Lab 1: Differential Privacy

Distributed Monday February 13.
Report due Monday March 12 at 11:59PM via websubmit.

Remember to include a list of your collaborators and references, per the collaboration policy in the course syllabus!

In this lab, you will produce a differentially private analysis of a real-world dataset of your choice **given a total privacy budget of $\epsilon = 5$**, and compare the accuracy of your result to a "noise-free analysis", *i.e.,* an in-the-clear analysis done without privacy.

Datasets. The URLs below contains some datasets you might consider using, but there are many other datasets floating around the Internet. Feel free to choose anything that interests you, subject to the restriction that your dataset must contain at least 5000 records.
`http://archive.ics.uci.edu/ml/`
`http://www.kdnuggets.com/datasets/`

Code. You may implement the differentially private analysis any way you like. One way to do this is to download PINQ and start using it with Visual Studio (Windows) or Mono (Linux). This has the advantage that PINQ will impose the privacy restrictions for you. Another option is to implement that code in some other language you prefer; however, note that by doing this you must be very careful not to fall into any privacy traps like those described in Ex 4 and 5 in homework 1. You can download PINQ from this url:
`http://research.microsoft.com/en-us/downloads/73099525-fd8d-4966-9b93-574e6023147f/`

Please make sure you sort out what environment you plan to write code in by Tuesday February 22, and see me if you are having trouble before this date.

Report. The deliverable for this project is a written report. The report must contain the following parts:

1. **Part 0: Meta information.** the name of the dataset, url where you found it, the number of records and attributes in the dataset, and the language (PINQ, C, MATLAB) that you used for your implementation.
2. **Part 1: The differentially private report.** This is the meat of the lab. You must produce an analysis containing *at least* 10 statistics and two figures (plots) computed DP'ly from your dataset. Your figures may be CDFs, pie charts, or any other figure you want (*e.g.,* even something like Figure 4 of McSherry'09). This portion of the report may NOT contain any figures or statistics that are not computed differentially-privately. Each figure and statistic must be accompanied with a number that will index into Part 2 of the report. So for example, if your report includes the sentence

> ...we find that 50.5% (6) of Enron executives sent emails containing the word "bad accounting"...

Then index (6) will point to the portion of Part 2 that will explain the differentially-private algorithm you used to obtain the statistic 50.5%.

Remember that your *total* privacy budget is $\epsilon = 5$! (For example, you might use ten 0.1-DP operations, and four 1-DP operations without exceeding your privacy budget.)

3. **Part 2: Privacy and utility analysis.** For every index of a statistic or figure you include (*e.g.,* index (6) for statistic 50.5%) you should include:

   (a) Pseudocode for the DP mechanism you used to obtain this statistic or figure.

   (b) Privacy budget used up by this mechanism.

   (c) For single statistics (*e.g.,* our example of 50.5%) , include the signal-to-noise ratio (SNR) of the statistic, computed as $\frac{v_{nf}}{\sigma}$ where σ is the standard deviation of the output of the differentially private mechanism, and v_{nf} is the noise-free value the of statistic (*e.g.,* suppose the actual fraction of Enron employees that mentioned the words "bad accounting" is 48%, and our DP mechanism introduce noise with standard deviation 2%. Then the SNR is $\frac{48\%}{2\%} = 24$.) Note also that σ is computed via a mathematical analysis of the DP mechanism, while v_{nf} is computed by empirically measuring the dataset.

   (d) For figures with multiple statistics, include the maximum, minimum, and average SNR. (*e.g.,* For a CDF that outputs n different points, compute the SNR for each individual point, and then take the max, min, and average.)

   (e) The root-mean-square error for the measurement in your report, computed as:

   $$\sqrt{\frac{1}{n}\sum_i \left(1 - \frac{v_p[i]}{v_{nf}[i]}\right)^2}$$

   where $v_{nf}[i]$ is the noise-free output at index i, and $v_p[i]$ is the value of the statistic at index i that was computed DP'ly and included in Part 1. (*e.g.,* for our running example, we only have a single index $i = 1$, so $v_p[1] = 50.5\%$ and $v_{nf}[1] = 48\%$, so RMSE $= \sqrt{\left(1 - \frac{50.5}{48}\right)^2}$.)

   (f) The 'total noise added' by this mechanism, *i.e.,* $\sigma[i]$ the standard deviation of the output of the mechanism at index i. Total noise is then $\sum_i \sigma[i]$.

4. **Part 3: Summary.** Provide a summary that includes the following:

   (a) Total privacy budget used in your analysis.

   (b) Total root-mean-square error. Note that you can add two RSME errors by taking

   $$\sqrt{(RSME_1)^2 + (RSME_2)^2}$$

   (c) The 'total noise added', the sum of the standard deviations of all the DP measurements used in your analysis.

A note on repeating measurements. Note that repeating a differentially private measurement multiple times (and taking the average or the median) will *change* your privacy budget! (Refer to HW 1 Ex 5.) So if you do this, please make sure you adjust your privacy budget accordingly. Also, is repeating measurements multiple times and keeping the "good one" differentially private? (Refer to HW 1 Ex 5 bullet 4). If it is not, don't do this, because you are cheating! And if it is, adjust your privacy budget accordingly!

Grading.

1. Correctness: 60%. Awarded for correctly adhering to the requirements, *i.e.,* not exceeding privacy budget, providing the right number of figures and statistics, and having correct DP algorithms, etc.

2. Telling a story: 15%. Awarding for choosing statistics and figures and writing your report in a manner that tells an interesting story about your dataset. Very few points will be awarded for reports that contain lists of meaningless statistics and boring figures.

3. Utility: 15%. Awarded for designing DP algorithms that limit the total noise and RSME.

4. Pushing the envelope: 10% + 20% bonus. Awarded if you stay below the privacy budget of $\epsilon = 5$, use a DP algorithm we did not discuss in class, or provided more than the just minimum number of statistics and figures, etc.