CS 281 Homework 3

Instructor: Carlos Guestrin

May 14, 2025

Deliverables: Please submit **a single pdf** containing the written answers to questions, along with the code wherever applicable. Starter code can be found here.

1 ε-DP mean estimation

Assume we are given a database of medical records $X \in \{0,1\}^{n \times d}$ for n patients, where each patient's record X_i is a vector $(x_{i1}, ..., x_{id})$ of d binary variables $x_{ij} \in \{0,1\}$ which represents if patient i has medical condition j (e.g., condition $j = 1$ could be if they have diabetes, $j = 1$ hypertension, $j = 2$ chronic kidney disease, etc.). Suppose we want to calculate the feature prevalence vector $\mathbb{E}[X]$, which is a $d-$dimensional vector where the $j$the element denotes the average prevalence of condition j in the database. However, in order to protect patient privacy, we are interested in identifying an ε-DP mechanism for calculating $f(X) = \mathbb{E}[X]$. In other words, we want to create a mechanism $M(X)$ which approximates the prevalence vector $f(X) = \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i$, such that, for any two datasets X, X' which differ in exactly one data point,

$$\frac{P(M(X) \in T)}{P(M(X') \in T)} \leq \exp(\varepsilon)$$

where $T \subseteq \mathbb{R}^d$ is any set of possible prevalence vectors.

1.1 Intuition [3 points]

For better clarity, consider $|T| = 1$, i.e. T is any single vector in $\mathbb{R}^d$. Explain, in plain language, what the ε-DP mechanism would guarantee in this setting with our specific f and X. (1-2 sentences)

1.2 Univariate case [12 points]

Consider a simplified, univariate version of this problem, where you're trying to estimate the prevalence of diabetes (i.e., we only care about the feature $X_{\cdot,j=0}$)[1]. We will use the Laplace mechanism defined over this single variable

$$M(X_{\cdot,j}) = f(X_{\cdot,j}) + W$$

where $f(X_{\cdot,j}) = \mathbb{E}[X_{\cdot,j}]$ for $j = 0$ and

$$W \sim \text{Laplace}\Big(\frac{\Delta f}{\varepsilon}\Big)$$

Using the same notation as in lecture, recall that Δf is the sensitivity of our function f, and the goal is identify the parameters that guarantee a ε-DP mechanism $M(X_{\cdot,j=0})$.

(a). Sensitivity and noise [2 points]. Calculate Δf and state the distribution of $W \sim \text{Laplace}(\cdot)$ as function of n, d, ε.

[1] We use the notation $X_{\cdot,j}$ to denote the jth variable in X for all i patients. In `numpy` notation, this would be `X[:,j]` for example.

(b). Implementation. **[8 points]** Using the simulated dataset in the starter code as the dataset X and your answer for sensitivity from part (a), implement the Laplace mechanism for calculating prevalence of a single variable $X_{\cdot,j=0}$ for three different ε values: 0.01, 0.1 and 1. For each ε, calculated the mechanism's mean prevalence 1000 different times by using 1000 different seeds. Then plot the resulting distributions of mean estimates. In each plot, only visualize the x-axis between $[0.3, 0.6]$ and set the number of bins to be 30. Finally, plot the true condition mean in each plot as a vertical red line.

(c). Interpretation of plots. **[2 points]** How do the plots change across the values of ε? What does this reflect about the change in privacy and why? (2 sentences max)

1.3 Multivariate case [15 points]

Now consider a full version of this problem using all d features using the Laplace mechanism

$$M(X) = f(X) + \boldsymbol{W}$$

where $\boldsymbol{W}$ is now the $d-$dimensional (zero-indexed) vector $(W_0, ..., W_{d-1})$ and each W_j are independent Laplace random variables $W_j \sim \text{Laplace}\left(\frac{\Delta f}{\varepsilon}\right)$

(a). Sensitivity and noise [2 points]. Calculate Δf and state the distribution of $W_j \sim \text{Laplace}(\cdot)$ using variables n, d, ε.

(b). Implementation. **[6 points]** Now, using your answer for sensitivity above, modify the univariate Laplace mechanism to work as a multivariate Laplace mechanism for all $d = 20$ variables in X. Again, calculate the mechanism for three different ε values: 0.01, 0.1 and 1. For each ε, calculate the mechanism's mean prevalence 1000 different times by using 1000 different seeds. Then, plot the first 5 ($j \in [0, 1, 2, 3, 4]$) distributions for each of the first 5 features. In each plot, only visualize the x-axis between $[-0.5, 1.5]$ and set the number of bins to be 30.

(c). Interpretation of plots. **[2 points]** Compare your plots to those generated in part 1.2. Explain how the plots for the first disease (diabetes) differ in the multivariate case from the univariate case, and why this is the case. (2 sentences max)

(d) Intuition for sensitivity. **[1 point]** Explain, in plain terms, how the sensitivity Δf influences the choice of ε. (1 sentence)

(e). Ethical considerations. **[4 points]** Suppose you are tasked with releasing your results for $f(X)$, where X is the dataset of all patients' medical records at Stanford Hospital. Name two reasons why you might want to use a low ε and two reasons you might want to use a high ε. (Use 4 bullet points, one for each reason.)

2 (ε, δ)-DP mean estimation

Recall another variation of ensuring differential privacy is by identifying a (ϵ, δ)-DP mechanisms. Similarly to above, let datasets X, X' differ in exactly one data point, $M(X)$ a mechanism that maps into the d-dimensional space $\mathbb{R}^d$, and $T \subseteq \mathbb{R}^d$ is any set of possible output. Then, for a given ε and δ, the mechanism M is (ϵ, δ)-DP if it satisfies

$$P(M(X) \in T) \leq \exp(\varepsilon) P(M(X') \in T) + \delta$$

In class, we saw that the Gaussian mechanism satisfies (ϵ, δ)-DP. That is, for a given function $f(X)$, its ℓ_2 sensitivity $\Delta_2 f$, and choice of δ, ε, the Gaussian mechanism outputs

$$M(X) = f(X) + \boldsymbol{Y}$$

where $\boldsymbol{Y}$ is the d-dimensional (zero-indexed) vector $(Y_0, \ldots Y_{d-1})$, the variable $Y_j \sim \mathcal{N}(0, \sigma^2)$, and

$$\sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta_2 f)^2}{\epsilon^2}$$

(Note, this is equivalent to $\boldsymbol{Y} \sim \mathcal{N}(0, \sigma^2 I_d)$). For this question, again define f as the prevalence vector $f(X) = \mathbb{E}[X]$ over the medical dataset X with d features and n patients.

2.1 The Gaussian mechanism [14 points]

(a). Sensitivity and noise. **[4 points]** Calculate the sensitivity $\Delta_2 f$ and state the distribution of $Y_j \sim \mathcal{N}(\cdot)$, using variables $n, d, \varepsilon, \delta$.

(b). Implementation. **[6 points]** Now, using your answer for sensitivity above (and the same simulated dataset for X), implement the Gaussian mechanism for calculating prevalence of all $d = 20$ condition variables in X for three different ε values: 0.01, 0.1 and 1, and $\delta = 0.01$. For each ε, calculate the mechanism's mean prevalence 1000 different times by using 1000 different seeds. Then, plot the first 5 ($j \in [0, 1, 2, 3, 4]$) distributions for each of the first 5 features. In each plot, only visualize the x-axis between $[-0.5, 1.5]$ and set the number of bins to be 30.

(d). Comparison and interpretation. **[2 points]** Fix $\varepsilon = 0.1$, and compare the plots for the first five variables to those generated for the multivariate Laplace mechanism from part 1.3. How do the mechanisms differ, and what does this say specifically about each of the two mechanisms' privacy? (1-2 sentences)

(e). Connection to fairness. **[2 points]** Suppose we are told by Stanford Hospital that we need a privacy budget of $\varepsilon = 0.1$ (and, if applicable $\delta = 0.01$). If we also care about fairness with respect to some sensitive attribute, should we use the aforementioned Laplace mechanism or the Gaussian mechanism, and why?

2.2 Comparing ℓ_2 and ℓ_1 sensitivity [6 points]

One reason for the above differences between Laplace and Gaussian mechanisms might be due to the fact that the Laplace mechanism (and the corresponding ε-DP guarantee) uses the ℓ_1 norm for computing the sensitivity Δf while the Gaussian mechanism (and the corresponding (ε, δ)-DP guarantee) uses the ℓ_2 norm for computing the sensitivity $\Delta_2 f$. In this question, we'll explore how the choice of sensitivity affects the mechanisms.[2]

Fix $\varepsilon = 0.1$, $\delta = 0.01$ and consider the multivariate case for all $d = 20$ conditions (i.e., diabetes). Then, for only the first variable $X_{\cdot,j=0}$ plot 4 plots: (1) the original Laplace mechanism using Δf as sensitivity, (2) the modified Gaussian mechanism using Δf as sensitivity, (3) the modified Laplace mechanism using $\Delta_2 f$ as sensitivity, (4) original Gaussian mechanism using $\Delta_2 f$ as sensitivity. For all plots, limit the x-axis to be [0,0.6] and use 30 bins. Then explain (i) how the definition of sensitivity impacts the mechanism and (ii) how this impacts privacy. (2 sentences max)

[2] In practice, the choice of the sensitivity metric is usually imposed by the characteristics of downstream tasks. For example, learning a model using gradients typically works better with ℓ_2 sensitivity, while counts or histograms tend to do well with ℓ_1.