CS480/680, Spring 2023

Assignment 4

Designer: Shufan Zhang; Instructor: Hongyang Zhang

Released: July 17; Due: July 31, noon

[IMPORTANT] Due to the due date changes, we adjust the marks for this assignment as follows:

- The marks for Question 1 are normalized to 100% for this assignment. If you answer Q1 correctly, you will get full marks for this assignment, which is 10 marks for the overall grades of this course.
- The marks for Question 2 change to bonus for all assignments in this course. If you get Q2 done correctly, you have 3 bonus marks for all the assignments, but the maximum marks you can get for the assignment are capped by 40 contributing to the overall grade of the course.

The new due date is **July 31, noon**.

1. **Writing: Differentially Private Data Analytics [100 Marks]**

Name	Age	Gender	Num. Matches
Henry	42	Male	8
Sarah	36	Female	16
Austin	22	Non-Binary	5
Adrian	44	Male	12
Natalie	30	Female	5
Chloe	23	Non-Binary	20
Tony	45	Male	13
Christine	28	Non-Binary	20
Olivia	39	Female	35

Table 1: Number of Matches Information

Tinker would like to use differential privacy to publish the data, since it is resilient to background knowledge. To do this, Tinker releases a differentially private histogram showing the number of users having a specific number of matches. To generate this histogram, Laplace noise is added to the true value. For instance, if 4 users in the dataset have 5 matches each, Laplace noise would be added to the total number of such users (4) to hide the true number of users with 5 matches.

The *histogram* representation of the dataset $x = (x_0, \ldots, x_{n-1})$, where $n = 36$, is a 36-dimensional vector (ranging from a minimum of 0 matches to the maximum number of matches observed in the dataset, 35) where the j-th entry is the number of x's rows whose number of matches is equal to j. For instance, according to the data in Table 1, $h_{20}(x) = 2$ since two users in the database (Chloe and Christine) have 20 Tinker matches, and thus:

$$h(x) := (h_0(x), \ldots, h_{n-1}(x)), n = 36$$

Consider the following *noisy histogram algorithm* output:

$$\hat{h}(x) := (h_0(x) + L_0, \ldots, h_{n-1}(x) + L_{n-1})$$

where every $L_j \sim Laplace(\lambda)$ is independent Laplace Noise.

Note: Wikipedia gives a good overview of differential privacy and differentially private mechanisms: `https://en.wikipedia.org/wiki/Differential_privacy`. You may also seek out additional resources to help answer this question.

(a) [15 Marks] What is the sensitivity of this query of releasing histograms?

- Sensitivity = $max(|h_i - h_{i-1}|)$ where $i = 1, 2, ..., 35$
- Sensitivity = 2

(b) [15 Marks] Tinker sets the parameters to $\epsilon = 0.01$, then what is λ in Laplace Mechanism?

- $\lambda = \frac{S}{\epsilon} = \frac{2}{0.01} = 200$

(c) [15 Marks] Please analyze the expected error of this mechanism. ($\mathcal{E} = \sum_{i=1}^{d} \mathbb{E}[(o_i - c_i)^2]$, where o_i is the ith entry of the noisy output, and c_i is the ith entry of the true answer.)

- Laplace noise has mean equal to 0. So the expected value of noisy output is equal to the true value plus Laplace noise. Therefore, $E[(o_i - c_i)^2] = E[L_i^2]$
- Note that $E[x^2] = Var[x] + E^2[x]$
- $E[L_i^2] = Var[L_i] + 0 = 2\lambda^2 = 2 \times 40000 = 80000$

(d) [25 Marks] Does this mechanism satisfy the definition of ϵ-differential privacy? Will the histogram output of this mechanism be useful? Justify.

- Definition stated as $Pr[M(X) \in S] \leq e^{\epsilon} Pr[M(X') \in S]$
- $\frac{Pr[M(X) \in S]}{Pr[M(X') \in S]} = \frac{Pr[h + Laplace(\lambda) \in S]}{Pr[h' + Laplace(\lambda) \in S]}$
- Define $S' = \{x - h : x \in S\}$
- $\frac{Pr[Laplace(\lambda) \in S']}{Pr[Laplace(\lambda) \in S' + (h - h')]} = \frac{\frac{1}{2\lambda} exp(-\frac{|x|}{\lambda})}{\frac{1}{2\lambda} exp(-\frac{|x-(h-h')|}{\lambda})}$ if we let Laplace(λ) = x
- $= exp(\frac{|x-(h-h')|-|x|}{\lambda}) \leq exp(\frac{h-h'}{\lambda}) \leq exp(\frac{S}{\lambda}) = exp(\epsilon)$
- Therefore, it satisfied the definition of ϵ-differential privacy. The output of this mechanism is useful.

(e) [30 Marks] Tinker would like to collect new user data with local differential privacy guarantee. Consider a domain $\Sigma = \{l_1, \ldots, l_k\}$ of k locations, please design a randomized response R that takes in a true location $l \in \Sigma$ and randomly outputs a location $o \in \Sigma$. (Describe your algorithm and show that the algorithm achieves ϵ-local differential privacy.)

- Let ϵ be a hyper-parameter which controls the level of privacy protection.
- For output l_i, output l_i directly with probability $exp(\frac{\epsilon}{2})$, otherwise, output a uniformly chosen random location in Σ
- To show this satisfied ϵ-differential privacy, let's consider two cases.
- Assume that the location in D is l_1 and in D' is l_2
- Case 1: if outputs the true location ($l_i = l_1 or l_2$)
- The probability of reporting the true location in D and D' is $exp(\frac{\epsilon}{2})$ in both cases, and the ratio is 1, which is $\leq e^{\epsilon}$
- Case 2: if outputs the random location:
- Both probabilities are $1 - exp(\frac{\epsilon}{2})$, and the ratio is 1, which $\leq e^{\epsilon}$
- Therefore, it satisfied ϵ-differential privacy

2. **Coding: Private Data Synthesis [Bonus 3 Marks for All Assignments]**

The US Census Bureau collects the geographic and demographic data of US residents. The data is anticipated to be used for performing several machine learning or analytic tasks to better understand the residents or incidents of residents, e.g., contact tracing, civic planning, natural disaster rapid response, etc. However, these types of data are considered highly sensitive that contain personally identifiable information (PII) of individuals. Due to privacy laws or regulations, some privacy enhancement techniques should be applied to guarantee individual privacy.

Now we consider using differential privacy (DP) as the means to protect individual privacy. One way to enforce differential privacy over the collected data is **private data synthesis**, meaning generating a synthetic dataset with DP guarantees. A **synthetic dataset** is a collection of artificially generated data that simulates real-world data. Instead of being collected from actual observations or measurements, synthetic data is created using various statistical and computational techniques to mimic the characteristics and patterns of the original data. Enforcing DP in data synthesis requires the data generation algorithm to be proved as differentially private.

Now you are given a dataset called *Adult* (which can be accessed via `https://archive.ics.uci.edu/dataset/2/adult`). This dataset has 48,842 rows and 14 attributes. These attributes (also called the *schema* of the dataset) include age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country. Each row thereby denotes a person. A typical machine learning task on this dataset is to use these 14 attributes as features to predict whether a person makes over 50K a year.

Your tasks for this question are the following:

(a) [1 Marks] Design a DP synthetic dataset generation algorithm (pseudo-code with step-by-step brief description), and show why it is DP. You can use any generative models you like, including GAN, BayesianNet, and your self-created ones.

(b) [1 Marks] Generate two DP synthetic datasets (with $\epsilon = \{1, 5\}$ respectively) for the Adult dataset, which should contain the exact same number of rows and 15 columns (14 features + 1 prediction class). Submit the code and the generated datasets.

(c) [1 Marks] Choose any classification models (e.g., decision tree, etc.) to perform the learning task (on predicting if a person has income over 50K a year), and for 3 datasets (the ground truth and 2 synthetic datasets), train the model on the first 2/3 of the dataset and test it on the rest of the data. Report the accuracy or other reasonable utility metrics (e.g., ROC) for three testings. You can plot the results or simply report the numbers with a brief justification.