COMP_SCI 397/497: Data Privacy (Winter 2022)

Homework 2 v1.1 (100 points)

Due date: February 17th, 2022, 3:29pm CST

Instructor: Sruti Bhagavatula

Your name: Hong Hong

Problem 1 Local differential privacy (45 points)

In this problem, you will get experience with implementing local differential privacy in a database and analyzing the related LDP guarantees.

Consider the following criminal records of a set of people.

Name	***Jail_Free***	***Type_of_crime***	***Jail_Free'***	***Type_of_crime'***
Liz Lemon	free	tax fraud	jail	1000
Patrick Star	jail	property theft	jail	0010
Ted Mosby	jail	DUI	jail	1000
Spongebob Squarepants	jail	DUI	jail	0001
Kumar Patel	free	property theft	jail	0100
Naruto Uzumaki	jail	shoplifting	jail	0100
Harold Lee	jail	property theft	jail	0100
Eugene Krabs	jail	tax fraud	jail	1000
Manabu Yukawa	free	shoplifting	free	0010

The columns `Jail_free` and `Type_of_crime` are sensitive attributes that the users want to protect. These columns contain their true sensitive attributes. The last two columns should contain the reported values of the first two columns according to the two randomized mechanisms described below.

Randomized response for reporting `Jail_Free`: Each user flips a private biased coin where the probability of flipping a Heads is p and the probability of Tails is $1-p$. If the user flips a Heads, they report their true value. If they flip a Tails, they flip a second private coin: if the second coin is Heads, they report "jail"; if the second coin is Tails, they report "free".

True counts of values of the sensitive attributes can be estimated according to the approach demonstrated in Lecture 6 (see page 9 of slides w/ notes).

Unary encoding for reporting `Type_of_crime`: Unary encoding is another type of randomized response mechanism which is suitable for multi-class categorical variables. Each user will encode their attribute value according to the Basic RAPPOR mechanism described in the paper by Erlingsson et al [1]. The steps for this mechanism are the following:

1. **Encode:** Represent each user's value of the `Type_of_crime` column by a one-hot encoding. A one hot encoding converts a multi-category categorical variable into multiple binary variables. For example, if a variable "color" has three possible values "red", "green", "purple", a one-hot encoding of this attribute value for a user might look like the following vector of three binary values of bits: $B = [color == red, color == green, color == purple]$. Here, the first element or bit indicates whether the their column value is red, the second bit indicates whether their column value is green, and the third bit indicates whether their column value is purple. Only one of these elements can be 1 and indicate a true value.

2. **Perturb:** Perturb each bit in the one-hot vector for each user generated in the previous step by flipping some of the bits. The probability that a bit gets flipped is determined by two probabilities q and p as follows:

$$Pr[B'[i] = 1] = \begin{cases} q, & \text{if } B[i] = 1 \\ p, & \text{if } B[i] = 0 \end{cases}$$

   Note that the mechanism determining P[B'[i] = 0] is symmetric to the above.

3. **Aggregate:** An analyst can estimate a count of each of the possible values of the sensitive attribute in the following way (i represents each possible value of the sensitive attribute, j represents each user's response, and n represents the number of user responses):

$$A[i] = \frac{(\sum_j B'_j[i]) - np}{q - p}$$

Exercise 1. [10 points] Assume you are locally implementing LDP for each user's `Jail_free` column value according to the randomized response mechanism above.

(a) (**8 pts**) Report a noisy version of each person's value of their `Jail_Free` column and enter these values in the `Jail_Free'` column. Consider p = 0.75. You may use the online python script here to generate unfair coin flips according to these probabilities.

(b) (**2 pts**) Explain in a few lines your process for reporting one user's `Jail_free` value.

Answer: Since the probabilities of each coin toss experiment are independent, I perform a coin toss on each column. If the first toss is Head, I report the true value, then move to next column. Otherwise, I do the second toss on the same column, and write the answer according to the result.

Exercise 2. [10 points] Assume you are locally implementing LDP for each user's `Type_of_crime` column value according to the unary encoding mechanism above.

(a) (**8 pts**) Report a noisy encoding of each person's value of their `Type_of_crime` column after the "perturb" step and enter these values in the `Type_of_crime'` column. This noisy encoding should be a 1-hot vector which you can enter in the table cell in the format: $[< is\_tax\_fraud >, < is\_property\_theft >, < is\_DUI >, < is\_shoplifting >]$. Consider $q = 0.75$ and $p = 0.25$. You may use again use the script here to generate unfair coin flips according to these probabilities to determine when to flip bits.[1]

[1] Flip an independent coin when deciding to flip each bit (i.e., run the script fresh each time).

(b) (**2 pts**) Explain in a few lines your process for reporting one user's `Type_of_crime` encoding.

Answer: I toss a coin for one bit each time, whenever B[i] equals to 1 or 0, that bit will be flipped when I toss a Tail. And for each column, I toss until the encoding become an one-hot vector.

Exercise 3. [14 points] Now put on your data analyst hat, and assume you are only given the two columns `Jail_free'` and `Type_of_crime'` containing noisy reports of users' data. You are aware of the mechanism that was used to add noise to the users' sensitive data even if you can't determine individual users' attributes.

(a) (**4 pts**) You want to count the people who are currently in jail. Provide an estimate of this count (not proportion) below and show your work.

Answer:
Noted that p is the true portion of people that who are currently in jail
$P(Jail\_Free' = \text{“}jail\text{”}|Jail\_Free = \text{“}jail\text{”}) = 0.75 + 0.25 * 0.75 = 0.9375$
$P(Jail\_Free' = \text{“}jail\text{”}|Jail\_Free = \text{“}free\text{”}) = 0.25 * 0.75 = 0.1875$
$P(\text{“}Jail\text{”}) = p * P(\text{“}jail\text{”}|\text{“}jail\text{”}) + (1-p) * P(\text{“}jail\text{”}|\text{“}free\text{”}) = 0.75p + 0.1875$
$p = \frac{4}{3} * P(\text{“}jail\text{”}) - \frac{1}{3}$
$\because P(\text{“}jail\text{”}) = \frac{8}{9}$
$\therefore p = \frac{23}{27}$
$\therefore number\ of\ people\ in\ jail = p * (total\ number\ of people) = \frac{23}{27} * 9 = 7$

(b) (**4 pts**) You want to determine the proportion of people who committed either tax fraud or property theft. Provide an estimate of this proportion below and show your work.

Answer:
$\because A[i] = \frac{(\sum_j B'_j[i]) - np}{q-p}$
$\therefore A[1] = \frac{6-18*0.25}{0.75-0.25} = 3$
$\therefore the\ proportion\ = \frac{3}{9} = \frac{1}{3}$

(c) (**3 pts**) How accurate is your estimate of the count of people in jail (i.e., how close to the true value is your estimate)? If the values are close, why do you think that is? If you are not satisfied with the accuracy, why do you think the accuracy might be low?

Answer: The true value is 6, my estimate of the count is 7, which I am not satisfied with. I think it is because the data points in this dataset is not enough.

(d) (**3 pts**) How accurate is your estimate of the proportion of people who committed either tax fraud or property theft (i.e., how close to the true value is your estimate)? If the values are close, why do you think that is? If you are not satisfied with the accuracy, why do you think the accuracy might be low?

Answer: The true value of proportion is $\frac{5}{9}$, my estimate is not close to the true value. Even though the bit is to flipped or not largely depends on the coin and the previous state of the bit, after so many times toss to form an one-hot encoding, the aggregate method is not going to estimate the true proportion of the sensitive attribute.

Exercise 4. [11 points] Analyze privacy guarantees.

(a) (**5 pts**) What local differential privacy guarantee can each user be sure their `Jail_free` attribute has? In other words, what is the minimal value of ϵ for which the randomized response mechanism is LDP? Show your work.

Answer:
$\because P(Jail\_Free' = \text{“}jail\text{”} | Jail\_Free = \text{“}free\text{”}) = 0.25 * 0.75 = 0.1875$
$\because P(Jail\_Free' = \text{“}jail\text{”} | Jail\_Free = \text{“}jail\text{”}) = 0.75 + 0.25 * 0.75 = 0.9375$
$\because P(Jail\_Free' = \text{“}free\text{”} | Jail\_Free = \text{“}jail\text{”}) = 0.25 * 0.25 = 0.0625$
$\because P(Jail\_Free' = \text{“}free\text{”} | Jail\_Free = \text{“}free\text{”}) = 0.75 + 0.25 * 0.25 = 0.8175$
$\because \frac{P(Y=y|X=x)}{P(Y=y|X=x')} \leq e^{\epsilon}$
$\therefore \epsilon \geq \ln \frac{0.9375}{0.0625} = \ln 15 = 2.708$
$\therefore the\ minimum\ value\ of \epsilon\ is\ 2.708$

(b) (**3 pts**) What local differential privacy guarantee can each user be sure their `Type_of_crime` attribute has? In other words, what is the minimal value of ϵ for which the unary encoding mechanism is LDP? Show your work. (Hint: see Section 3.2 of the RAPPOR paper: last equation in the first column of the page; consider h = 1)

Answer:
$\because q^* = P(S_i = 1 | b_i = 1) = 0.75$
$\because p^* = P(S_i = 1 | b_i = 0) = 0.25$
$\because \epsilon_1 = h \log(\frac{q^*(1-p^*)}{p^*(1-q^*)}) = 1 * \log(9) = 0.954$

(c) (**3 pts**) As an analyst, you've computed both the count of people in jail and the proportion of people who've committed either tax fraud or property theft. What differential privacy guarantee do the users' sensitive attributes have given that an analyst may query both the `Jail_free` and `Type_of_crime` columns, compute statistics on each column, and compute a statistic based on the first two statistics? In other words, what is the minimal value of ϵ for which the composition of the two mechanisms is LDP? Show your work.

Answer:
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 1 | Jail\_Free = \text{“}free\text{”}, b_i = 1) = 0.1875 * 0.75 = 0.1406$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 1 | Jail\_Free = \text{“}jail\text{”}, b_i = 1) = 0.9375 * 0.75 = 0.7031$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 1 | Jail\_Free = \text{“}jail\text{”}, b_i = 1) = 0.0625 * 0.75 = 0.0469$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 1 | Jail\_Free = \text{“}free\text{”}, b_i = 1) = 0.8175 * 0.75 = 0.6131$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 1 | Jail\_Free = \text{“}free\text{”}, b_i = 0) = 0.1875 * 0.25 = 0.0469$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 1 | Jail\_Free = \text{“}jail\text{”}, b_i = 0) = 0.9375 * 0.25 = 0.2344$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 1 | Jail\_Free = \text{“}jail\text{”}, b_i = 0) = 0.0625 * 0.25 = 0.0156$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 1 | Jail\_Free = \text{“}free\text{”}, b_i = 0) = 0.8175 * 0.25 = 0.2044$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 0 | Jail\_Free = \text{“}free\text{”}, b_i = 1) = 0.1875 * 0.25 = 0.0469$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 0 | Jail\_Free = \text{“}jail\text{”}, b_i = 1) = 0.9375 * 0.25 = 0.2344$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 0 | Jail\_Free = \text{“}jail\text{”}, b_i = 1) = 0.0625 * 0.25 = 0.0156$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 0 | Jail\_Free = \text{“}free\text{”}, b_i = 1) = 0.8175 * 0.25 = 0.2044$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 0 | Jail\_Free = \text{“}free\text{”}, b_i = 0) = 0.1875 * 0.75 = 0.1406$
$\because P(Jail\_Free' = \text{“}jail\text{”}, S_i = 0 | Jail\_Free = \text{“}jail\text{”}, b_i = 0) = 0.9375 * 0.75 = 0.7031$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 0 | Jail\_Free = \text{“}jail\text{”}, b_i = 0) = 0.0625 * 0.75 = 0.0469$
$\because P(Jail\_Free' = \text{“}free\text{”}, S_i = 0 | Jail\_Free = \text{“}free\text{”}, b_i = 0) = 0.8175 * 0.75 = 0.6131$
$\because \frac{P(Y=y|X=x)}{P(Y=y|X=x')} \leq e^{\epsilon}$
$\therefore \epsilon \geq \ln \frac{0.7031}{0.0156} = \ln 45 = 3.807$
$\therefore the\ minimum\ value\ of \epsilon\ is\ 3.807$

Problem 2 Private machine learning (35 points)

In this problem, you will be implementing privacy attacks on machine learning and building privacy-preserving machine learning models.

You be using IBM's two python libraries: `ai-privacy-toolkit` and `diffprivlib`. The paper for diffprivlib can be found here. The GitHub repositories for each library can be found here and here.

The exercises in this problem will be based on two Jupyter notebooks:

- https://github.com/IBM/differential-privacy-library/blob/main/notebooks/logistic_regression.ipynb
- https://github.com/IBM/ai-privacy-toolkit/blob/main/notebooks/membership_inference_diffpriv_nursery.ipynb.

Downloading the Jupyter notebooks is not required. You can either compile the snippets into your own `.py` file or use the notebook itself. You will need to save and submit all code written for this exercise whether built from the notebook or implemented on your own (see submission instructions at the end of the handout).

You will implement your code for the two sub-problems below in `python3`. You can install the python libraries by running the following commands:

```
> pip install diffprivlib
> pip install ai-privacy-toolkit
```

Note that depending on your setup, you may need to run `pip3 install...` to install the libraries to use with `python3`.

2.1 Differentially-private classifiers (22 points)

Refer to the Jupyter notebook here: https://github.com/IBM/differential-privacy-library/blob/master/notebooks/logistic_regression.ipynb.

Exercise 5. [2 points] As implemented in the Jupyter notebook in cell `[10]:`, when training the differentially private logistic regression model, it is shown that by setting their ϵ value equal to `float(inf)`, they can "produce the same result as the non-private logistic regression classifier". Assuming setting ϵ to `inf` is allowed, why does this happen? Are there any DP guarantees in this situation?

Answer: Because higher epsilon means more privacy loss, when the epsilon is inf, which means there are no privacy, and there is no DP guarantees in this situation.

Exercise 6. [20 points] Now you will train a new model on a different dataset and evaluate how a differentially private classifier compares to a regular classifier on this dataset. The dataset you will use for this is the Breast Cancer Wisconsin dataset. It is included with `scikit-learn` and can be initialized as below:

```
from sklearn import datasets
dataset = datasets.load_breast_cancer()
```

(a) (**7 pts**) Implement both a non-private classifier using `scikit-learn` and a private classifier using `diffprivlib`. Refer to the following resources for help using this dataset:

- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html
- https://www.kaggle.com/leemun1/predicting-breast-cancer-logistic-regression

(b) (**8 pts**) Choose a range of ϵ starting at a value less than 1 and ending above 1, and plot a graph comparing the accuracy of the above two classifiers. A similar graph and accompanying code can be found near the bottom of the example `diffprivlib` notebook. Save your graph (submission instructions at the end).

(c) (**5 pts**) Describe in a few sentences your interpretation of your graph. Specifically, how do the two accuracies compare and how does this relate to ϵ? Note: this will be different for every student depending on the range of ϵ chosen.

Answer: The non-private model can always perform high accuracy, which is the baseline(non-private) in the graph. However, the accuracy of private classifier is always less than baseline, which means the classifier protects the privacy. We can see from the graph that lower epsilon has higher privacy, higher epsilon means more privacy loss, cause the accuracy of private classifier in the graph becomes closer to the baseline when epsilon higher. The discontinuty observed near epsilon = 10 is an artifact of the model. Because of the norm-clipping applied to the dataset before training (data norm=100), the accuracy plateaus without reaching the non-private baseline.

2.2 Defending against membership inference attacks (13 points)

Refer to the Jupyter notebook here: https://github.com/IBM/ai-privacy-toolkit/blob/main/notebooks/membership_inference_diffpriv_nursery.ipynb.

Exercise 7. Train a regular scikit-learn logistic regression classifier on the same Breast Cancer Wisconsin dataset from the previous problem.

(a) (**3 pts**) Execute a black-box membership inference attack on the logistic regression model you trained. How effective is the attack in terms of how much better or worse it is than a random coin flip?

> **Answer:** For 60.7% of the data, membership status is inferred correctly, just a little bit better than random coin filp

(b) (**3 pts**) Build a privacy-preserving model by training a differentially-private version of the logistic regression model (as you did in the previous problem) with $\epsilon = 2$. Is this model more resilient to membership inference attacks? If so, by how much (in terms of change in attack accuracy)?

> **Answer:** No, the differentially-private version of my model is not more resilient to attacks in terms of change in the overall accuracy.

(c) (**7 pts**) Using the same approach as at the bottom of the notebook, what is an optimal value of ϵ where the model accuracy is high and the attack accuracy is low?

> **Answer:** The optimal value of epsilon is 100 in my program.

Problem 3 Evaluating fairness (20 points)

A long time ago, nine celebrities were caught protesting and rioting in Evanston. The police came to the scene and arrested all nine celebrities. Each celebrity then went through a trial with a judge and a prosecutor and was found guilty or not guilty. The following table has information about each celebrity's demographics, whether they've had previous felonies, and whether they were found guilty or not guilty.

name	***gender***	***age***	***prev_felonies***	***guilty***
Bruce Willis	Male	45	Yes	Yes
Zooey Deschanel	Female	3	Yes	No
Andy Samberg	Male	90	No	Yes
Kal Penn	Male	56	No	Yes
Alec Baldwin	Male	67	Yes	Yes
Tina Fey	Female	89	Yes	Yes
Steve Carell	Male	26	No	No
John Cho	Male	74	No	Yes
Will Smith	Male	15	No	No

Following the big celebrity riot, the court got tired of bringing each person newly arrested for rioting in for a full trial. Therefore, to reduce their workload, the court decided to not have a trial

for each arrested person but rather to use the outcomes of the famous celebrity riot to automatically predict if someone should be found guilty or not. To do this, they built a machine learning model using the above table as training data to predict if a new arrestee is **guilty** based on the three features (**gender**, **age**, **prev_felonies**).

The following table consists of predictions made by the model on six ordinary citizens arrested for rioting.

name	***gender***	***age***	***prev_felonies***	***guilty (true)***	***guilty (pred)***
Sruti Bhagavatula	Female	11	Yes	No	Yes
Olivia Escousse	Female	87	No	Yes	Yes
Misha Ivaniuk	Male	58	No	No	No
Ronald McDonald	Male	1	Yes	No	Yes
Sonic the hedgehog	Male	90	Yes	Yes	No
Princess Peach	Female	25	No	No	No

Exercise 8. [4 points] What can the protected attribute(s) here be?

Answer: name, gender, age

Exercise 9. [4 points] For the protected attribute you determined in the previous step (select one if you said there were multiple), determine the privileged and unprivileged group based on the training data. In this instance, you can do this by looking at what traits the people who were found "not guilty" in the training data often had in common. Describe the condition or rule on the attribute value you determined that separates people into the two groups.

Answer: I think for the people who were found "not guilty", they were always much more younger, for their age were all under 30. The people who were found "guilty" were always much more older. And for the name and gender, I can not find some specific characteristics.

Exercise 10. [12 points] Is the classifier fair with respect to any of the four group fairness metrics we learned about in class? Which one(s)? Show that the classifier is fair for each metric that is satisfied and show that it is not fair for the remaining metrics. (Remember that, in contrast to the other metrics, satisfying disparate impact implies unfairness and that *not* satisfying disparate impact implies fairness.)

Answer:
priviledged group: age under 30
1)disparate imapct:
$\because P(Y'=1|X \notin S) = \frac{2}{3} \because P(Y'=1|X \in S) = \frac{1}{3} \therefore \frac{P(Y'=1|X \notin S)}{P(Y'=1|X \in S)} = 2 > 0.8$
$\therefore$ the classifier is fair for this metric
2)demographic parity:
$\because P(Y'=1|X \notin S) \neq P(Y'=1|X \in S)$
$\therefore$ the classifier is not fair for this metric
3)equal opportunity:
$\because P(Y'=1|X \notin S, Y=1) = 1 \because P(Y'=1|X \in S, Y=1) = \frac{1}{3}$
$\therefore P(Y'=1|X \notin S, Y=1) \neq P(Y'=1|X \in S, Y=1)$
$\therefore$ the classifier is not fair for this metric
4)equalized odds:
we could see from the above, this metric is not fair

Submission instructions

You will need to submit two sets of files to Canvas under Homework 2 (you can submit multiple files by selecting "Add another file"):

1. One PDF file (named `<name>_<netID>_hw2.pdf`) with answers to problems 1, 2 (parts of it), and 3. You can add your answers to this PDF in the provided answer spaces (recommended), in the LaTex template provided (recommended), or a separate PDF with only answers.

2. A zip file (named `<name>_<netID>_hw2.zip`) containing:

   - The code you wrote for Problem 2.1 (named `<name>_<netID>_prob_2.1.py`)
   - The code you wrote for Problem 2.2 (named `<name>_<netID>_prob_2.2.py`)
   - The graph you generated for Exercise 6(b) (named `<name>_<netID>_graph_6b.py`)
   - The model and attack accuracy graphs you generated for Exercise 7(c) (named `<name>_<netID>_graph_7c-model.py` and `<name>_<netID>_graph_7c-attack.py`)

References

[1] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *ACM SIGSAC Conference on Computer and Communications Security*, 2014.