Pascal Jutras-Dubé

Introduction

The problem of privacy-preserving data analysis has a long history spanning multiple disciplines. In particular, large and representative datasets may be crowdsourced and contain sensitive information. For example, in the medical field, individual data from clinical studies may contain sensitive information; an individual could be affected if their medical data were shared.

One question arises: does sharing the probability distribution of the data instead of the original sensitive data guarantee confidentiality? Ideally, with the generative distribution in hand, it would be possible to sample a synthetic population that respects the statistical properties of the true population. Generative adversarial network (GAN) has recently attracted intensive research interests due to its excellent empirical performance as a generative model. One common issue in GANs is that the density of the learned generative distribution could concentrate on the training data points, meaning that they can easily remember training samples due to the high model complexity of deep networks. This becomes a major concern when GANs are applied to private or sensitive data.

The goal of the proposed academic project is to replicate the differentially private WGAN framework proposed by Xu et al. [2019] and Xie et al. [2018] and to study its efficiency on a different, non-public dataset. To do so, I have access to Maryland demographics from the US Census. Pumas (Public Use Microdata Area), are geographic units used by the U.S. census to provide statistical and demographic information.

To generate discrete variables

Since the generator *G* is trained by the error signal from the discriminator *D* via backpropagation, the original GAN can only learn to approximate discrete features with continuous values. In medGAN [Choi et al., 2017], the authors propose to learn an embedding to deal with the generation of discrete variables. The idea is to pretrain an autoencoder to learn a continuous representation of the discrete variables that can be applied to decode the continuous output of *G*.

With the pretrained autoencoder, we can allow the GAN to generate a representation in the space of the output of the encoder *Enc*, rather than generating in the original data space. Then the pretrained decoder *Dec* can pick up the right signals from *G(z)* to convert it *Dec(G(z))*. The discriminator *D* is trained to determine whether the given input is a synthetic sample *Dec(G(z))* or a real sample *x*.

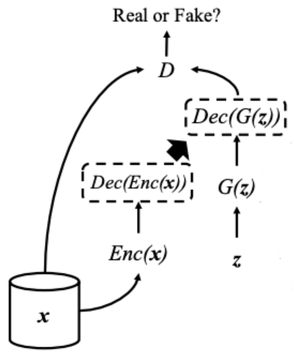

Figure 1: Architecture of medGAN.

Differential privacy

Differential confidentiality [Dwork and Roth, 2014] refers to the process by which a mechanism modifies a set of sensitive data by means of randomization.

A randomized mechanism $\mathscr{M} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ satisfies (ϵ, δ) - differential privacy if for any two adjacent (differing in one entry) databases $d_1, d_2 \in \mathscr{D}$ and for any subset of output $S \subseteq \mathscr{R}$ it holds that

$$P[\mathscr{M}(d_1) \in S] \leq e^{\epsilon} P[\mathscr{M}(d_2) \in S] + \delta$$

if $\delta = 0$ we say that $\mathscr{M}$ is ϵ - *DP*. Intuitively, ϵ quantifies the worst case of loss of confidentiality: an attacker cannot know more than e^{ϵ} times what he might have expected to know if one of the entry had been removed from the dataset. When $\delta > 0$, the mechanism satisfies ϵ - DP with probability $1 - \delta$.

Differentially private GAN

Abadi et al. [2016] present a modified SGD algorithm that respects differential privacy by controlling the influence of the training data during the training process. At each step of the SGD, the idea is to compute the gradient for a random subset of examples, clip the norm of each gradient, compute the average, add gaussian noise in order to protect privacy, and take a step in the opposite direction of this average noisy gradient.

To keep track of a bound on the moments of the privacy loss random variable, Abadi et al. [2016] propose the *moments accountant* method which uses the composability properties of differential privacy. This method needs the gradients to be bounded. In the WGAN [Arjovsky et al., 2017], framework, weight clipping is a way to enforce a Lipschitz constraint. In [Xu et al., 2019], they prove that the gradient can be bounded at same time, which avoids unnecessary distortion of the gradient. This not only keeps the loss function with Lipschitz property but also provides a sufficient privacy guarantee.

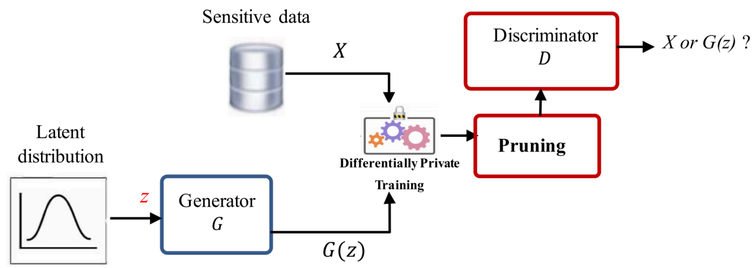

Figure 2: Training a Gan with differential privacy.

Evaluation metrics

- **Dimension-wise prediction:** We choose one dimension to be the label and train a model on the synthetic data to predict it. The closer the performance is to the one of a model train on the real data, the better the quality of the GAN. It indirectly measures how well the model captures the inter-dimensional relationships of the real samples.
- **Attribute disclosure:** It occurs when attackers can derive additional attributes about *x* based on a subset of attributes they already know about *x*. We sample *m* examples of the training set. For each example *r*, we suppose that the attacker knows *s* random attributes. The attacker tries to find the unknown attributes by finding the *k* nearest neighbours of *r* in the synthetic data set and make them vote to estimate the unknown attributes. To evaluate the classification, we can compute the accuracy of the attributes found compared to the total number of the attributes we tried to guess.

Results and challenges

Generation performance

The first challenge is to succeed in generating data without Differential privacy. The GAN framework may not be the right one for this type of data. A first simple implementation can generate examples that seems to respect some statistics on the marginal distributions such as the mean and the range. However, the dimension-wise prediction performance is not satisfying for most of the dimensions in comparison to the predictive power of the training set.

Privacy considerations

A row is a low dimensional vector of count variables (a ten of features) so it is probable for a training example to also appear in a synthetic dataset. For a training set of approximately 10000 samples and a synthetic dataset of the same size, there are a ten of rows in the intersection of the sets*.

For the attribute disclosure, it turns around 25% for the synthetic dataset while being much higher for the original training set (around 60%).

Finally, since the current « vanilla » GAN architecture hasn't an impressive performance, adding a satisfying level of noise in term of privacy budget ruins the utility.

* Should be mean over multiple simulations

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Oct 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, Proceedings of the 34th International Conference on Machine Learning , volume 70 of Proceedings of Machine Learning Research , pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html. C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy, volume 9. 2014.

E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. arXiv e-prints , art. arXiv:1703.06490, Mar. 2017.

C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy, volume 9. 2014.

L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. Differentially Private Generative Adversarial Network. arXiv e-prints , art. arXiv:1802.06739, Feb. 2018.

C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren. Ganobfuscator: Mitigating information leakage under gan via differential privacy. IEEE Transactions on Information Forensics and Security, PP:1–1, 02 2019. doi: 10.1109/TIFS.2019.2897874.