CIS 4230/5230 Homework 2: The Laplace Mechanism

Prof Michael Kearns

Spring 2024

Due Date: 11:59PM Tue April 16. Every student should work alone and submit their own notebook.

Last updated: April 11, 2024 2PM (will be refined/added to so please make sure you have latest version)

In this assignment, you are asked to develop (from scratch) a Python notebook that implements the Laplace Mechanism (LM) of differential privacy, applies it to several standard functions of numerical data, and visualizes the tradeoffs between the sensitivity of the function, the privacy parameter epsilon, and the size of the dataset.

More specifically, you will be asked to implement the following functionality:

1. Read in a numerical dataset
2. Plot histograms of the columns with mean, standard deviation, median, max and min values annotated
3. Normalize the dataset so that each column has a range of values in [0,1]; this will allow us to use the sensitivity bounds (which depend only on the number of datapoints used, not the data itself) that we derived in lectures
4. Implement the LM as a subroutine
5. For each of mean, standard deviation, median, max and min values, call your LM subroutine to sample many values from the LM at different values of epsilon and the number of data points n
6. Use these samples to plot/visualize the tradeoffs between epsilon, accuracy of LM, function sensitivity and n
7. In comments/text at the bottom of your notebook, write a brief essay discussing what your results/analyses show, how they vary with the function being computed, the column, the value of n, etc. Discuss whether/how the results align with the theory discussed in lectures, and be sure to discuss which functions are in the "sweet spot" for DP and why, and how your results reflect this.

More details on these steps (to be gradually refined/added):

a) Read in a numerical dataset: we'll use this dataset, which is a standard Google Colab dataset: https://www.cis.upenn.edu/~mkearns/teaching/EADSpring24/california_housing_train.csv
b) The dataset has 9 real-valued columns of 17,000 entries. Plot a 3 by 3 grid of subplots in which you plot histograms of each of the columns, with the mean, median, max and min values clearly marked by vertical lines through the histograms, and with standard deviation indicated by error bars around the mean. Make the resolution/number of bins in your histogram at least 100 so it is not too coarse.
c) Normalize each column of the dataset separately so that the range of each column is [0,1], i.e. map the smallest value in the column to 0, and the largest to 1, and the intermediate values to the corresponding place in [0,1]. Thus x in [0,1] should be mapped to |x-a|/|b-a|.
d) Implement the Laplace Mechanism as a subroutine, which should take the following input

parameters:

- The true/noise-free value of a function (e.g. mean, standard deviation, etc.) which you will need to compute and pass to the subroutine
- The desired value of epsilon, which you will set and vary outside the subroutine
- The sensitivity of the function in question, which you will again need to compute and pass to the subroutine; it should depend only on the number of datapoints used, not the data itself
- The number of samples from the LM to return

Your LM subroutine will use the first three parameters above to set the value of b, the parameter which determines the variance of the Laplace distribution as discussed in lecture and notes. It will then repeatedly sample noise from this Laplace distribution, adding the values to the true value of the function, and return a vector of samples whose length is equal to the number of samples requested.

e) For each of the 9 columns, and each of the 5 functions (mean, standard deviation, etc.), there will be a separate plot (so 9 x 5 = 45 plots total). For each column/function pair, the corresponding plot will have the privacy parameter epsilon on the x axis, and the accuracy on the y axis. The values of epsilon will be varied by your code systematically, and used to make repeated calls to the LM subroutine; you should use values of epsilon varying from 0.1 to 5.0 in reasonably small increments, e.g. 0.1. You will also vary the number of datapoints n of the column you use, from about n = 100 points to all n = 17,000, in increments of about 2500. Choose a collection of values of n that show an "interesting" range of behaviors, meaning they might not be equally spaced.

f) When you are using less than the full column, use a prefix of the column so that all of you are running on the same subset of the column. (Note that once you take a prefix of a column, the range of values might be a subinterval of [0,1], but the sensitivity of a function over the full range [0,1] is always an upper bound on the sensitivity over a subinterval.)
The plot for a given column/function pair will thus have multiple curves on it, each one corresponding to different choices of n. For each such curve, the x-axis will indicate the value of epsilon, and on the y-axis you should plot the average (over multiple LM samples) of the absolute difference between the true value of the function and the noisy LM samples. Use some healthy amount of LM samples (last parameter of subroutine) like 1000 to get an accurate estimate of these absolute differences. You plots should err on the side of specifying/labeling EVERYTHING: which column, which function, values of n on each curve, etc.

g) In comments/text at the bottom of your notebook, write a brief essay discussing what your results/analyses show, how they vary with the function being computed, the column, the value of n, etc. Discuss whether/how the results align with the theory discussed in lectures, and be sure to discuss which functions are in the "sweet spot" for DP and why, and how your results reflect this.