Project 2: Building a Classifier with Differential Privacy

Duc Viet Le

April 28, 2018

1 Written Tasks

Question 1: Suppose that all the n values are in the range $[a, b]$, what is the sensitivity for the following functions? (1) sum, (2) mean, (3) median

Answer:

(a) **Sum:** In bounded differential privacy, one can replace extract one element with other element to obatin two neighbor datasets. Thus, one can replace the smallest value with the largest value; we have the sensitivity value for max computed as follow:

$$\Delta_f = \max_{D \simeq D'} |f(D) - f(D')| = |(b + \Sigma_1^{n-1} e_i) - (a + \Sigma_1^{n-1} e_i)| = b - a \tag{1}$$

Where $\Sigma_1^{n-1} e_i$ denotes the sum of other $n-1$ elements.

(b) **Mean:** Similarly, we have the sensitivity value for mean compute as follow:

$$\Delta_f = \max_{D \simeq D'} |f(D) - f(D')| = \left| \frac{b + \Sigma_1^{n-1} e_i}{n} - \frac{a + \Sigma_1^{n-1} e_i}{n} \right| = (b-a)/n \tag{2}$$

Where $\Sigma_1^{n-1} e_i$ denotes the sum of other $n-1$ elements.

(c) **Median:**
- *n is odd:* Then, if we sort n values, the median will be element $(n+1)/2$. Thus, Let $D = \{a, ..., a, b, ..., b\}$ where first $(n+1)/2$ elements are a and the rest are b. Thus, we obtain D' by replace the $(n+1)/2$ element with b. It is not difficult to see that the median of D is a, and the median of D' is b. Therefore, we have the sensitivity value for median computed as follow:

$$\Delta_f = \max_{D \simeq D'} |f(D) - f(D')| = |a - b| = b - a \tag{3}$$

- *n is even:* Consider a dataset D that first $n/2 + 1$ elements are a and last $n/2 - 1$ elements are b. We obtain D' by replace the $(n/2 + 1)$-th with value b. It's not difficult to see that median of D is a while the median of D' is $(a + b)/2$. Therefore, we have the sensitivity value for median computed as follow:

$$\Delta_f = \max_{D \simeq D'} |f(D) - f(D')| = |a - (b+a)/2| = (b-a)/2 \tag{4}$$

Since we only consider the max, then the sensitivity is $b - a$

Question 2: Suppose that all the n values are either a or b. Suppose your DP algorithm outputs the number of a values and b values, respectively (with Laplace mechanism under privacy budget ϵ), and you now calculate the mean of these n values. What is the variance of this estimation?

Answer: The sensitivity is computed as the L_1 norm of two vector. Since the global sensitivity for a count function is 1, in this case we have two count functions. If we replace one element with other element, one count will decrease by 1 and the other will increase by 1, thus we have the sensitivity:

$$\Delta_{\mathsf{count}} = \max_{D \simeq D'} ||f(D) - f(D')||_1 = 2 \tag{5}$$

Consider the function $\mathsf{count}(D)$ thats output a vector $(\mathsf{c_a}, \mathsf{c_b})$ where $\mathsf{c_a}, \mathsf{c_b}$ denotes the number of a and b in D, respectively. We apply Laplace mechanism to obtain:

$$A_{\mathsf{count}}(D) = \mathsf{count}(D) + \langle X_1, X_2 \rangle = \langle c_a + X1, c_B + X_2 \rangle$$

Where X_1, X_2 are random variable from $\mathsf{Lap}(\Delta_{\mathsf{count}}/\epsilon) = \mathsf{Lap}(2/\epsilon)$.
We have an estimation for the mean:

$$\mathsf{EstMean}(A_{\mathsf{count}}(D)) = ((c_a + X_1) \cdot a + (c_b + X_2) \cdot b)/n = (c_a \cdot a + c_b \cdot b)/n + (aX_1 + bX_2)/n \tag{6}$$

Thus we have the variance:

$$\begin{aligned}
\mathsf{Var}(\mathsf{EstMean}(A_{\mathsf{count}}(D)) &= \mathsf{Var}((c_a \cdot a + c_b \cdot b)/n + (aX_1 + bX_2)/n) \\
&= \mathsf{Var}(aX_1 + bX_2)/n^2 \\
&= (\mathsf{Var}(aX_1) + \mathsf{Var}(bX_2))/n^2 \\
&= (a^2\mathsf{Var}(X_1) + b^2\mathsf{Var}(X_2))/n^2 \text{ since } X_1, X_2 \text{ are idependent} \\
&= (8a^2/\epsilon^2 + 8b^2/\epsilon^2)/n^2 \text{ since } \mathsf{Var}(X_1) = \mathsf{Var}(X_2) = 8/\epsilon^2 \\
&= 8(a^2 + b^2)/(\epsilon n)^2
\end{aligned} \tag{7}$$

Question 3: Suppose that the n values each has two attributes, age and gender. You are going to publish a histogram (with Laplace mechanism and ϵ) of both attribute, with age bucketized into $[0-49]$ and $[50-100]$ (so there will be four numbers for: male-$[0-49]$, male-$[50-100]$, female-$[0-49]$, female-$[50-100]$). Now you want to estimate the number of male users, what is the variance of this estimation? What is the variance if you just use the gender attribute and ignore age when you publish the histogram? If each value has d binary attributes, what is the size of your histogram?

Answer: Similarly, in bounded DP, the sensitivity of $\mathsf{Count}(\cdot)$ is 2 because a decrement in one bin implies an increment in another bin. We have:

$$\Delta_{\mathsf{count}} = \max_{D \simeq D'} ||f(D) - f(D')||_1 = 2 \tag{8}$$

Consider the function $\mathsf{Count(D)} = \langle c_{m[0-49]}, c_{m[50-100]}, c_{f[0-49]}, c_{f[50-100]} \rangle$ that outputs four numbers for: male-$[0-49]$, male-$[50-100]$, female-$[0-49]$, female-$[50-100]$. We have the Laplace Mechanism as follow:

$$A_{\mathsf{count}}(D) = \langle c_{m[0-49]}, c_{m[50-100]}, c_{f[0-49]}, c_{f[50-100]} \rangle + \langle X_1, X_2, X_3, X_4 \rangle \tag{9}$$

Where X_1, X_2, X_3, X_4 are random variable from $\mathsf{Lap}(\Delta_{\mathsf{count}}/\epsilon) = \mathsf{Lap}(2/\epsilon)$

- Now you want to estimate the number of male users, what is the variance of this estimation?

$$\begin{aligned}
\mathsf{Var}(\mathsf{EstMale}(A_{\mathsf{count}}(D)) &= \mathsf{Var}(c_{m[0-49]} + X_1 + c_{m[50-100]} + X_2) \\
&= \mathsf{Var}(X_1) + \mathsf{Var}(X_2) \\
&= 8/\epsilon^2 + 8/\epsilon^2 = 16/\epsilon^2
\end{aligned} \tag{10}$$

- What is the variance if you just use the gender attribute and ignore age when you publish the histogram? The variance will just be:

$$\mathsf{Var}(\mathsf{EstMale}(A_{\mathsf{count}}(D)) = \mathsf{Var}(X) = 8/\epsilon^2 \tag{11}$$

- If each value has d binary attributes, what is the size of your histogram? The size of histogram will be: 2^d

Question 4: Suppose that all the n values are in the range $[a, b]$, and your task is to publish the 25th, 50th, and 75th percentiles (assume $n > 100$). Now you are given an algorithm that adds independent Laplace noise $\mathsf{Lap}\,(\beta_{25})$, $\mathsf{Lap}\,(\beta_{50})$, and $\mathsf{Lap}\,(\beta_{75})$, to the real answers, respectively ($\beta_{25} < \beta_{50} < \beta_{75}$). Your task is to find out (1) what is the sensitivity of this problem, (2) what is the final minimal ϵ this algorithm can achieve?

Answer:

- What is the sensitivity of this problem?
  Since we consider bounded, we can only replace elements. Consider a function that output 3 element 25th, 50th, 75th, in the worst case, only one element can change at either 25th, 50th, or 75th. Therefore, Let f to be the function that publishes the 25th, 50th, and 75th percentiles, then:

$$\begin{gathered}\Delta_f = \max_{D \simeq D'} ||f(D) - f(D')||_1 = |b - a| = b - a \\ \text{where 25th,50th percentiles are } a \text{ and 75th percentile is b in } D \text{ and } a \text{ in } D'\end{gathered} \tag{12}$$

- What is the final minimal ϵ this algorithm can achieve?
  Let $f(D) = \langle e_{25}, e_{50}, e_{75} \rangle$, $f(D') = \langle e'_{25}, e'_{50}, e'_{75} \rangle$. We define the Laplace mechanism as follow:

$$\mathbf{A}(D) = f(D) + \langle X_0, X_1, X_2 \rangle \tag{13}$$

Where $X_0, X_1.X_2$ are random variable drawn from $\mathsf{Lap}(\beta_{25}), \mathsf{Lap}(\beta_{50})$ and $\mathsf{Lap}(\beta_{75})$, respectively. Thus, we have:

$$\begin{aligned}\mathsf{Pr}\left[\mathbf{A}(D) = t\right] &= \mathsf{Pr}\left[f(D) + \langle X_0, X_1, X_2 \rangle = t\right] \\ &= \mathsf{Pr}\left[(X_0 = t_0 - e_{25}) \wedge (X_1 = t_1 - e_{50}) \wedge (X_2 = t_2 - e_{75})\right] \\ &= \mathsf{Pr}\left[X_0 = t_0 - e_{25}\right] \mathsf{Pr}\left[X_1 = t_1 - e_{50}\right] \mathsf{Pr}\left[X_2 = t_2 - e_{75}\right] \\ &= \frac{1}{8\beta_{25}\beta_{50}\beta_{75}} \exp(-|t_0 - e_{25}|/\beta_{25} - |t_1 - e_{50}|/\beta_{50} - |t_2 - e_{75}|/\beta_{75})\end{aligned} \tag{14}$$

Similarly, we have:

$$\mathsf{Pr}\left[\mathbf{A}(D') = t\right] = \frac{1}{8\beta_{25}\beta_{50}\beta_{75}} \exp(-|t_0 - e'_{25}|/\beta_{25} - |t_1 - e'_{50}|/\beta_{50} - |t_2 - e'_{75}|/\beta_{75}) \tag{15}$$

From (14),(15) we have:

$$\begin{aligned}\frac{\mathsf{Pr}\left[\mathbf{A}(D) = t\right]}{\mathsf{Pr}\left[\mathbf{A}(D') = t\right]} &= \exp\left(\frac{|t_0 - e'_{25}| - |t_0 - e_{25}|}{\beta_{25}} + \frac{|t_1 - e'_{50}| - |t_1 - e_{50}|}{\beta_{50}} + \frac{|t_2 - e'_{75}| - |t_2 - e_{75}|}{\beta_{75}}\right) \\ &\le \exp\left(\frac{|e_{25} - e'_{25}|}{\beta_{25}} + \frac{|e_{50} - e'_{50}|}{\beta_{50}} + \frac{|e_{75} - e'_{75}|}{\beta_{75}}\right) \text{ because of triangle inequality} \\ &\le \exp\left(\frac{||f(D) - f(D')||_1}{\beta_{25}}\right) \le \exp(\Delta_f/\beta_{25}) = \exp((b - a)/\beta_{25})\end{aligned}$$

Therefore, the minimal ϵ is $(b - a)/\beta_{25}$

Question 5: If there is no public dataset available, and you instead use 10% of your sensitive data, sampled randomly, to find the desired histogram, without differential privacy. The remaining 90% of data is used to calculate the exact values in each cell of the histogram (and then add Laplace noise $\mathsf{Lap}\left(\frac{1}{\epsilon}\right)$). What will be the worst case ϵ for the whole process?

Answer: Given dataset D, there are two algorithms $\mathbf{A}_1(\cdot)$ and $\mathbf{A}_2(\cdot)$. $\mathbf{A}_1$ takes as input the dataset D outputs T that contains some attributes and some dataset D_2 such that $D_2 \subset D$ and $|D_2|/|D| = .9$. The algorithm $\mathbf{A}_2$ uses D_2 and outputs the count for each attributes output by $\mathbf{A}_1$ and adds $\mathsf{Lap}(1/\epsilon)$ noise to each count. We want to compute the worst case for the sensitivity ϵ of the process $\mathbf{A}_2(\mathbf{A}_1(D))$ in bounded setting:

- Algorithm $\mathbf{A}_1$ samples 10% data uniformly at random, I assume that the algorithm is 0-DP on unbounded setting.
- Algorithm $\mathbf{A}_2$ applies the count and add $\mathsf{Lap}(1/\epsilon)$, this algorithm is ϵ-DP on unbounded setting

Using the sequential composition, $\mathbf{A}_2(\mathbf{A}_1(D))$ achieves ϵ-unbounded-DP. However, in this project we consider bounded DP, the $\mathbf{A}_2(\mathbf{A}_1(D))$ will achieve 2ϵ-bounded-DP because one can assume that we transform the data by remove 1 and add 1. Therefore, the worst case is 2ϵ.

2 Programming Part

- A brief report of the programming task is required. The report should describe your algorithm, the reason for your design, and the argument why your algorithm is ϵ-DP.

Answer:

Selecting Grid: The algorithm uses 6 attributes which are `age`,`capital-gain`, `education-num`, `marital-status`. Here are few methods that I used:

- ∗ **Semantic Meaning:** Based on the semantic meaning, I think it's reasonable to choose `capital-gain` because this attribute is highly correlated to the income.
- ∗ **Decision Tree:** For each of attributes, for example `education-num` and `marital-status`, I plot those attributes to find out which range are more likely to have income higher than 50k and divide those bins accordingly.

Finally, the smaller the number of attributes and bins, the lower the noise that we need to add.

Adding Laplace noise: I use the equation from wikipedia[a] to generate noise:

$$X = \mu - \beta \mathsf{sgn}(U) \ln(1 - 2|U|) \tag{16}$$

where U is the random variable drawn from the uniform distribution on the interval $(-.5, .5]$ In this work, $\mu = 0$ and $\beta = \Delta_f/\epsilon = 2/\epsilon$ (bounded-DP)

The algorihm is ϵ-DP: The algorithm satisfies ϵ-DP. The reason is that the algorithm output a histogram as an vector. Using result of theorem 2.15 (Laplace Mechanism, the vector case) from the book "Differential Privacy: From Theory to Practice", we can claim that by adding noise independently for each element of the vector, our algorithm still satisfies ϵ-DP

[a] `https://en.wikipedia.org/wiki/Laplace_distribution#Generating_random_variables_according_to_the_Laplace_distribution`

- Performance $\epsilon = .1$

```
python3 exp.py --epsilon=.1 --file=adult.data
```

```
1   0.8304620203602192
2   0.847821762482594
3   0.8417481540610656
4   0.8435496794871795
5   0.8337006611901423
6   0.8466561139691333
7   0.8361863057324841
8   0.8419580419580419
9   0.8451548451548452
10   0.8354579329280251
Average: 0.8402695517323728
```

- Performance $\epsilon = .5$

```
python3 exp.py --epsilon=.5 --file=adult.data
1   0.8484787830264211
2   0.8376921541225794
3   0.8287027783330002
4   0.8447862564922094
5   0.842546397924566
6   0.8413173652694611
7   0.8414123279473369
8   0.8472444089456869
9   0.8438936638017189
10   0.8426203315358498
Average: 0.841869446739883
```

- Performance $\epsilon = 1.$

```
python3 exp.py --epsilon=1.0 --file=adult.data
1   0.8451239008792966
2   0.8436378141204627
3   0.8447242206235012
4   0.8296725239616614
5   0.8517853580690206
6   0.8437874550539353
7   0.8421157684630739
8   0.8357970435477428
9   0.8457850579304834
10   0.8431450808221912
Average: 0.8425574223471368
```