Differential Privacy

Protecting privacy while performing statistical analysis is quite challenging. On the one had, the goal of statistics and machine learning is to be as informative as possible. Protecting privacy is the opposite goal.

How do we formally define privacy? Can we protect privacy and still do an informative analysis? We address these questions in these notes.

1 Introduction

The definition of privacy that has become most common lately is *differential privacy* (Dwork et al 2006).

2 Randomized Response

The predecessor to differential privacy is *randomized response* which is a method used in surveys. It was proposed by Warner in 1965.

I want to know how many students have ever cheated on a test. Suppose that proportion is p. If I ask this question I will not get truthful responses. I tell everyone to flip a coin C with $P(C = 1) = \theta$ and $P(C = 0) = 1 - \theta$. To protect their privacy, I tell them: if the coin is tails answer YES and if the coin is heads answer the question "have you every cheated?" The observation Y is thus $Y = (1 - C) + CZ$ where $Z = 1$ if they have cheated and $Z = 0$ otherwise. So $\pi \equiv P(Y = 1)$ is $\pi = (1 - \theta) + \theta p$ so that $p = (\pi - 1 + \theta)/\theta$. I can then estimate p by estimating π.

3 Differential Privacy

Suppose we have a dataset $X_1, \ldots, X_n$ where $X_i \in \mathcal{X}$. *Knowing the sample space $\mathcal{X}$ explicitly is critical for differential privacy.* The data set $D = \{X_1, \ldots, X_n\}$ is in $\mathcal{X}^n$. Our goal is to report some function $Z = T(D)$ of the data. We will be using some sort of randomization to do this. That is, we will take $Z \sim Q(\cdot|X_1, \ldots, X_n)$.

Two datasets D and D's are neighbors if they differ in one random variable. In other words $D = \{X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n\}$ and $D' = \{X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\}$. In this case we write $D \sim D'$.

We say that Q satisfies ϵ-differential privacy if

$$Q(Z \in A|D) \leq e^{\epsilon} Q(Z \in A|D')$$

for all A and all pairs $D \sim D'$. If Q has density q this means that

$$\sup_z \frac{q(z|D)}{q(z|D')} \leq e^{\epsilon}.$$

What does the definition mean? It means that whether you are in or not in the database has little affect on the output Z. For example, suppose I think you are person i in the database and I want to guess if you value is $X_i = a$ or $X_i = b$. Before I see any information, suppose my odds are $P(X_i = a)/P(X_i = b)$. After I see Z,

$$\frac{P(X_i = a|Z)}{P(X_i = b|Z)} = \frac{p(z|X_i = a)}{p(z|X_i = b)} \frac{P(X_i = a)}{P(X_i = b)}$$

and so

$$e^{-\epsilon} \frac{P(X_i = a)}{P(X_i = b)} \leq \frac{P(X_i = a|Z)}{P(X_i = b|Z)} \leq e^{\epsilon} \frac{P(X_i = a)}{P(X_i = b)}.$$

Since $e^{\epsilon} \approx 1 + \epsilon$ and ϵ is small, we see that knowing Z does not change my odds much. It is also possible to show that we cannot construct any test with non-trivial power about what your value of X_i is.

So, when differential privacy holds, we cannot learn much about whether a particular person is in the dataset.

4 Queries

The computer science model of privacy is that some curator keeps the data and users send queries about the data. The goal is to release answers that are differentially private.

Let f be a function of the data and define *the sensitivity*

$$\Delta = \sup_{D \sim D'} |f(D) - f(D')|.$$

Suppose we release

$$Z = f(D) + W$$

where $f(w) \propto e^{-w/\lambda}$. Note that W has a Laplace distribution with standard deviation $\sqrt{2}\lambda$. If we set $\lambda = \Delta/\epsilon$ then

$$\frac{p(z|D)}{p(z|D')} \leq e^{|f(D) - f(D')|/\lambda} \leq e^{\epsilon}$$

so that differential privacy holds.

For example, suppose that $X_1, \ldots, X_n \in [-B, B]$ and that $f(D) = \overline{X}$. Then $\Delta = 2B/n$ so we need to add noise with standard deviation $O(B/(n\epsilon))$. As a function of n this is good. As a function of B it is bad.

5 How Informative is Z?

Obviously we lose information when we use differential privacy.

As an extreme example, suppose the data are on $[0, 1]$ and suppose the true distribution F is a point mass at $x \in [0, 1]$. So the dataset is $X = (X_1, \ldots, X_n) = (x, x, \ldots, x)$. Suppose we output $Z_1, \ldots, Z_k$ from a differentially private $Q(Z|X)$. Let $\widehat{F}$ be the empirical distribution of Z. Then it can be shown that $\widehat{F}$ must be inconsistent, that is, there exists $\delta > 0$ such that,

$$\liminf_{n\to\infty} P(\sup_s |F(s) - \widehat{F}(s)| > \delta) > 0.$$

See Blum, Ligett and Roth (2008) and Wasserman and Zhou (2010).

Barber and Duchi (2014) studied differential privacy from the minimax point of view. Suppose we observe $X_1, \ldots, X_n \sim P$ where $X_i \in [0, 1]^d$. Consider the simple task of estimating the mean μ. If we ignore privacy, then we can use $\overline{X}$ which is risk $\mathbb{E}[||\overline{X} - \mu||^2] \preceq d/n$. They showed that any differentially private estimator $\widetilde{\mu}$ satisfies the lower bound

$$\mathbb{E}[||\widetilde{\mu} - \mu||^2] \succeq \frac{d}{n} + \frac{d^3}{n^2\epsilon^2} = \frac{d}{n}\left[1 + \frac{d^2}{n\epsilon^2}\right].$$

So the price we pay for privacy is $\frac{d^3}{n^2\epsilon^2}$ which is quite steep.

6 Releasing a Whole Dataset

Statisticians have been less enthusiastic about differential privacy than computer scientists. One of the reasons for this is the heavy dependence on the notion of using privatized queries. The idea that we would analyze data by sending queries to a curator is unrealistic. Real data analysis involves: looking at the data, fitting models, testing fit, making predictions, constructing confidence sets etc. This requires access to the whole data set. This leads to the following questions. Can we release a privatized version of the whole dataset? In fact, there are several ways to do this.

6.1 Exponential Mechanism

The exponential mechanism, due to McSherry and Talwar (2007), is a general method for preserving differential privacy. Here, I'll discuss the special case where we want to release a private data set $Z = (Z_1, \ldots, Z_k)$. Let $\xi(x, y)$ be some function that measures the distance between two data sets $x = (x_1, \ldots, x_n)$ and $z = (z_1, \ldots, z_k)$. Define the sensitivity

$$\Delta = \sup_{x \sim y} \sup_{z} |\xi(x, z) - \xi(y, z)|.$$

Now draw $Z = (Z_1, \ldots, Z_k)$ from the density

$$q(z|x) \propto \exp\left(-\frac{\epsilon \xi(x, z)}{2\Delta}\right).$$

It is easy to check that this satisfies ϵ-differential privacy.

As an example, suppose that $\mathcal{X}$ is compact and define $\xi(x, z) = \sup_t |F_x(t) - F_z(t)| = ||F_x - F_z||_\infty$ where F_x is the empirical cdf of $x = (x_1, \ldots, x_n)$ and F_z is the empirical cdf of $z = (z_1, \ldots, z_k)$. So ξ is the Kolmogorov-Smirnov distance. In this case, $\Delta = 1/n$ and so we draw z from the density

$$q(z|x) \propto \exp\left(-\frac{n\epsilon ||F_x - F_z||_\infty}{2}\right).$$

Wasserman and Zhou (2010) showed that, for this scheme, the optimal choice of k is $k \asymp n^{2/3}$ and that $||F - F_z||_\infty = O_P(n^{-1/3})$. Without privacy we have $||F - F_x||_\infty = O_P(n^{-1/2})$. So we see that we have lost accuracy.

More generally, they showed that

$$P\left(||F - F_z||_\infty > \delta\right) \leq \frac{(\sup_x p(x))^k e^{-3\epsilon\delta n/16}}{S(k, \delta/2)}$$

where $S(k, \delta/2)$ is the *small ball probability*, that is $P(||F - F_z|| \leq \delta/2)$. However, it is not known if these bound are tight.

6.2 Density Estimation I

Another way to release a privatized dataset is to compute a privatized density estimate $\widehat{p}$. Then we can draw a sample $Z_1, \ldots, Z_N \sim \widehat{p}$. It is easy to show that if $\widehat{p}$ is differentially private then so is $Z = (Z_1, \ldots, Z_N)$.

Dwork et al (2006) suggested using a privatized histogram which was analyzed in Wasserman and Zhou (2010). Suppose that the data are on $[0, 1]^d$. Divide the space into $m = 1/h^d$ bins

$B_1, \ldots, B_m$ and form the usual histogram estimator

$$\widehat{p}(x) = \sum_j \frac{\widehat{p}_j}{h^d} I(x \in B_j)$$

where $\widehat{p}_j = C_j/n$ and C_j is the number of observations in bin B_j. To privatize $\widehat{p}$, define

$$\widehat{q}(x) = \sum_j \frac{\widehat{q}_j}{h^d} I(x \in B_j)$$

where $\widehat{q}_j = \widetilde{D}_j / \sum_s \widetilde{D}_s$, $\widetilde{D}_j = \max\{D_j, 0\}$ and $D_j = C_j + \nu_j$ where ν_j is drawn from a Laplace density with mean 0 and variance $8/\epsilon^2$. Wasserman and Zhou (2010) showed that, if X has a Lipschitz density, and if we choose $m \asymp n^{d/(2+d)}$ the histogram of the privatized density achieves the minimax rate $n^{-2/(2+d)}$. So in this case, there is no loss in rate by releasing the privatized data or the privatized histogram.

However, the minimax rate is not the whole story. Suppose that the original histogram $\widehat{p}$ is sparse i.e. has many empty cells. The privatized histogram $\widehat{q}$ is forced to "fill in" these empty cells. So in these cases, $\widehat{q}$ will look very different from $\widehat{p}$. In particular, much of the clustering structure will be lost. And if there is any lower dimensional structure in the data, this will be destroyed.

6.3 Density Estimation II

A second approach is based on orthogonal series. For simplicity assume that $\mathcal{X} = [0, 1]$. Write

$$p(x) = 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x)$$

where $\{1, \psi_1, \psi_2, \ldots, \}$ is an orthonormal basis. Suppose that $\sum_j \beta_j^2 j^{2\gamma} \leq C^2$. This is a Sobolev ellipsoid. The minimax rate is $n^{-2\gamma/(2\gamma+1)}$.

The usual density estimator in this framework is

$$\widehat{p}(x) = 1 + \sum_{j=1}^{m} \widehat{\beta}_j \psi_j(x)$$

where $m = n^{1/(2\gamma+1)}$ and $\widehat{\beta}_j = n^{-1} \sum_i \psi_j(X_i)$ which achieves the minimax rate. A privatized estimator is

$$\widehat{q}(x) = 1 + \sum_{j=1}^{m} (\widehat{\beta}_j + \nu_j) \psi_j(x)$$

where ν_j is Laplace with mean 0 and standard deviation $mc_0/(n\epsilon)$ where $c_0 = \sup_j \sup_x |\psi_j(x)|$. It turns out that $\widehat{q}$ also achieves the minimax rate.

6.4 Density Estimation III

The most commonly used density estimator is the kernel density estimator

$$\widehat{p}(x) = \frac{1}{n}\sum_i \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right).$$

Is there a way to privatize $\widehat{p}$?

This is trickier than histograms and orthogonal series estimators since $\widehat{p}$ is not easily described by a finite set of parameters. In principle, we want to draw a random function g such that $P(g \in A|D) \leq e^\epsilon P(g \in A|D')$. But the sets A are now subsets of some function space and it is not immediately clear how to do this.

So far, I know of only two ways to do this. Hall, Rinaldo and Wasserman (2013) suggested using

$$g = \widehat{p} + \frac{C}{n\epsilon h^{d/2}} G$$

where C is an appropriate constant and G is a mean 0 Gaussian process with a certain covariance structure. The resulting density estimator g is very wiggly but it does satisfy differential privacy. Moreover, it has the same rate of convergence as the original density estimator.

A second approach was recently presented by Alda and Rubinstein (2017). The first create a grid on the sample space. Next, the approximate $\widehat{p}$ using Bernstein polynomials. Then the add Laplace noise to the coefficients of the polynomials.

I should add that the methods in Hall, Rinaldo and Wasserman (2013) and Alda and Rubinstein (2017) are quite general and can be used for the private release of fairly general functions. In fact, Alda and Rubinstein (2017) also apply their approach to classification, logistic regression and empirical risk minimization. They also provide a lower bound which shows that we must, in general, introduce an error of size Δ/ϵ when privately releasing a function, where Δ is the sensitivity defined earlier.

7 Conclusion

Differential privacy (DP) is a very active area of research. Here is a summary of the strengths and weaknesses of this approach:

Strengths:

1. DP gives a very rigorous, precise notion of privacy.

2. Many methods in machine learning and statistics can be made differentially private.
3. DP can be used for other purposes. For example, Dwork et al (2015) created a method called *reusable holdout* that allows and interactive approach to data analysis while making repeated looks at the data without introducing too much bias. The hear of the method is to impose a sort of differential privacy on each step of the analysis.

Weaknesses:

1. DP has dominated the research in privacy. It seems that there is not much research in other approaches.
2. DP is very strong. You need to add a lot of noise to the data.
3. When there is a structure in the data, such as voids, manifolds etc, it is destroyed by DP.
4. I have not seen it really used in much practical data analysis.