A Statistical View of Differential Privacy

January 2011

Larry Wasserman
Dept of Statistics
and
Machine Learning Department
Carnegie Mellon University

Joint work with Shuheng Zhou
Ongoing work with Shuheng Zhou and Alessandro Rinaldo

Overview

I CS versus Statistics: a short story.

II Statistical Thinking.

III Sanitized databases.

IV Problems with differential privacy.

I: A Short Story

- I live in several worlds:

-statistical theory
-applications (astronomy, biology, genomics)
-machine learning (theory and methdology)

- Steve Fienberg got me and a post-doc (Shuheng Zhou) intersted in privacy. I dabbled a bit and we wrote a paper.

- I was invited to IPAM: there was statisticians and computer scientists.

- The statisticians were mostly applied statisticians working on real problems.

- The CS people were mainly theoreticians doing very interesting theory.

A Short Story

- There was a huge cultural divide.

- The CS people wanted precise definitions of privacy and theorems guranteeing that privacy (mostly differential privacy) held. I liked that.

- The statisticians wanted methods that worked on real, complex, data sets. I liked that.

- With a few notable exceptions (Steve, Adam, Cynthia, ...) they ignored each other.

A Short Story

- CS view: receive a query, return a private answer.

- Statistics view: give me data. Then I can: draw plots, fit models, test fit, estimate parameters, make predictions, ...

- The moral of the story: statisticians want a sanitized data base, not answers to specific queries.

- I have a dream: statisticians will read the CS literature and CS people will read the statistics literature. There is a huge opportunity for collaboration.

- You will notice many unfinished items marked: in progress.

Overview

I CS versus Statistics: a short story.

II Statistical Thinking.

III Sanitized databases.

IV Problems with differential privacy.

Statistical Thinking

or Some Statistical Concepts

- What do statisticians do?

-Exploratory methods
-Model fitting
-Looking at residuals
-Assessing fit
-Develop new methods
-Theory

- We view these as very inter-connected.

Some Statistical Concepts

- Data $D = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n \sim P$.

- Often (but not always) we view the data base as a sample from a population. The goal is not just to summarize the database; we want to infer (learn) about the population.

- Formally, the goal is to infer P or some functions of P (means, correlations, etc.) or predict a new observation.

- A model is a set of distributions $\mathcal{P}$.

- Can be parametric: Example: $\mathcal{P} =$ Normal distributions.

- Can be nonparametric: Example: $\mathcal{P} =$ all distributions.

Point estimation

- Let $\theta = T(P)$. (Example, $T(P)$ is the mean of P.)
- Estimator: $\widehat{\theta} = g(X_1, \ldots, X_n)$.
- Loss function $\ell(\widehat{\theta}, \theta)$. Example: $\ell(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$.
- Optimality: find $\widehat{\theta}$ that achieves the minimax risk:

$$\boxed{R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\widehat{\theta}, \theta)]}$$

- How is R_n affected by privacy?

- Example: If θ is the mean and $\mathcal{P}$ is all Normals, then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is minimax for all bowl-shaped loss functions.

- $R_n = O(1/\sqrt{n})$. Extra loss from differential privacy is $O(1/n)$.

$$\frac{|R_n - R_n^*|}{R_n} = 1 + O\left(\frac{1}{\sqrt{n}}\right).$$

$\star$ This is not quite true. The real statement requires working on a bounded domain and involved a few complications.

- Example: Estimating a probability density function.
- Observe from $X_1, \ldots, X_n \sim p$, where $X_i \in \mathbb{R}^d$. Estimate p.
- Loss $\int (p(x) - \widehat{p}(x))^2 dx$ where $\widehat{p}$ is the estimator.
- Minimax risk:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p \int (p(x) - \widehat{p}(x))^2 dx = \frac{C}{n^{4/(4+d)}}$$

where $\mathcal{P}$ is all smooth densities.

- We have several methods for estimating p that are optimal. This will be useful when we discuss sanitized databases. More later.

Confidence Intervals

- Possibly the most important and most common statistical calculation.

- Find a (random) set $C = C(X_1, \ldots, X_n)$ such that

$$P(\theta \in C) \geq 1 - \alpha \quad \text{for all } P.$$

- Again, can be parameteric or nonparametric.

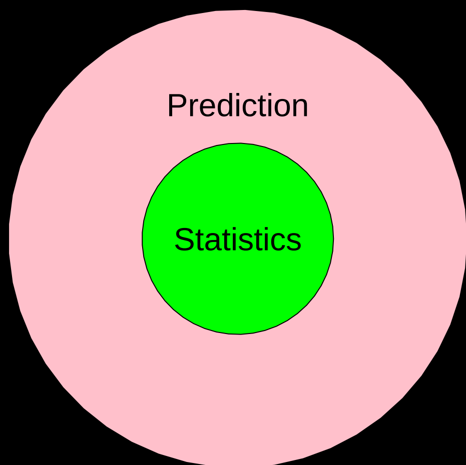
- Optimality: want C to be as small as possible.

- Example: correlation between a SNP and a disease.

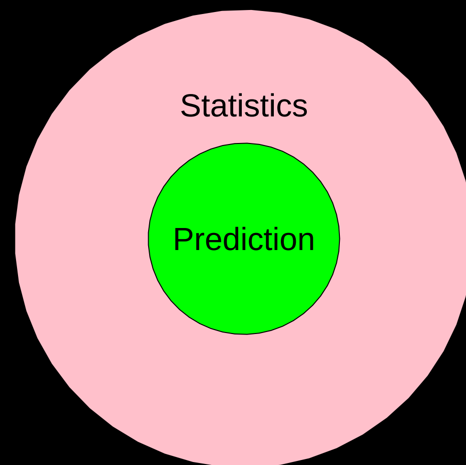
Prediction

- Data $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Observe new X_{n+1}. Predict Y_{n+1}.
- If $Y \in \mathbb{R}$ this is regression. If Y is discrete this is classification.
- Classification is a small (but important) part of statistics.

CS View



Statistics View



Other Things Statisticians Do

- Clustering.
- PCA.
- Curve Estimation.
- Experimental design.
- Graphical models.
- Networks.
- Manifold methods.
- Hypothesis testing.
- Yada yada yada.

Overview

I CS versus Statistics: a short story.

II Statistical Thinking.

III Sanitized databases.

IV Problems with differential privacy.

How To Make a Sanitized Database

(And How To Measure Accuracy)

- Data $X = (X_1, \ldots, X_n)$. $X \in \mathcal{X}$ = set of possible databases.

$$\boxed{\bullet\ P \longrightarrow X = (X_1, \ldots, X_n) \longrightarrow Q \longrightarrow Z = (Z_1, \ldots, Z_k)}$$

where $Q(Z_1, \ldots, Z_k | X_1, \ldots, X_n)$ generates random (synthetic) data.

- Require:

$$\boxed{Q(Z \in B | X) \leq e^{\alpha}\, Q(Z \in B | X') \quad \text{for all } B}$$

whenever $X \sim X'$ (differ by one observation).

Method I: Density Estimation.

(1) Choose a basis: $\psi_1, \psi_2, \ldots,$ (Fourier, polynomial, wavelets, ...)

Expand: $p(x) = \sum_j \beta_j \psi_j(x)$ where $\beta_j = \int \psi_j(x) p(x) dx$.

(2) Estimate p with

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

where

$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \psi_j(X_i).$$

Method I: Density Estimation.

(3) Privatize:

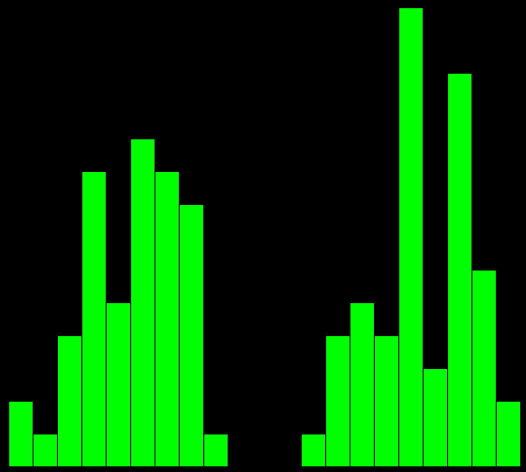
$$\widehat{\beta}_j^* = \widehat{\beta}_j + \frac{cJ}{n} L_j$$

where $L_1, \ldots, L_J$ are Laplace random variables. Let

$$\widehat{p}^*(x) = \sum_{j=1}^{J} \widehat{\beta}_j^* \, \psi_j(x).$$

(4) Sample from p^*: $Z_1, \ldots, Z_k \sim \widehat{p}^*$.

- Histograms:

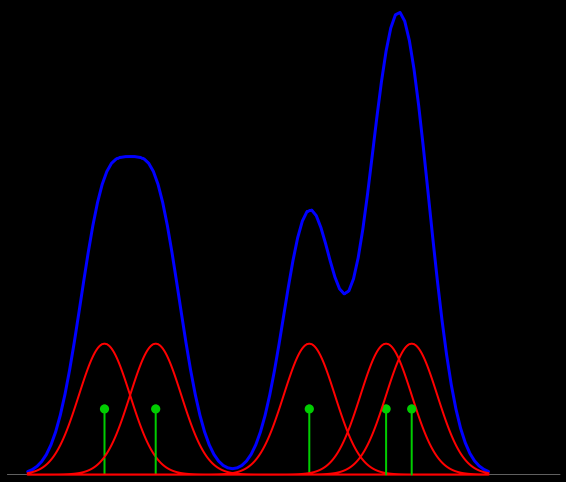


- Add Laplace noise to the counts, sample from the histogram. (This is less accurate.)

- Kernel density estimator:

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{k} \frac{1}{h^d} K\left(\frac{||x - X_i||}{h} \right)$$

where $h > 0$ is a bandwidth and $K(\cdot)$ is a kernel.



- Privatize (sample and aggregate) then sample from $\widehat{p}^*$. (in progress).

- Technical point:

- Each density estimate requires careful choice of tuning parameter:

-series estimator: number of terms
-histogram: size of the bins
-kernel estimator: bandwidth

- We do this using risk estimation.

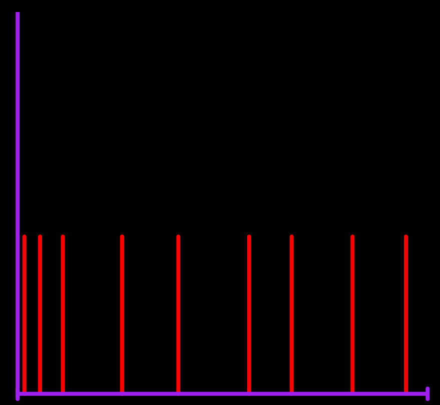
- This needs to be privatized too.

- In progress.

Method II: Exponential Mechanism (McSherry and Talwar 2007)

- First need to explain: (i) empirical distributions and (ii) metrics on distributions.

- Let P_X be the distribution that puts mass $1/n$ at each X_i.



Some Metrics d

- Supremum distance: $d(P,Q) = \sup_{A\in\mathcal{A}} |P(A) - P(A)|$.

- Wasserstein: (Earth-mover):

$$d(P,Q) = \inf_J \mathbb{E}_J ||X - Z||^p$$

where $X \sim P$, $Y \sim Q$ and the infimum is over all joint distributions J with marginals P and Q.

- L_2 distance on counts: $||x - z|| = \sqrt{\sum_j (x_j - z_j)^2}$.

- L_1 distance on densities: $\int |p - q|$.

Exponential mechanism

- Draw $Z = (Z_1, \ldots, Z_k)$ from

$$q(z_1, \ldots, z_k | x_1, \ldots, x_n) \propto \exp\left(-\frac{\alpha\, d(P_x, P_z)}{2\Delta}\right)$$

where

$$\Delta = \Delta(n, k) = \sup_{x,x'} \sup_{z} |d(P_x, P_z) - d(P_{x'}, P_z)|.$$

- Can do the sampling by importance sampling but it is difficult.

ACCURACY

Let $\mathcal{Q}_\alpha$ be all Q's that satisfy α-differential privacy.

$$
\begin{aligned}
R(\mathcal{X}) &= \inf_{Q\in\mathcal{Q}_\alpha} \sup_{x\in\mathcal{X}} \mathbb{E}_Q(d(P_x, P_Z)) \\
R(\mathcal{P}) &= \inf_{Q\in\mathcal{Q}_\alpha} \sup_{P\in\mathcal{P}} \mathbb{E}_P\mathbb{E}_Q(d(P_X, P_Z)).
\end{aligned}
$$

This has been in a few cases. (See Hardt and Talwar 2007, Roth 2010, and Kasiviswanathan, Rudelson, Smith and Ullman 2010 for example.) Mostly, it is work in progress.

Less ambitious is to compute:

$$
R(\mathcal{P}, Q) = \sup_{P\in\mathcal{P}} \mathbb{E}_P\mathbb{E}_Q(d(P_X, P_Z))
$$

for some specific Q's.

An Accuracy Bound

- Draw $Z = (Z_1, \ldots, Z_k)$ from

$$q(z|x) \propto e^{-\alpha\, d(P_x, P_z)/(2\Delta)}$$

where

$$\Delta = \Delta(n,k) = \sup_{x,x'} \sup_{z} |d(P_x, P_z) - d(P_{x'}, P_z)|.$$

- Theorem (WZ 2010):

$$\mathbb{P}(d(P, P_Z) > \epsilon) \leq \frac{(\sup_x p(x))^k e^{-3\alpha\epsilon/(16\Delta)}}{R(k, \epsilon/2)}$$

where

$$R(k,\epsilon) = \mathbb{P}(d(P, P_S) \leq \epsilon) \qquad S = (S_1, \ldots, S_k) \sim P$$

is the small ball probability.

dimension r

Distance	Data Release mechanism			minimax rate
	smoothed histogram	perturbed histogram	exponential mechanism	
L_2	$n^{-2/(2r+3)}$	$n^{-2/(2+r)}$	NA	$n^{-2/(2+r)}$
KS	$\sqrt{\log n} \times n^{-2/(6+r)}$	$\log n \times n^{-2/(2+r)}$	$n^{-1/3}$	$n^{-1/2}$

dimension $r = 1$

	exponential mechanism	perturbed orthogonal series estimator	minimax rate
L_2	$n^{-\gamma/(2\gamma+1)}$	$n^{-2\gamma/(2\gamma+1)}$	$n^{-2\gamma/(2\gamma+1)}$

Main point: general picture on optimality not yet clear.

Overview

I CS versus Statistics: a short story.

II Statistical Thinking.

III Sanitized databases.

IV Problems with differential privacy.

Problems With Differential Privacy

- Differential privacy is a precise and strong guarantee.

- But there are two problems:

- First, recall that $\mathcal{X}$ = set of possible databases. What is $\mathcal{X}$? It is ambiguous.

- And as Avrim pointed out, the notion of neighboring databases can be ambiguous.

- Also, $\mathcal{X}$ can be exotic. For example, in functional data analysis, each data point is a function living in some function space.

Problems With Differential Privacy

- Consider histogram counts: $X = (c_1, \ldots, c_k)$ where $\sum_j c_j = n$.

- Lower bounds:

Hardt and Talwar (2009): $O(k^{3/2})$. We got: $O(k^{3/2}/n)$.

- Depends on whether $\mathcal{X}$ is all histograms or $\mathcal{X}$ histograms with sample size n.

- In many real problems, it simply is not clear what $\mathcal{X}$ is.

- Need to know $\mathcal{X}$ to even implement differential privacy.

Problems With Differential Privacy

- Second, it is too strong.

- Consider a high dimensional contingency table. The counts are very sparse. There are many zeroes.

- The sample size is n is much smaller than the number of cells.

- Creating a synthetic database subject to differential privacy leads to a very noisy database. (Mostly noise.)

CONCLUSION

- Differential privacy is a precise, mathematical guarantee.
- This precision is useful theoretically but makes it somewhat impractical.
- Popular in CS. Mostly ignored in statistics.
- Need modified version of differential privacy?