Homework 1
CSCI 699: Privacy-Preserving Machine Learning

Instructor: Sai Praneeth Karimireddy
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?

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

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

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.

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.

(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.

(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.

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:

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