Homework 1
CSCI 699: Privacy-Preserving Machine Learning

Instructor: Sai Praneeth Karimireddy
Student Name: Shuying Cao
USCID:6773049268
Due: Sep 20, 2024

Instructions: Answer the following questions clearly and concisely. Justify your answers with precise reasoning where necessary. Points for each question are indicated. Type your answers in Latex and submit the pdf.

Questions

Question 1: K-Anonymity Interpretation (3 points)

Consider an anonymized dataset that has been released under the notion of k-anonymity. Explain if k-anonymity protects against each of the following privacy attacks.

(a) **Membership inference:** Can an attacker determine whether a specific individual is part of the dataset?

Answer: Yes. From the definition of K-Anonymity, every row in the database, there should be (k-1) others with the exact same attributes. Therefore, attackers can know whether a specific person is in the dataset by deducing.

(b) **Sensitive attribute disclosure:** Can an attacker deduce whether a specific individual has a particular sensitive attribute (e.g., COVID positive/negative)?

Answer: Yes. Attackers can deduce the sensitive information of a specific person. Because if everyone in the dataset had covid-19 before, then attackers could know the person in the dataset once infected by covid-19,leading to the sensitive attribute disclosure.

(c) **Identity disclosure:** Can an attacker identify which specific data record corresponds to a particular individual?

Answer: No. Because every row in the database, there should be (k-1) others with the exact same attributes, attackers can not identify a specific data corresponds to a particular person.

Question 2: Differential Privacy for Datasets with Multiple Differences (2 points)

Let $A(D)$ be an algorithm that satisfies ε-differential privacy (DP) when the notion of "similar datasets" refers to datasets that differ in exactly one datapoint. Prove that the same algorithm $A(D)$ satisfies $k\varepsilon$-differential privacy when we redefine neighboring datasets to be those that differ in up to k datapoints.

Proof: If we have k Datasets $D_0, D_1, D_2, ..., D_k$, and these datasets are similar datasets to each other, which means the dataset and the dataset next to it just differ in exactly one datapoint. And D_1 and D_k differ in k datapoints.

For the ε-differential privacy (DP)

$$\frac{Pr[A(D) = y]}{Pr[A(D') = y]} \leq e^{\varepsilon} \tag{1}$$

we can get

$$\begin{aligned} Pr[A(D_0) = y] &\leq e^{\varepsilon} * Pr[A(D_1) = y] \\ &\leq e^{\varepsilon} * e^{\varepsilon} * Pr[A(D_2) = y] \\ &\leq e^{\varepsilon} * e^{\varepsilon} * e^{\varepsilon} * Pr[A(D_3) = y] \\ &... \\ &\leq e^{k\varepsilon} * Pr[A(D_k) = y] \end{aligned} \tag{2}$$

Proofed.

Question 3: Trade-off Curves for Randomized Classifiers (3 points)

Given an algorithm A and two datasets D, D', we output $Y = A(D)$ or $A(D')$ with equal probability (0.5). An adversary sees the output Y but does not know which dataset was used, and they construct a classifier $c(Y)$ to distinguish whether Y came from D or D'. The classifier c has the following properties:

- Type I error ($Pr[c(Y) = D'|Y = A(D)]$): 0.2
- Type II error ($Pr[c(Y) = D|Y = A(D')]$): 0.4

Now, consider a modified classifier $c'(Y)$ defined as follows:

(a) **First modification:** If $c(Y)$ predicts D, the modified classifier $c'(Y)$ outputs D with probability $(1 - p)$ and flips the prediction to D' with probability p. If $c(Y) = D$, then $c'(Y) = D$. Derive the type I and type II errors of the modified classifier $c'(Y)$ as a function of p.

Answer: From the question, we can know:

$$\begin{aligned} Pr[c'(Y)] = D|c(Y) = D] &= 1 - p \\ Pr[c'(Y)] = D'|c(Y) = D] &= p \\ Pr[c'(Y)] = D'|c(Y) = D'] &= 1 \end{aligned} \tag{3}$$

Type I Error:

$$\begin{aligned} Pr[c'(Y) = D'|Y = A(D)] &= Pr[c(Y) = D'|Y = A(D)] + Pr[c(Y) = D|Y = A(D)] * p \\ &= 0.2 + 0.8 * p \end{aligned} \tag{4}$$

Type II Error:

$$Pr[c'(Y) = D|Y = A(D')] = (1 - p) * 0.4 \tag{5}$$

(b) **Second modification:** Further modify the classifier as follows: when $c(Y) = D$, flip the prediction with probability p as before. Additionally, when $c(Y) = D'$, flip the prediction to D with probability q. Derive the new type I and type II errors of this modified classifier as functions of both p and q.

Answer: From the question, we can know:

$$\begin{aligned} Pr[c''(Y)] = D'|c(Y) = D] &= p \\ Pr[c''(Y)] = D|c(Y) = D] &= 1 - p \\ Pr[c''(Y)] = D|c(Y) = D'] &= q \\ Pr[c'(Y)] = D'|c(Y) = D'] &= 1 - q \end{aligned} \tag{6}$$

Type I Error:

$$\begin{aligned} Pr[c''(Y) = D'|Y = A(D)] &= Pr[c''(Y)] = D'|c(Y) = D] * Pr[c(Y) = D|Y = A(D)] \\ &+ Pr[c''(Y)] = D|c(Y) = D] * Pr[c(Y) = D'|Y = A(D)] \\ &= 0.8 * p + (1 - q) * 0.2 \\ &= 0.2 + 0.8p - 0.2q \end{aligned} \tag{7}$$

Type II Error:

$$\begin{aligned} Pr[c''(Y) = D|Y = A(D')] &= Pr[c''(Y)] = D|c(Y) = D'] * Pr[c(Y) = D'|Y = A(D')] \\ &+ Pr[c''(Y)] = D|c(Y) = D] * Pr[c(Y) = D|Y = A(D')] \\ &= q * 0.6 + (1 - p) * 0.4 \\ &= 0.4 + 0.6q - 0.4p \end{aligned} \tag{8}$$

(c) **Optimization:** For each value of $\alpha \in [0, 1]$, compute the optimal values of p and q to minimize the weighted error function:

$$\min_{p,q} \alpha \cdot \text{Type I Error}(p, q) + (1 - \alpha) \cdot \text{Type II Error}(p, q).$$

Based on this, plot the trade-off curve between the type I and type II errors.

Answer:

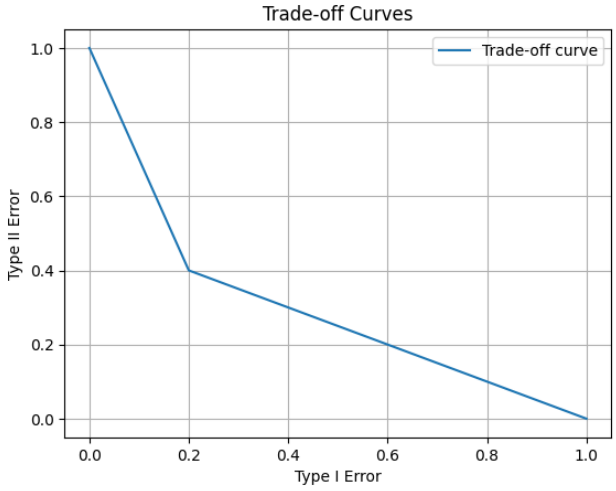

Figure 1: Trade-off Curve

Question 4: Hypothesis testing and Differential Privacy (2 points)

Let A be an algorithm that satisfies ε-differential privacy. Prove the following lower bound relationship between the type I and type II errors of any hypothesis test based on the output of A:

$$e^{\varepsilon} \cdot \text{Type I Error} + \text{Type II Error} \geq 1.$$

Proof: we could write Type I Error and Type II Error as following:

$$\begin{aligned} TypeIError &= Pr[guessD|D'] \\ TypeIIError &= Pr[guessD'|D] \end{aligned} \tag{9}$$

Then, using this we can transform the formula we want to prove.

$$\begin{aligned} e^{\varepsilon} * Type1Error + Type2Error &\geq 1 \\ e^{\varepsilon} * Pr[guessD|D'] + Pr[guessD'|D] &\geq 1 \\ e^{\varepsilon} * Pr[guessD|D'] &\geq 1 - Pr[guessD'|D] \\ e^{\varepsilon} * Pr[guessD|D'] &\geq Pr[guessD|D] \\ e^{\varepsilon} &\geq \frac{Pr[guessD|D]}{Pr[guessD|D']} \\ e^{\varepsilon} &\geq \frac{Pr[A(D) = y]}{Pr[A(D') = y]} \end{aligned} \tag{10}$$

Proofed.