Differential Privacy - An Overview #1

Jingchen (Monika) Hu

Vassar College

Data Confidentiality

Outline

Outline

Recap of synthetic data

- Synthetic microdata
  1. Bayesian synthesis models (Lectures 3 and 4)
  2. Methods for utility evaluation (Lectures 5 and 6)
  3. Methods for risk evaluation (Lectures 7, 8 and 9)
- Synthetic data is driven by modeling, i.e. from the angle of utility
- Risk evaluation methods make assumption about intruder's knowledge and behavior

Recap of synthetic data

- Synthetic microdata
    1. Bayesian synthesis models (Lectures 3 and 4)
    2. Methods for utility evaluation (Lectures 5 and 6)
    3. Methods for risk evaluation (Lectures 7, 8 and 9)
- Synthetic data is driven by modeling, i.e. from the angle of utility
- Risk evaluation methods make assumption about intruder's knowledge and behavior
- Can we approach data privacy from the angle of risk?

Differential privacy

- Dwork et al. (2006), computer science
- A formal mathematical framework to provide privacy protection guarantees
- Main initial focus is on summary statistics, not microdata nor tabular data

Outline

Adding noise for privacy protection

- Key idea: add noise to the output of queries made to databases
- Added noise is random; depends on a predetermined privacy budget and the type of queries

Definitions: database

- Databases are datasets that data analysts use for analysis
- Databases are confidential, whether and how can the data analyst gets information of quantities of interest?
- Whether and how the database holder to provide information to the data analyst: useful and privacy-protected

Definitions: database cont'd

- Example: CE sample

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.

- A quantity of interest: the number of rural CUs in this sample

Definitions: query

- Denote numeric queries as functions $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, mapping databases to k real numbers, $\mathbb{R}^k$
- Example: the data analyst can send the following query to the CE database
  - how many rural CUs are there in this sample?
- Discussion: As the database holder, can we give out the actual values? Why or why not?

Definitions: query

- Denote numeric queries as functions $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, mapping databases to k real numbers, $\mathbb{R}^k$
- Example: the data analyst can send the following query to the CE database
    - how many rural CUs are there in this sample?
- Discussion: As the database holder, can we give out the actual values? Why or why not?
- We will add noise to the query output for privacy protection, how?

Definitions: Hamming-distance

- Given databases $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$, let $\delta(\mathbf{x}, \mathbf{y})$ denote the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ by:

$$\delta(\mathbf{x}, \mathbf{y}) = \#\{i : x_i \neq y_i\}. \tag{1}$$

- Under differential privacy, we add noise by considering the scenario where two databases differ by one record, i.e. $\delta(\mathbf{x}, \mathbf{y}) = 1$

Definitions: ℓ_1-sensitivity

- The ℓ_1-sensitivity is the magnitude a single individual's data can change the ℓ_1 norm of the function f in the worst case
- Formally, the ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ is:

$$\Delta f = \max_{\mathbf{x},\mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}, \delta(\mathbf{x},\mathbf{y})=1} ||f(\mathbf{x}) - f(\mathbf{y})||_1. \quad (2)$$

- The ℓ_1 norm between $f(\mathbf{x})$ and $f(\mathbf{y})$ is the absolute difference between $f(\mathbf{x})$ and $f(\mathbf{y})$, denoted as $||f(\mathbf{x}) - f(\mathbf{y})||_1$
- Δf is the maximum change in the function 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. $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}, \delta(\mathbf{x}, \mathbf{y}) = 1$)

Definitions: ℓ_1-sensitivity cont'd

- Example: CE database
  - $\mathbf{x}$ is the confidential CE sample, $\mathbf{y}$ is the database where one data entry is different from $\mathbf{x}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$)
- Query f: How many rural CUs are there in this sample?
  - question: what is the ℓ_1-sensitivity for query f?

Definitions: ℓ_1-sensitivity cont'd

- Example: CE database
  - $\mathbf{x}$ is the confidential CE sample, $\mathbf{y}$ is the database where one data entry is different from $\mathbf{x}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$)
- Query f: How many rural CUs are there in this sample?
  - question: what is the ℓ_1-sensitivity for query f?
  - answer: $\Delta f = 1$

Definitions: ℓ_1-sensitivity cont'd

- Example: CE database
  - $\mathbf{x}$ is the confidential CE sample, $\mathbf{y}$ is the database where one data entry is different from $\mathbf{x}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$)
- Query f: How many rural CUs are there in this sample?
  - question: what is the ℓ_1-sensitivity for query f?
  - answer: $\Delta f = 1$
- Another query fy: what is the average income of this sample?
  - question: what is the ℓ_1-sensitivity for query f?

Definitions: ℓ_1-sensitivity cont'd

- Example: CE database
  - $\mathbf{x}$ is the confidential CE sample, $\mathbf{y}$ is the database where one data entry is different from $\mathbf{x}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$)
- Query f: How many rural CUs are there in this sample?
  - question: what is the ℓ_1-sensitivity for query f?
  - answer: $\Delta f = 1$
- Another query fy: what is the average income of this sample?
  - question: what is the ℓ_1-sensitivity for query f?
  - answer: $\Delta f = \frac{b-a}{n}$ ($b - a$ is the range, and n is the sample size)

Definitions: ℓ_1-sensitivity cont'd

- Example: CE database
  - $\mathbf{x}$ is the confidential CE sample, $\mathbf{y}$ is the database where one data entry is different from $\mathbf{x}$ ($\delta(\mathbf{x}, \mathbf{y}) = 1$)
- Query f: How many rural CUs are there in this sample?
  - question: what is the ℓ_1-sensitivity for query f?
  - answer: $\Delta f = 1$

- Another query fy: what is the average income of this sample?
  - question: what is the ℓ_1-sensitivity for query f?
  - answer: $\Delta f = \frac{b-a}{n}$ ($b - a$ is the range, and n is the sample size)

- In sum, the ℓ_1-sensitivity depends on the database and the query sent to the database by the data analyst

Definitions: $\epsilon-$differential privacy

- We want to guarantee that a mechanism (aka technology) behaves similarly (i.e. giving similar outputs) on similar inputs (e.g. when two databases differ by one)
- One approach:
    - bound the log ratio of the probabilities of the outputs from above
    - give an upper bound on the noise added to the output to preserve privacy

Definitions: $\epsilon-$differential privacy

- We want to guarantee that a mechanism (aka technology) behaves similarly (i.e. giving similar outputs) on similar inputs (e.g. when two databases differ by one)
- One approach:
  - bound the log ratio of the probabilities of the outputs from above
  - give an upper bound on the noise added to the output to preserve privacy
- 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 $\delta(\mathbf{x}, \mathbf{y}) = 1$:

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

Definitions: $\epsilon-$differential privacy cont'd

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

- The ratio $\ln \left(\frac{Pr[\mathcal{M}(\mathbf{x}) \in \mathcal{S}]}{Pr[\mathcal{M}(\mathbf{y}) \in \mathcal{S}]} \right)$
    - is the log of the ratio of the probability of the output undergone mechanism $\mathcal{M}$ from the database $\mathbf{x}$, and that from the database $\mathbf{y}$
    - can be considered as the difference in the outputs
- Bound the ratio above by ϵ, the privacy budget (to be defined next), i.e. setting the maximum difference
- $\epsilon-$differential privacy provides us a means to perturb the output by adding noise, so that similar inputs produce similar outputs under the mechanism $\mathcal{M}$

Definitions: privacy budget

- The term ϵ is the privacy budget, that is to be spent by the database holder when answering queries

Implications

- With given privacy budget, we can then add noise according to the $\epsilon-$differential privacy definition to the output, in order to preserve privacy
- Relationships among: database, query, sensitivity, privacy budge and added noise
- Two important implications:
  1. the added noise is positively related to the sensitivity
  2. the added noise negatively related to the privacy budget

Implications: sensitivity and added noise

- The ℓ_1-sensitivity of query (function) f is to capture the magnitude a single individual's data can change the ℓ_1 norm of the query f in the worse case, denoted as Δf
- ℓ_1-sensitivity depends on
  1. the database
  2. the query
- Examples:
  1. a count query, $\Delta f = 1$ (regardless of the database)

Implications: sensitivity and added noise

- The ℓ_1-sensitivity of query (function) f is to capture the magnitude a single individual's data can change the ℓ_1 norm of the query f in the worse case, denoted as Δf
- ℓ_1-sensitivity depends on
  1. the database
  2. the query
- Examples:
  1. a count query, $\Delta f = 1$ (regardless of the database)
  2. an average query, $\Delta f = \frac{b-a}{n}$ (depends on the database: a, b, n)

Implications: sensitivity and added noise cont'd

- For a query f with large ℓ_1-sensitivity, Δf, larger noise is needed for the same level of privacy protection (i.e. given fixed privacy budget), and vice versa
- Consider two queries:
  1. what is the average income of this sample (income before taxes in past 12 months)?
  2. what is the average expenditure of this sample (total expenditures in last quarter)?
- Question # 1: given fixed privacy budget ϵ, which query has a larger sensitivity?

Implications: sensitivity and added noise cont'd

- For a query f with large ℓ_1-sensitivity, Δf, larger noise is needed for the same level of privacy protection (i.e. given fixed privacy budget), and vice versa

- Consider two queries:
  1. what is the average income of this sample (income before taxes in past 12 months)?
  2. what is the average expenditure of this sample (total expenditures in last quarter)?

- Question # 1: given fixed privacy budget ϵ, which query has a larger sensitivity?

- Answer # 1: 1

Implications: sensitivity and added noise cont'd

- For a query f with large ℓ_1-sensitivity, Δf, larger noise is needed for the same level of privacy protection (i.e. given fixed privacy budget), and vice versa
- Consider two queries:
  1. what is the average income of this sample (income before taxes in past 12 months)?
  2. what is the average expenditure of this sample (total expenditures in last quarter)?
- Question # 1: given fixed privacy budget ϵ, which query has a larger sensitivity?
- Answer # 1: 1
- Question # 2: given your answer to Question # 1, which query needs a larger noise to be added?
- Answer # 2: 1

Implications: sensitivity and added noise cont'd

- In sum, the sensitivity and the added noise are positively related: given fixed privacy budget ϵ, larger sensitivity results in larger added noise

Implications: privacy budget and added noise

- $\epsilon-$differential privacy provides an upper bound on the noised necessary to be added to the output for privacy protection
- The upper bound is ϵ, the privacy budget
- The privacy budget ϵ does not depend on the database or the query

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

- Discussion: what's the relationship between the privacy budget and added noise, given fixed ℓ_1-sensitivity?
  - what happens to the added noise when ϵ increases?
  - what happens to the added noise when ϵ decreases?

Implications: privacy budget and added noise cont'd

- In sum, the privacy budget and the added noise are negatively related: given fixed sensitivity Δf, larger privacy budget results in smaller added noise

Summary

- Key idea: add noise to the output of queries made to databases
- Added noise is random; depends on a predetermined privacy budget and the type of queries
- Two important implications:
  1. the added noise is positively related to the sensitivity
  2. the added noise negatively related to the privacy budget
- We will explore the Laplace Mechanism, which satisfies $\epsilon-$differential privacy, and add Laplace noise to summary statistics such as count and average

References

- 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.