Differentially Private Synthetic Tabular Data

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

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 synthetic data

- A conjugate Bayesian model for categorical variables: Dirichlet-multinomial; now for contingency tables

1. Suppose we have a count vector $\mathbf{y}$ of length I, with total number of records $y.$ (the sum of $\mathbf{y}$). The multinomial sampling model follows:

$$\mathbf{y} \mid \boldsymbol{\theta} \sim \text{Multinomial}(y.; \boldsymbol{\theta}). \tag{1}$$

Recap of synthetic data

- A conjugate Bayesian model for categorical variables: Dirichlet-multinomial; now for contingency tables

1. Suppose we have a count vector $\mathbf{y}$ of length I, with total number of records $y.$ (the sum of $\mathbf{y}$). The multinomial sampling model follows:

$$\mathbf{y} \mid \boldsymbol{\theta} \sim \text{Multinomial}(y.; \boldsymbol{\theta}). \tag{1}$$

2. A conjugate prior for θ is Dirichlet:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}). \tag{2}$$

Recap of synthetic data

- A conjugate Bayesian model for categorical variables: Dirichlet-multinomial; now for contingency tables

1. Suppose we have a count vector $\mathbf{y}$ of length I, with total number of records $y.$ (the sum of $\mathbf{y}$). The multinomial sampling model follows:

$$\mathbf{y} \mid \boldsymbol{\theta} \sim \text{Multinomial}(y.; \boldsymbol{\theta}). \tag{1}$$

2. A conjugate prior for θ is Dirichlet:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}). \tag{2}$$

3. Due to conjugacy, we come to a Dirichlet posterior for θ:

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \text{Dirichlet}(\mathbf{y} + \boldsymbol{\alpha}). \tag{3}$$

Overview

- Can we make the Dirichlet-multinomial synthesizer satisfy $\epsilon-$differential privacy?
- The original $\epsilon-$differential privacy definition: a mechanism $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is $\epsilon-$differentially private for all $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$ and for all $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$ such that $\delta(\mathbf{x}, \mathbf{y}) = 1$:

$$\left| \ln \left(\frac{Pr[\mathcal{M}(\mathbf{x}) \in \mathcal{S}]}{Pr[\mathcal{M}(\mathbf{y}) \in \mathcal{S}]} \right) \right| \leq \epsilon. \tag{4}$$

- What to do for synthetic tabular data?

Overview cont'd

- Now in synthetic tabular data:

Let $\mathbf{y}$ denote the true count vector of length I, and $\mathbf{x}$ denote another count vector with Hamming distance 1 from $\mathbf{y}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$) and $\sum_{i=1}^{I} x_i = \sum_{i=1}^{I} y_i$. Let $y.^*$ denote an $\epsilon-$differentially private synthetic count vector, and $\boldsymbol{\theta}$ denote model parameters vector. In such setting, $\epsilon-$differential privacy requires

$$\left| \ln \left(\frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\theta})} \right) \right| \leq \epsilon. \tag{5}$$

Overview cont'd

- Now in synthetic tabular data:

Let $\mathbf{y}$ denote the true count vector of length I, and $\mathbf{x}$ denote another count vector with Hamming distance 1 from $\mathbf{y}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$) and $\sum_{i=1}^{I} x_i = \sum_{i=1}^{I} y_i$. Let $y.^*$ denote an $\epsilon-$differentially private synthetic count vector, and $\boldsymbol{\theta}$ denote model parameters vector. In such setting, $\epsilon-$differential privacy requires

$$\left| \ln \left(\frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\theta})} \right) \right| \leq \epsilon. \tag{5}$$

- If we can the Dirichlet-multinomial synthesizer satisfy $\epsilon-$differential privacy, we produce differentially private synthetic tabular data

Outline

The procedure

Abowd and Vilhuber (2008) and Machanavajjhala et al. (2008)

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

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

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \mathrm{Dirichlet}(\mathbf{y} + \boldsymbol{\alpha}), \tag{6}$$

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

The procedure

Abowd and Vilhuber (2008) and Machanavajjhala et al. (2008)

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

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

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \mathrm{Dirichlet}(\mathbf{y} + \boldsymbol{\alpha}), \tag{6}$$

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

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

$$\mathbf{y}^* \mid \boldsymbol{\theta}^* \sim \mathrm{Multinomial}(y.^*; \boldsymbol{\theta}^*), \tag{7}$$

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

Why it works?

- The posterior predictive distribution is:

$$
\begin{aligned}
p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\alpha}) &= \int p(\mathbf{y}^* \mid \boldsymbol{\theta}, \boldsymbol{\alpha}) \times p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\alpha}) d\boldsymbol{\theta} \\
&= \int \frac{y.^*!}{\prod_{i=1}^I y_i^*!} \times \prod_{i=1}^I \theta_i^{y_i^*} \times \frac{\Gamma(\sum_{i=1}^I y_i + \alpha_i)}{\prod_{i=1}^I \Gamma(y_i + \alpha_i)} \times \prod_{i=1}^I \theta_i^{y_i + \alpha_i - 1} d\boldsymbol{\theta} \\
&= \frac{y.^*!}{\prod_{i=1}^I y_i^*!} \times \frac{\Gamma(\sum_{i=1}^I y_i + \alpha_i)}{\prod_{i=1}^I \Gamma(y_i + \alpha_i)} \times \frac{\prod_{i=1}^I \Gamma(y_i^* + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^I y_i^* + y_i + \alpha_i)}. \quad (8)
\end{aligned}
$$

Why it works? cont'd

- To satisfy $\epsilon-$differential privacy, we require

$$\left|\log\left(\frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\alpha})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\alpha})}\right)\right| = \left|\log\left(\frac{\prod_{i=1}^{I}\Gamma(\alpha_i + x_i)}{\prod_{i=1}^{I}\Gamma(\alpha_i + y_i)} \times \frac{\prod_{i=1}^{I}\Gamma(y_i^* + \alpha_i + y_i)}{\prod_{i=1}^{I}\Gamma(y_i^* + \alpha_i + x_i)}\right)\right| \leq \epsilon \tag{9}$$

- $\mathbf{x}$ has Hamming distance 1 from $\mathbf{y}$ (i.e. $\delta(\mathbf{x}, \mathbf{y}) = 1$)
- $\sum_{i=1}^{I} x_i = \sum_{i=1}^{I} y_i$ (total sum is fixed)

Why it works? cont'd

- Assume the the only differences in $\mathbf{x}$ and $\mathbf{y}$ exist between the pairs $(x_i, x_{i'})$ and $(y_i, y_{i'})$
- Without loss of generality, assume $x_i = y_i - 1$ and $x_{i'} = y_{i'} + 1$

$$\frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\alpha})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\alpha})} = \frac{\alpha_i + y_i}{\alpha_{i'} + y_{i'} - 1} \times \frac{y_{i'}^* + \alpha_{i'} + y_{i'} - 1}{y_i^* + \alpha_i + y_i}, \tag{10}$$

where $y_i^* + y_{i'}^* \leq y.^*$.

Why it works? cont'd

$$\frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\alpha})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\alpha})} = \frac{\alpha_i + y_i}{\alpha_{i'} + y_{i'} - 1} \times \frac{y^*_{i'} + \alpha_{i'} + y_{i'} - 1}{y^*_i + \alpha_i + y_i}$$

- Maximized when $y_{i'} = 1, y^*_i = 0$ and $y^*_{i'} = z$.
- Minimized when $y_{i'} = 0, y^*_{i'} = 0$ and $y^*_i = z$.

$$\frac{\alpha_i}{y.^* + \alpha_i} \leq \frac{p(\mathbf{y}^* \mid \mathbf{y}, \boldsymbol{\alpha})}{p(\mathbf{y}^* \mid \mathbf{x}, \boldsymbol{\alpha})} \leq \frac{y.^* + \alpha_{i'}}{\alpha_{i'}}. \tag{11}$$

Why it works? cont'd

- Now to satisfy $\epsilon-$differential privacy where $\left|\log\left(\frac{p(\mathbf{y}^*|\mathbf{y},\boldsymbol{\alpha})}{p(\mathbf{y}^*|\mathbf{x},\boldsymbol{\alpha})}\right)\right| \leq \epsilon$, we require

$$\epsilon = \log\left(\frac{y_{\cdot}^* + \min(\alpha_i)}{\min(\alpha_i)}\right), \tag{12}$$

which results in

$$\min(\alpha_i) \geq \frac{y_{\cdot}^*}{\exp(\epsilon) - 1}. \tag{13}$$

Outline

CE data

Table 1: Variables used in the CE database. Data taken from the 2017 CE public use microdata samples.

Variable Name	Variable information
UrbanRural	Binary; the urban / rural status of CU: 1 = Urban, 2 = Rural.
Income	Continuous; the amount of CU income bfore taxes in past 12 months.
Race	Categorical; the race category of the reference person: 1 = White, 2 = Black, 3 = Native American, 4 = Asian, 5 = Pacific Islander, 6 = Multi-race.
Expenditure	Continuous; CU's total expenditures in last quarter.

The contingency table of Race categories

```r
require(readr)
CEdata <- read_csv("CEdata.csv")
Race_Count <- CEdata %>% count(Race)
Race_Count
```

```
## # A tibble: 6 x 2
##    Race     n
##   <dbl> <int>
## 1     1   816
## 2     2   109
## 3     3     7
## 4     4    39
## 5     5     6
## 6     6    17
```

Step 1: calculate α

- Expression for α

$$\min(\alpha_i) \geq \frac{y_{\cdot}^{*}}{\exp(\epsilon) - 1}$$

- Use privacy budget $\epsilon = 5$

```
epsilon <- 5

alpha_min <- sum(Race_Count$n)/(exp(epsilon) - 1)
alpha_min
```

```
## [1] 6.742953
```

```
alpha_vector <- rep(alpha_min, dim(Race_Count)[1])
alpha_vector
```

```
## [1] 6.742953 6.742953 6.742953 6.742953 6.742953 6.742953
```

Step 2: sample $\boldsymbol{\theta}^*$

- $\mathbf{y}$ is the original count vector of `Race` categories.

```
y_vector <- Race_Count$n
y_vector
```

```
## [1] 816 109   7  39   6  17
```

- With the calcuated `alpha_vector`, sample $\boldsymbol{\theta}^*$ from $\mathrm{Dirichlet}(\mathbf{y} + \boldsymbol{\alpha})$

```
require(gtools)
set.seed(123)
theta_DPsyn <- rdirichlet(n = 1, alpha = y_vector + alpha_vector)
theta_DPsyn
```

```
##           [,1]      [,2]        [,3]       [,4]       [,5]       [,6]
## [1,] 0.7804585 0.1242675 0.007558385 0.04464151 0.01837218 0.02470187
```

Step 3: sample $\boldsymbol{y}^*$

- With the sampled `theta_DPsyn`, we can sample $\mathbf{y}^*$ from $\text{Multinomial}(y.^*; \boldsymbol{\theta}^*)$

```
y_DPsyn <- rmultinom(n = 1, size = sum(y_vector),
                     prob = theta_DPsyn)
```

Step 3: sample $\boldsymbol{y}^*$ cont'd

- Put original and synthetic side-by-side

```
y_DPsyn <- rmultinom(n = 1, size = sum(y_vector),
                     prob = theta_DPsyn)
y_both <- data.frame(y_vector, y_DPsyn)
names(y_both) <- c("original", "DPsynthetic")
y_both
```

```
##   original DPsynthetic
## 1      816         776
## 2      109         118
## 3        7           6
## 4       39          53
## 5        6          22
## 6       17          19
```

Summary and discussion

- The choice of privacy budget ϵ has great influence on the resulted differentially private synthetic contingency table
- What are your thoughts?

Summary and discussion

- The choice of privacy budget ϵ has great influence on the resulted differentially private synthetic contingency table
- What are your thoughts?

- The higher the value of ϵ, the higher the utility

Summary and discussion

- The choice of privacy budget ϵ has great influence on the resulted differentially private synthetic contingency table
- What are your thoughts?

- The higher the value of ϵ, the higher the utility

- Other differentially private synthetic tabular data models:
  1. beta-binomial (McClure and Reiter, 2012)
  2. gamma-Poisson (Quick, 2019)

References

- Abowd, J. M., and L. Vilhuber. (2008). How Protective Are Synthetic Data? Privacy in Statistical Databases, 239–46.
- Machanavajjhala, A., D. and Kifer, J. and Abowd, J.and Gehrke, and L. Vilhuber. (2008). Privacy: Theory Meets Practice on the Map. The IEEE 24th International Conference on Data Engineering, 277–86.
- McClure, D., and J. P. Reiter. (2012). Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data. Transactions on Data Privacy 5: 535–52.
- Quick, H. (2019). Generaing Poisson-distributeed differentially private synthetic data. arXiv:1906.00455.