Differentially Private Synthetic Microdata

Jingchen (Monika) Hu

Vassar College

Data Confidentiality

Outline

Outline

Recap of differential privacy

- Definitions: database, query, output, sensitivity, privacy budge, and added noise
- Implications of key terms in differential privacy: the relationship between sensitivity (Δf), privacy budget (ϵ), and added noise
- The Laplace Mechanism
  - adds random noise according to $\epsilon-$differential privacy guarantee
  - the noise is drawn from a Laplace distribution centered at 0, with scale $\frac{\Delta f}{\epsilon}$
- DP properties: example queries to the confidential CE databse

Recap of differntially private synthetic tabular data

- Based on Dirichlet-multinomial conjugate models
- To generate a differentially private synthetic count vector $\mathbf{y}^*$ given $y.^* = y.$ (the total sum is fixed):

1. Sample $\boldsymbol{\theta}^*$ from

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \mathrm{Dirichlet}(\mathbf{y} + \boldsymbol{\alpha}), \tag{1}$$

where $\min(\alpha_i) \geq \frac{y.^*}{\exp(\epsilon)-1}$.

2. Sample $\mathbf{y}^*$ from

$$\mathbf{y}^* \mid \boldsymbol{\theta}^* \sim \mathrm{Multinomial}(y.^*; \boldsymbol{\theta}^*), \tag{2}$$

and the generated count vector $\mathbf{y}^*$ satisfies $\epsilon-$differential privacy.

Differentially private synthetic microdata

- Respondent-level data: the focus of our synthetic data approach
- Synthetic data has certain levels of privacy protection
  - Identification disclosure and IR risks
  - Attribute disclosure and AR risks

Differentially private synthetic microdata

- Respondent-level data: the focus of our synthetic data approach
- Synthetic data has certain levels of privacy protection
  - Identification disclosure and IR risks
  - Attribute disclosure and AR risks
- However the privacy protection does not satisfy $\epsilon-$differential privacy
  - Original definition:

$$\left| \ln \left(\frac{Pr[\mathcal{M}(\mathbf{x}) \in \mathcal{S}]}{Pr[\mathcal{M}(\mathbf{y}) \in \mathcal{S}]} \right) \right| \leq \epsilon \tag{3}$$

  - Updated definition in the context of tabular data

$$\left| \ln \left(\frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\theta})} \right) \right| \leq \epsilon \tag{4}$$

Outline

- The Exponential Mechanism (McSherry and Talwar, 2007)
  - it turns a non-private mechanism (e.g. a Bayesian synthesis model) into a private mechanism (e.g. a Bayesian synthesis model satisfying differential privacy)

Outline

- The Exponential Mechanism (McSherry and Talwar, 2007)
  - it turns a non-private mechanism (e.g. a Bayesian synthesis model) into a private mechanism (e.g. a Bayesian synthesis model satisfying differential privacy)

- Three mechanisms based on the Exponential Mechanism
  - pMSE Mechanism (Snoke and Slavovic, 2018)
  - Posterior Mechanism (Dimitrakakis et al., 2017)
  - Pseudo Posteror Mechanism (Savitsky et al., 2019)

Outline

The background

- Dwork et al. (2006) and Nissim et al. (2007) show that any function of an $\epsilon-$differentially private algorithm also satisfies $\epsilon-$differential privacy

The background

- Dwork et al. (2006) and Nissim et al. (2007) show that any function of an $\epsilon-$differentially private algorithm also satisfies $\epsilon-$differential privacy

- In synthetic data generation: if parameters satisfy $\epsilon-$differential privacy, synthetic data generated based on the $\epsilon-$differentially private parameters are also differentially private

$$\hat{\theta} \sim g(\theta), \quad (5)$$
$$\mathbf{x}^* \sim f(\mathbf{x} \mid \hat{\theta}), \quad (6)$$

where $\pi(\cdot)$ is the mechanism that makes parameter draws $\hat{\theta}$ differentially private, and $f(\cdot)$ is the sampling model.

- Question: how to make Equation (5) happen?

The EM

- Proposed by McSherry and Talwar (2007)
- The EM inputs non-private parameters θ and generates private parameters $\hat{\theta}$ (i.e. satisfying differential privacy)
- In the context of generating differentially private synthetic data from a Bayesian perspective, we follow the general framework proposed by Zhang et al. (2016)

The EM cont'd

The Exponential Mechanism generates private parameters $\hat{\theta}$ from:

$$\hat{\theta} \propto \exp\left(\frac{\epsilon\, u(\mathbf{x}, \theta)}{2\Delta_u}\right) \pi(\theta) \tag{7}$$

- ϵ is the privacy budget
- $u(\mathbf{x}, \theta)$ is the utility function
- Δ_u is the sensitivity of the utility function
- $\pi(\theta)$ is the base distribution to ensure proper density function (one can think of $\pi(\theta)$ as the prior distribution for θ)

The utility function

- In the DP overview lecture:
  - Δ_f is defined as the ℓ_1-sensitivity of a query function f
  - which is the maximum change in the in fuction f on $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$ and differ by a single observation (i.e. $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}, \delta(\mathbf{x}, \mathbf{y}) = 1$)

The utility function

- In the DP overview lecture:
  - Δ_f is defined as the ℓ_1-sensitivity of a query function f
  - which is the maximum change in the in fuction f on $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$ and differ by a single observation (i.e. $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}, \delta(\mathbf{x}, \mathbf{y}) = 1$)
- Here:
  - Δ_u in the Exponential Mechanism is the global sensitivity
  - defined as the maxium change in the utility function $u(\mathbf{x}, \theta)$ for $\mathbf{x}$ and $\mathbf{y}$
  - where $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ and differ by a single observation (i.e. $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n, \delta(\mathbf{x}, \mathbf{y}) = 1$)
- Formally:

$$\Delta_u = sup_{\theta \in \Theta} sup_{\mathbf{x}, \mathbf{y}: \delta(\mathbf{x}, \mathbf{y}) = 1} |u(\mathbf{x}, \theta) - u(\mathbf{y}, \theta)| \tag{8}$$

Summary

- We wish to generate private θ, from which we can ultimately generate and release private $\mathbf{x}$
- The Exponential Mechanism defines a distribution from which private samples, $\hat{\theta}$ can be simulated
- The keys to the Exponential Mechanism are the utility function $u(\mathbf{x}, \theta)$ and its global sensitivity Δ_u

Summary

- We wish to generate private θ, from which we can ultimately generate and release private $\mathbf{x}$
- The Exponential Mechanism defines a distribution from which private samples, $\hat{\theta}$ can be simulated
- The keys to the Exponential Mechanism are the utility function $u(\mathbf{x}, \theta)$ and its global sensitivity Δ_u
- Next we introduce three mechanisms based on EM to generate differentially private synthetic microdata

Outline

The pMSE Mechanism

- Snoke and Slavovic (2018)
- Based on the propensity score measure we have learned:
  - stack up the original dataset and the synthetic dataset resulting in a merged dataset of size $2n$
  - and use a classification algorithm (e.g. logistic regression) to predict whether an observation belongs to the original dataset or the synthetic dataset
  - return a summary statistic U_p, which measures overall how close the predicted probability of each observation $\hat{p}_i$ is to $\frac{1}{2}$:

$$U_p = \frac{1}{2n} \sum_{i=1}^{2n} (\hat{p}_i - \frac{1}{2})^2. \tag{9}$$

- High level of similarity between the original and the synthetic datasets results in $U_p \approx 0$; low level of similarity results in $U_p \approx \frac{1}{4}$

The pMSE Mechanism cont'd

- One way to turn the *pMSE* into a utility function that is a function of θ, the parameters, is to take the expectation of *pMSE* given θ:

$$u(\mathbf{x}, \theta) = \mathrm{E}[pMSE(\mathbf{x}, \mathbf{x}^*) \mid \mathbf{x}, \theta], \tag{10}$$

where $\mathbf{x}$ is the private database, and $\mathbf{x}^*$ is the synthetic database and generated from a Bayesian synthesis model $f(\theta)$, i.e. $\mathbf{x}^* \sim f(\theta)$

- The sensitivity of the utility function is bounded

$$\Delta_u = sup_{\theta} sup_{\delta(\mathbf{x},\mathbf{y})=1} |u(\mathbf{x}, \theta) - u(\mathbf{y}, \theta)| \leq \frac{1}{n} \tag{11}$$

The Posterior Mechanism

- Dimitrakakis et al. (2017)
- Use the log-likelihood function as the utility function

$$u(\mathbf{x}, \theta) = \log \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta), \tag{12}$$

where the sensitivity of the log-likelihood function is bounded by

$$\Delta_u = sup_{\theta} sup_{\delta(\mathbf{x},\mathbf{y})=1} |u(\mathbf{x}, \theta) - u(\mathbf{y}, \theta)| \leq \Delta, \tag{13}$$

where Δ is called a Lipschitz bound (which can be infinite in some cases, such as normal, exponential, Poisson, geometric)

The Posterior Mechanism cont'd

- That is, we can draw private parameter draws $\hat{\theta}$ from:

$$\hat{\theta} \propto \exp\left(\frac{\epsilon \, log \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta)}{2\Delta_u}\right)\pi(\theta) \tag{14}$$

- This Posterior Mechanism achieves an $\epsilon = 2\Delta-$differential privacy guanratee for each posterior draw of θ

The Pseudo Posterior Mechanism

- Savistsky et al. (2019)
- Generalize the Posterior Mechanism to ensure $\Delta < \infty$
- Key: add weights in the likelihood function

$$\log \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta)^{\alpha_i}, \tag{15}$$

where

$$\alpha_i \propto \frac{1}{sup_{\theta \in \Theta} \log(\mathbf{x}_i \mid \theta)}. \tag{16}$$

- Weight-added likelihood is called pseudo likelihood

The Pseudo Posterior Mechanism cont'd

- Use the log-pseudo likelihood function as the utility function

$$u(\mathbf{x}, \theta) = \log \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta)^{\alpha_i}, \tag{17}$$

where the sensitivity of the log-pseudo likelihood function is bounded by

$$\Delta_u = sup_{\theta} sup_{\delta(\mathbf{x},\mathbf{y})=1} |u(\mathbf{x}, \theta) - u(\mathbf{y}, \theta)| \leq \Delta^{\alpha}, \tag{18}$$

where $\Delta^{\alpha} < \infty$

The Pseudo Posterior Mechanism cont'd

- That is, we can draw private parameter draws $\hat{\theta}$ from:

$$\hat{\theta} \propto \exp\left(\frac{\epsilon \, log \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta)^{\alpha_i}}{2\Delta_u}\right)\pi(\theta) \tag{19}$$

- This Pseudo Posterior Mechanism achieves an $\epsilon = 2\Delta^{\alpha}-$differential privacy guanratee for each posterior draw of θ

Example: estimating Poisson-distributed data cont'd

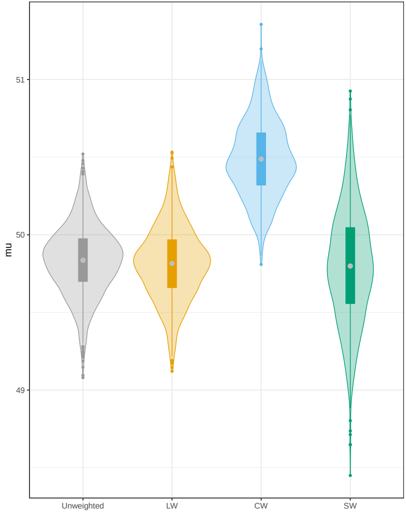

Figure 1: Violin plots of mu

References

- Dimitrakakis, C. and Nelson, B. and Zhang, Z. and Mitrokotsa, A. and Rubinstein, B. I. P. (2017). Differential privacy for Bayesian inference through posterior sampling. Journal of Machine Learning Research, 18(1), 343-381.
- Dwork, C. and McSherry, F. and Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Proceedings of the Third Conference on Theory of Cryptography, 265-284.
- McSherry, F., and K. Talwar. (2007). Mechanism Design via Differential Privacy. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, 94–103.
- Nissim, K., S. Raskhodnikova, and A. Smith. (2007). Smooth Sensitivity and Sampling in Private Data Anlaysis. Proceedings of the 39th Annual ACM Symposium on Theory of Computing, 75–83.

References

- Savitsky, T. D. and Williams, M. R. and Hu, J. (2019), Bayesian pseudo posterior mechanism under differential privacy, arXiv 1909.11796
- Snoke, J., and A. Slavkovic. (2018). pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity. Privacy in Statistical Databases, 138–159.
- Zhang, Z., B. I. P. Rubinstein, and C. Dimitrakakis. (2016). On the Differential Privacy of Bayesian Inference. Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2365–2371.