CIS 6261: Trustworthy Machine Learning (Spring 2023)
Homework 2 — Data Privacy & Differential Privacy

Name: Anol Kurian Vadakkeparampil

March 22, 2023

This is an individual assignment. Academic integrity violations (i.e., cheating, plagiarism) will be reported to SCCR! The official CISE policy recommended for such offenses is a course grade of E. Additional sanctions may be imposed by SCCR such as marks on your permanent educational transcripts, dismissal or expulsion.

Reminder of the Honor Pledge: On all work submitted for credit by Students at the University of Florida, the following pledge is either required or implied: *"On my honor, I have neither given nor received unauthorized aid in doing this assignment."*

Instructions

Please read the instructions and questions carefully. Write your answers directly in the space provided. Compile the tex document and hand in the resulting PDF.

For this assignment, you will solve several data privacy problems. The third problem asks you to implement a differential privacy mechanism using Python3. Use the code skeleton provided and submit the completed source file(s) alongside with the PDF. *Note: bonus points you get on this assignment *do* carry across assignments.*

Problem 1: Syntactic Metrics (20 pts)

Consider the data set depicted in Table 2. Answer the following questions. (Justify your answers as appropriate.)

Age	Zip Code	Sex	Diagnosis
40-49	32605	M	COVID-19
30-39	32607	M	Broken Leg
30-39	32607	M	Cancer
40-49	32611	F	Heart Disease
20-29	32607	F	Asthma
20-29	32607	F	Heart Disease
40-49	32611	M	Hypertension
40-49	32611	F	COVID-19

Table 1: Anonymized Data Set 1.

1. (5 pts) What are the quasi-identifier(s)? What are the sensitive attribute(s)?

   The quasi-identifiers are Age, Zip Code, and Sex. The sensitive attribute is Diagnosis.

2. (5 pts) What is the largest integer k such that the data set satisfies k-anonymity? What is the largest integer l such that the data set satisfies l-diversity?

   Due to the presence of quasi-identifier Zip : 32605, which is unique ;we have a constraint of $k = 1$ for k-anonymity and $l = 1$ for l-diversity.

3. (10 pts) Modify the data set using generalization and suppression to ensure that it satisfies 4-anonymity and 3-diversity. Here we are looking for a solution that minimally affects the utility of the data. Write the modified data set below.

   To achieve 4-anonymity, we can generalize the age to be in the range of 20-39 or 40-49 and suppress the zip code partially and Sex completely. On doing this we end up with a dataset that is in itself contains 3-diversity, i.e. 3 unique sensitive entries for each group. The modified data set is:

Age	Zip Code	Sex	Diagnosis
40-49	326**	*	COVID-19
20-39	326**	*	Broken Leg
20-39	326**	*	Cancer
40-49	326**	*	Heart Disease
20-39	326**	*	Asthma
20-39	326**	*	Heart Disease
40-49	326**	*	Hypertension
40-49	326**	*	COVID-19

Table 2: Modified Dataset.

*Note: * denotes suppressed data.*

Problem 2: Differential Privacy Mechanisms (40 pts)

Social science researchers at the University of Florida want to conduct a study to explore the prevalence of crime among students. They have a list of crimes and for each they want to know if the proportion of students that committed a specific crime exceeds some threshold $t \in (0, 1)$. For example: is the proportion of students who have ever stolen a bike larger than $t = 0.05$?

Researchers are ethical so they want to carefully design the study to ensure that participants respond truthfully and that privacy is protected. They reached out to you, a CIS 6261 student, to evaluate their methods.

Suppose X is the database of students' binary answers (yes (1) or no (0)) to the question of whether they have ever committed a specific crime and let t be a threshold $t \in (0, 1)$.

Define $g(X)$ to be number of yes (1) answers in the database X and let $n = |X|$ denote the number of students in the database. Also define the following thresholding function parameterized by a threshold t: $\text{thresh}_t : \mathbb{R} \to \{0, 1\}$ such that $\text{thresh}_t(x) = 1$ if $x \geq t$ and $\text{thresh}_t(x) = 0$ otherwise (i.e., if $x < t$).

Let f_t be the query function. It is defined as follows. On input database X, let $f_t(X) = \text{thresh}_t(\frac{g(X)}{|X|})$. In other words: $f_t(X) = 1$ if the proportion of yes (1) in X is equal or larger than t (i.e., if $\frac{g(X)}{n} \geq t$) and $f_t(X) = 0$ otherwise ((i.e., if $\frac{g(X)}{n} < t$)).

Answer the following questions.

1. (5 pts) What is Δg, the global sensitivity of g? Give the definition of global sensitivity and then give the answer.

   Definition:The maximum amount that the results of a function can vary when a single person's data is added to or deleted from a dataset is measured by global sensitivity. It is a fundamental idea in differential privacy, a framework for creating data analysis algorithms that protect user privacy. The biggest absolute difference between the results of any two nearby datasets, where two datasets are deemed neighboring if they differ only by the data of a single individual, is known as global sensitivity.

   Removing any one individual from the database can change g(X) by at most 1, so the global sensitivity of g is $\Delta g = 1$

2. (5 pts) Now let's turn to f_t the actual query function. What is Δf_t, the global sensitivity of f_t?

   $\Delta f_t = 1$. Since f_t returns value of either 1 or 0 based on threshold value.

3. (5 pts) Does Δf_t depend on the thresold t? (Why or why not?)

   No, because no matter the threshold value, f_t always outputs 0 or 1 based on whether the ratio is above or below the threshold value.

4. (5 pts) Give an example of a database X (with a least 5 entries) such that the local sensitivity of f_t given threshold $t = 0.3$ is 0. First give the definition of local sensitivity then answer the question.

   Definition: The local sensitivity of a function is defined as the maximum amount by which the output of the function changes when a single individual's data is added or removed from the dataset.

   X = (0, 0, 0, 0, 0) It does not matter what data is removed or added to this dataset; since value of f_t generated for an entry of 0 or 1 would be 0.

Now let's consider the following mechanisms. For each of them, you are asked to prove or disprove whether they satisfy differential privacy. (To prove differential privacy you need to show that the mechanism satisfies the definition. To disprove you can simply give a counterexample.)

Recall the definition of (pure) ε-differential privacy.

Definition 1. *A randomized algorithm $\mathcal{F}$ satisfies* ε-differential privacy *(for $\varepsilon > 0$) if for any two neighboring databases X, X' and any $S \subseteq \text{Range}(\mathcal{F})$:*

$$\Pr\{\mathcal{F}(X) \in S\} \leq e^{\varepsilon} \Pr\{\mathcal{F}(X') \in S\} \ .$$

5. (5 pts) Mechanism A: $\mathcal{F}(X) = f_t(X)$. That is the mechanism simply returns the output of f_t.

   This mechanism does not satisfy differential privacy.

   Counterexample: Let X and X' be two neighboring databases that differ only in one student's response. Suppose that in database X, the proportion of students who have committed the crime is slightly below t, while in database X' the proportion is slightly above t. Then, the output of f_t on X is 0 and the output on X' is 1, which violates differential privacy.

6. (5 pts) Mechanism B: $\mathcal{F}(X) = \text{thresh}_{0.5}(f_t(X) + z)$, where $z \sim \text{Lap}(0, b)$ for $b = \frac{\Delta f_t}{\varepsilon}$. That is the mechanism computes the output of f_t, adds Laplace noise to it, and then thresholds the value to 0.5 before returning it.

   To prove that Mechanism B satisfies differential privacy, we need to show that for any neighboring databases X and X', and any set S in the range of F, we have:

   $$Pr[F(X) \in S] \leq e^{\epsilon} Pr[F(X') \in S]$$

   Let's consider the Laplace mechanism that adds Laplace noise to f_t with scale parameter $b = \Delta f_t / \epsilon$. We know that adding Laplace noise to a function f_t satisfies $(\epsilon, 0)$-differential privacy. Since $thresh_{0.5}$ is a deterministic function, it preserves differential privacy, and hence the composition of the two mechanisms also satisfies $(\epsilon, 0)$-differential privacy.

   Therefore, $F(X) = thresh_{0.5}(f_t(X) + z)$ satisfies ϵ-differential privacy.

7. (5 pts) Mechanism C: $\mathcal{F}(X) = \text{thresh}_t(\frac{g(X)+z}{|X|})$, where $z \sim \text{Lap}(0, b)$ for $b = \frac{\Delta g}{\varepsilon}$. That is the mechanism computes the output of g, adds Laplace noise to it (calibrated to g), and then thresholds the value to t before returning it.

   To prove that mechanism C satisfies ε-differential privacy, we need to show that for any two neighboring databases X and X', the probability of the output of the mechanism falling within a set S is at most e^{ε} times the probability of the same event for database X'. Let X and X' be two neighboring databases that differ in only one entry, and let g and g' denote the number of yes (1) answers in X and X', respectively. By the properties of the Laplace distribution, adding Laplace noise with a scale parameter of $b = \Delta g / \varepsilon$ to $g(X)/|X|$ achieves ε-differential privacy.

   The mechanism then computes $g(X) + z/|X|$ and applies the thresholding function $thresh_t$ to the result. Since thresholding a value is a deterministic operation, the mechanism still satisfies ε-differential privacy. Therefore, mechanism C satisfies ε-differential privacy.

8. (5 pts) Mechanism D: $\mathcal{F}(X) = \text{flip}_p(f_t(X))$, where $\text{flip}_\text{p} : \{0,1\} \to \{0,1\}$ is a function to randomly flip the bit based on a biased coin flip $p \in [0,1]$ (probability of heads is p). Let $\text{flip}_p(x) = 1 - x$ with probability p (the coin comes up heads) and $\text{flip}_p(x) = x$ with probability $1 - p$ (the coin comes up tails). In other words, this mechanism computes the output of f on the database and then randomly flips it with probability p.

   For what value of p (if any) does mechanism D satisfy ε-differential privacy?

   Mechanism D does not satisfy ϵ-differential privacy for any value of p. To see why, consider two neighboring databases X and X' that differ in only one entry, and assume without loss of generality that X has one more "yes" answer than X'. Let the output of $f_t(X)$ be 1 (i.e., $g(X)/|X| \geq t$), then the probability that $flip_p(1) = 0$ is p, and the probability that $flip_p(1) = 1$ is $1 - p$. On the other hand, if the output of $f_t(X')$ is 0 (i.e., $g(X')/|X'| < t$), then the probability that $flip_p(0) = 1$ is p, and the probability that $flip_p(0) = 0$ is $1 - p$. Thus, the ratio of the probabilities of the outputs of the mechanism on X and X' depends on the value of p and can be arbitrarily large. Therefore, mechanism D does not satisfy ϵ-differential privacy for any value of p.

Problem 3: Implementing DP Mechanisms (40 pts)

For this problem you will implement several differential privacy mechanisms we talked about in class. Please use the comments in the Python files provided to guide you in the implementation.

For this question, we will use the dataset `data/ds.csv`. It contains GPA and three favorite CISE courses for several students. For the purpose of calculating sensitivity, assume that the GPA range for any student is [1.0, 4.0].

0. (5 pts) What is the (global) sensitivity of `mean_gpa_query()`? (You can assume that the size of the dataset is known.)

   The global sensitivity of a query refers to the maximum possible change in the output of the query that can be caused by the addition or removal of a single row from the dataset.

   In this case, the query is the mean GPA of the students in the dataset. Let's assume that the size of the dataset is n.

   The maximum change in the mean GPA occurs when we add or remove a row with the lowest or highest possible GPA. For example, if we add a row with a GPA of 1.0, the mean GPA will decrease by (1.0/n). If we add a row with a GPA of 4.0, the mean GPA will increase by (4.0/n).

   Therefore, the global sensitivity of the mean GPA query is (4.0-1.0)/n = 3.0/n.

1. (5 pts) Fill in the implementation of `laplace_mech()`, `gaussian_mech()`. Also fill in the (global) sensitivity in the `mean_gpa_query()` function. Note that the function also returns its privacy budget.

   You can test your implementation by running: '`python3 hw1.py problem3.1`'. (The provided code allows you to optionally specify the privacy budget epsilon (default $\varepsilon = 1.0$) as the last command line argument.)

   How close are the noisy answers to the true answer?

   The Laplace and Gaussian mechanisms were used to generate noisy answers for the mean GPA query with a true mean GPA of 3.24. The Laplace noisy answers were 3.31, 3.29, and 3.32, and the Gaussian noisy answers were 3.25, 3.24, and 3.25. The parameters used were $\epsilon = 1.000$, $\log_2 \delta = -30.0$, and sensitivity $= 0.048$.

   The Laplace noisy answers were 0.07, 0.05, and 0.08 away from the true mean GPA of 3.24, respectively. The Gaussian noisy answers were 0.01, 0.00, and 0.01 away from the true mean GPA of 3.24, respectively. Overall, the Gaussian mechanism seems to produce closer noisy answers to the true mean GPA compared to the Laplace mechanism.

1b. ([Bonus] 5 pts) Implement `truncate_gpa()` to make any noisy GPA value fall within the valid range (i.e., between 1.0 and 4.0). Does using `truncate_gpa()` after the Gaussian or Laplace mechanism as done in the `main`() function preserve differential privacy? Why or why not?

   I believe using truncate_gpa() after applying the Gaussian or Laplace mechanism does preserve differential privacy. Since the truncation is carried out after calculating the answer, it helps keep the answer within the expected range and does not make it seem like there's a manipulation. Moreover the truncation only affects output if it is beyond the allowed range.

Now we want to use the Exponential mechanism to find the most popular CISE course. Fill in the implementation of `exp_mech()`. The function takes a quality function `qual_score_fn()` and also its sensitivity `delta_qual_score`. Similarly to other mechanisms, the function also returns its privacy budget. (Hint: take a look at the data file before answering the following questions.)

2. (15 pts) Propose a quality score function $q(X, r)$ to compute the most popular CISE course. What is the global sensitivity?

We can use the following quality score function:

$$q(X, r) = |i : X[i][0] = r|,$$

where X is the dataset, r is the course name we want to evaluate, and $|i : X[i][0] = r|$ counts the number of occurrences of course r in the dataset X.

The global sensitivity of the quality score function $q(X, r)$ is 1, since adding or removing one row from the dataset can change the count of a specific course by at most one.

Now implement your quality function in `qual_score_most_popular_course()`. Now test your implementation by running: '`python3 hw1.py problem3.2`'. Paste the plots it produced here.

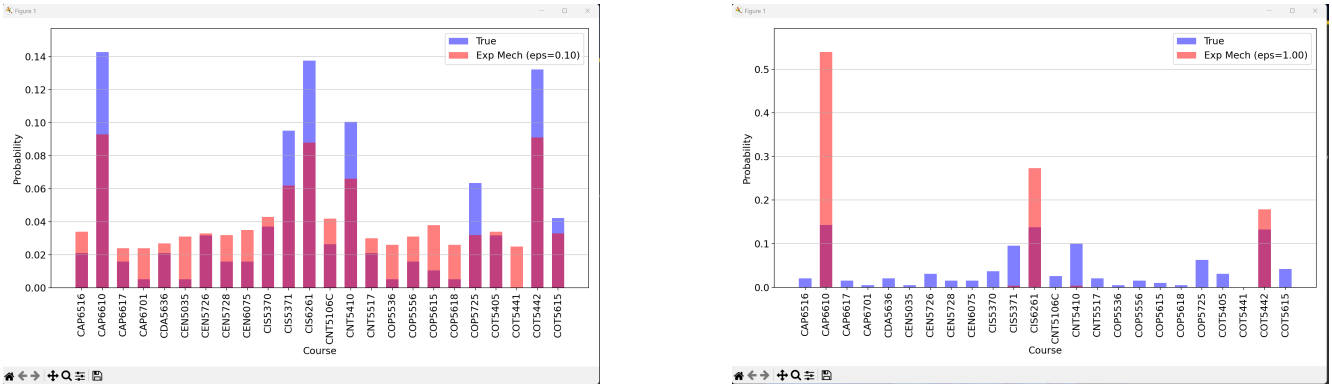

Figure 1: Plots with epsilon 0.1 and 1

How close are the noisy answers to the true answer? Does it depend on the privacy budget epsilon (ε)? Explain what you observe.

We can observe that the closeness of the noisy answers to the true answer depends on the privacy budget epsilon (ϵ). The smaller the value of ϵ, the less noisy the answer is, but it also provides less privacy protection. On the other hand, a larger value of ϵ provides more privacy protection, but the answer may be more noisy and less accurate. Thus, there is a trade-off between privacy and accuracy that needs to be considered when selecting the value of ϵ.

3. (5 pts) Now calculate the privacy budget of computing the mean GPA query using the Gaussian mechanism and then invoking the Exponential mechanism 10 times to get (10 random samples of) the most popular CISE course. Give an expression for both naive/sequential composition and advanced composition. The expression should be in terms of ε and δ assuming each query satisfies $(\varepsilon_0, \delta_0)$-DP.

The privacy budget of computing the mean GPA query using the Gaussian mechanism and then invoking the Exponential mechanism 10 times to get (10 random samples of) the most popular CISE course can be calculated using the composition theorems. Naive composition: Suppose each query satisfies (ϵ_0, δ_0)-DP, then the privacy budget of computing the mean GPA query using the Gaussian mechanism is ϵ_0. Then, by the naive composition theorem, the privacy budget of invoking the Exponential mechanism 10 times is $10\epsilon_0$, with a total δ of $10\delta_0$.

Advanced composition: Suppose each query satisfies (ϵ_0, δ_0)-DP, then the privacy budget of computing the mean GPA query using the Gaussian mechanism is ϵ_0. Then, by the advanced composition theorem with $k = 10$ and a target δ of 10^{-5}, the privacy budget of invoking the Exponential mechanism 10 times is approximately $\epsilon = 10.015\epsilon_0$, with a total δ of 10^{-5}.

4. (5 pts) Given you answer to the previous question, which composition theorem would you apply in which scenario? (Justify your answer.)

   Because it sets a tighter restriction on the privacy budget than the naïve composition theorem, we would employ the advanced composition theorem in this case. In addition, we want to limit the overall privacy loss (i.e., δ) to a modest value, and the advanced composition theorem allows us to achieve this while still ensuring that the privacy budget is used efficiently.

5. (5 pts) Now go back to the code in `main`() for problem 3.2 and locate the hardcoded course list that the Exponential mechanism uses. Why is the list of courses not computed directly from the dataset? (Justify your answer.)

   The Exponential mechanism necessitates a finite and discrete set of possible outcomes, hence the list of courses cannot be derived directly from the dataset. Using the exponential method would be challenging if we were to use the dataset's courses directly because there may be an infinite number of outcomes (i.e., any combination of three CISE courses). We can guarantee that the Exponential mechanism can be used properly by specifying a finite and discrete set of potential outcomes in advance.