Introduction to Differential Privacy

Audra McMillan

November 19, 2015
Introduction to Differential Privacy

Audra McMillan

November 19, 2015
Motivation

Suppose Sebastian knows everyone in the departments belieber status and he has had a request to make the fraction of the department that likes the Beibs public. Michael is upset because he doesn’t want anyone to EVER be able to work out whether or not be likes Justin Beiber.

Should Michael be worried?
Yes. Imagine Michael’s nemesis, Erin, has stealthily accumulated everyone but Michael’s belieber status over the years. If Sebastian releases the average that includes Michael then Erin can easily work out Michael’s status from the information that she has.
But most people don’t have nemeses.

Suppose you are an insurance company and you have a large amount of medical data that you think would be of interest to the medical community. How can you release the data, while maintaining the privacy of the patients?
But most people don’t have nemeses.

Suppose you are an insurance company and you have a large amount of medical data that you think would be of interest to the medical community. How can you release the data, while maintaining the privacy of the patients?

ANONYMISE THE DATA!
The Massachusett’s Group Insurance Commision anonymised their data and released it to the public.
Case Study

The Massachusett’s Group Insurance Commision anonymised their data and released it to the public.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Ethnicity</th>
<th>Marital Status</th>
<th>Years Married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>Single</td>
<td>32</td>
</tr>
<tr>
<td>Male</td>
<td>50</td>
<td>White</td>
<td>Married</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>23</td>
<td>White</td>
<td>Single</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>75</td>
<td>White</td>
<td>Divorced</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>Black</td>
<td>Married</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>61</td>
<td>Asian</td>
<td>Widowed</td>
<td>20</td>
</tr>
<tr>
<td>Female</td>
<td>45</td>
<td>White</td>
<td>Married</td>
<td>28</td>
</tr>
<tr>
<td>Male</td>
<td>63</td>
<td>White</td>
<td>Married</td>
<td></td>
</tr>
</tbody>
</table>

Voting registration data + news reports that the Governor of Massachusett’s had had the flu on a certain date and been admitted to hospital.
Case Study

The Massachusetts’s Group Insurance Commission anonymized their data and released it to the public.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Ethnicity</th>
<th>Marital Status</th>
<th>Years Married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>Single</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>56</td>
<td>White</td>
<td>Married</td>
<td>32</td>
</tr>
<tr>
<td>Female</td>
<td>23</td>
<td>White</td>
<td>Single</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>75</td>
<td>White</td>
<td>Divorced</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>Black</td>
<td>Married</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>61</td>
<td>Asian</td>
<td>Widowed</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>45</td>
<td>White</td>
<td>Married</td>
<td>20</td>
</tr>
<tr>
<td>Male</td>
<td>63</td>
<td>White</td>
<td>Married</td>
<td>28</td>
</tr>
</tbody>
</table>

 Voting registration data + news reports that the Governor of Massachusetts’s had had the flu on a certain date and been admitted to hospital.

The Governor of Massachusetts’s medical records.
Privacy is often not very well thought out...

Hopefully you are convinced by now that there is a need for a proper definition of what it means for data analysis to be “private”.

If not, talk to me after.
The VC dimension.

Suppose I have a set $H$ of binary functions with domain $X$. Then $H$ is said to **shatter** a set of points $\{x_1, \ldots, x_m\}$ if for every possible assignment of labels in $\{0, 1\}^m$, there exists a function in $H$ that agrees with these labels. That is

$$\forall \ell \in \{0, 1\}^m \exists f \in H \forall i \in \{1, \ldots, m\} \ f(x_i) = \ell_i.$$ 

The set $H$ is said to have **VC-dimension** $d$ if there exists a set of size $d$ that $H$ shatters but not set of size $d + 1$ that is shattered by $H$. 
Example 1: Linear Threshold functions

Consider the class of linear classifiers on $\mathbb{R}^2$. 
Example 1: Linear Threshold functions

Consider the class of linear classifiers on $\mathbb{R}^2$.

VC dimension = 3
How should I think about the VC-dimension?

Example 2: Let $H$ be a single function. The VC dimension $H =$ ?
How should I think about the VC-dimension?

Example 2: Let $H$ be a single function. The VC dimension $H = 0$. 
Example 2: Let $H$ be a single function. The VC dimension $H = 0$.

The VC dimension is a measure of how “complex” the class of functions $H$ is. Intuitively, it captures what kind of complicated patterns the functions in $H$ can have. This is a pretty combinatorial notion but it is going to relate to how easy it is the “learn” a function from $H$ and “differentially privately compute” all the functions in $H$. 
A simple learning problem

Suppose I have a class of functions $H$ and you choose $h \in H$. You then choose a random sample $S = \{x_1, \cdots, x_m\}$ from the domain $X$.

You then tell me $H$ and $\{(x_1, h(x_1)), \cdots, (x_m, h(x_m))\}$ and ask me to guess $h$. 
What does the VC-dimension have to do with anything?

Hint: The VC-dimension is going to turn up in how many samples you need to give me before I have a good chance of guessing a function that is pretty close to the right answer.
What does the VC-dimension have to do with anything?

Let’s consider how many samples we might need to determine a pretty good linear classifier.

Really, we just need a couple of sample points from each side to be able to determine the classifier pretty well. With a pretty small number of samples, we have pretty high probability of choosing a sample like this.
What does the VC-dimension have to do with anything?

In comparison, suppose $H$ is the class of threshold functions defined by cubic functions. Intuitively, we need samples from more areas in order to make a good guess.
Let $\mathcal{A}$ be an algorithm that takes as input a function class $H$ with VC-dimension $d$ and a random sample $S$ of size $m$. Then there exists such an algorithm that has the property that if

$$
m \geq C \left( \frac{d + \ln(1/\delta)}{\varepsilon} \right)$$

for some constant $C$ then with probability $1 - \delta$ the random sample $S$ will be such that the algorithm outputs a function that is wrong on at most an $\varepsilon$ fraction of the domain.
Back to Sebastian and Michael!

We’ve agreed that Michael has reason to be worried but Sebastian REALLY wants to release the statistic. What kind of guarantee should Sebastian give Michael?
What kind of guarantee are we giving Michael?

Option 1: Sebastian could promise Michael the statistic that he releases will not allow anyone to learn anything about him.
What kind of guarantee are we giving Michael?

Option 1: Sebastian could promise Michael the statistic that he releases will not allow anyone to learn anything about him.

Problem: Even if Michael had never trusted Sebastian, this would be an unrealistic promise. Suppose Sebastian computes the fraction without including Michael’s data and it turns out that 99% of the department like Justin Beiber. Then there is a good chance that Michael also likes Justin Beiber so I’ve learnt something about him but have I invaded his privacy?
What kind of guarantee are we giving Michael?

Option 1: Sebastian could promise Michael the statistic that he releases will not allow anyone to learn anything about him.

Problem: Even if Michael had never trusted Sebastian, this would be an unrealistic promise. Suppose Sebastian computes the fraction without including Michael’s data and it turns out that 99% of the department like Justin Beiber. Then there is a good chance that Michael also likes Justin Beiber so I’ve learnt something about him but have I invaded his privacy?

Differential privacy says that this is not an invasion of Michael’s privacy so this is not the promise Sebastian will be giving Michael.
What kind of guarantee are we giving Michael?

Option 2: Sebastian could promise Michael that the inclusion of his data will not affect the statistic he outputs at all.
What kind of guarantee are we giving Michael?

Option 2: Sebastian could promise Michael that the inclusion of his data will not affect the statistic he outputs at all.

Problem: Presumably Sebastian should guarantee privacy for all his confidantes, in which case any statistic with this guarantee will be completely useless.
What kind of guarantee are we giving Michael?

Option 3: Sebastian could promise Michael that the answer will reveal (almost) exactly as much about him whether or not Sebastian includes Michael’s data in his calculation. That is, Sebastian has an algorithm that he is going to input his data into that will output something that looks like the fraction of the department that like Justin Beiber. Sebastian tells Michael that I can include your data or not, but I promise you that Erin will not be able to tell which one of these options I chose.
Option 3: Sebastian could promise Michael that the answer will reveal (almost) exactly as much about him whether or not Sebastian includes Michael’s data in his calculation. That is, Sebastian has an algorithm that he is going to input his data into that will output something that looks like the fraction of the department that like Justin Beiber. Sebastian tells Michael that I can include your data or not, but I promise you that Erin will not be able to tell which one of these options I chose.

THIS IS THE GUARENTEE THAT DIFFERENTIAL PRIVACY GIVES!
Differential Privacy

An Algorithm $\mathcal{A}$ is $\epsilon$-differentially private if for all databases $D, D'$ that differ on a single individual and for all subsets $K$ of the range

$$\Pr(\mathcal{A}(D) \in K) \leq e^\epsilon \Pr(\mathcal{A}(D') \in K).$$
An Algorithm $\mathcal{A}$ is $\epsilon$-differentially private if for all databases $D, D'$ that differ on a single individual and for all subsets $K$ of the range

$$\Pr(\mathcal{A}(D) \in K) \leq e^{\epsilon}\Pr(\mathcal{A}(D') \in K).$$

An Algorithm $\mathcal{A}$ is $(\epsilon, \delta)$-differentially private if for all databases $D, D'$ that differ on a single individual and for all subsets $K$ of the range

$$\Pr(\mathcal{A}(D) \in K) \leq e^{\epsilon}\Pr(\mathcal{A}(D') \in K) + \delta.$$
Why is this definition so great?

In terms of privacy...
Why is this definition so great?

In terms of privacy...

1. It makes sense as a definition for privacy. (It gives Michael plausible deniability)
Why is this definition so great?

In terms of privacy...

1. It makes sense as a definition for privacy. (It gives Michael plausible deniability)
2. It quantifies privacy loss. So Sebastian has something concrete he can say to Michael and Michael can make concrete demands of Sebastian.
Why is this definition so great?

In terms of privacy...

1. It makes sense as a definition for privacy. (It gives Michael plausible deniability)
2. It quantifies privacy loss. So Sebastian has something concrete he can say to Michael and Michael can make concrete demands of Sebastian.
3. It is additive in a nice way. If Sebastian performs an $\epsilon$-DP algorithm then an $\epsilon'$-DP algorithm then the composition of the algorithms is $(\epsilon + \epsilon')$-DP. EVEN if the choice of the second algorithm is dependant on the answer to the first algorithm.
Why is this definition so great?

In terms of privacy...

1. It makes sense as a definition for privacy. (It gives Michael plausible deniability)
2. It quantifies privacy loss. So Sebastian has something concrete he can say to Michael and Michael can make concrete demands of Sebastian.
3. It is additive in a nice way. If Sebastian performs an $\epsilon$-DP algorithm then an $\epsilon'$-DP algorithm then the composition of the algorithms is $(\epsilon + \epsilon')$-DP. EVEN if the choice of the second algorithm is dependant on the answer to the first algorithm.
4. There is no way that Erin can do post-processing to get more information about Michael. This method ensures that regardless of how much outside information Erin has, she can't determine anything about Michael.
Why is this definition so great?

In terms of utility...
Why is this definition so great?

In terms of utility...

1. It turns out that you can still do useful data analysis.
Why is this definition so great?

In terms of utility...

1. It turns out that you can still do useful data analysis.
2. This is actually a pretty reasonable thing to expect of an data analysis algorithm whose answer is supposed to generalise to an entire population.
So what should Sebastian do?
So what should Sebastian do?

ADD NOISE!
So what should Sebastian do?

ADD NOISE! But how much?

Suppose these two distributions are centred at the answers with and without Michael’s data. Then we can choose the variance so that they look pretty similar.
So what should Sebastian do?

ADD NOISE! But how much?

Suppose these two distributions are centred at the answers with and without Michael’s data. Then we can choose the variance so that they look pretty similar.

If Sebastian samples from one of these distributions and outputs that as his answer then Erin can’t tell which distribution Sebastian sampled from. Also, with high probability the output will be close to the mean (the answer he wanted to output).

This is called the Laplace Mechanism.
Is adding noise the end of the story?

NO!
Is adding noise the end of the story?

NO!

Definitely not. There are loads of examples of when adding noise is not a reasonable thing to do. In particular, if the output is not a metric space.
Is adding noise the end of the story?

NO!

Definitely not. There are loads of examples of when adding noise is not a reasonable thing to do. In particular, if the output is not metric space.

Also, even if adding noise is a reasonable thing to do, if we do it naively then the privacy loss adds up very quickly. Is there any way of answering a lot of queries?
Oh look, the VC dimension.

Let $H$ be a set of functions that assign to each individual in our database a binary value. Suppose for every function $h \in H$ we want to know the fraction of the population that are given the value 1 by $h$ (these are called statistical queries).
Let $H$ be a set of functions that assign to each individual in our database a binary value. Suppose for every function $h \in H$ we want to know the fraction of the population that are given the value 1 by $h$ (these are called statistical queries).

It turns out the VC dimension can be related to how much data we need in order to be able to answer all these queries accurately while maintaining privacy.

The VC dimension is related to what kind of complex patterns the functions in $H$ can distinguish. So, intuitively, a class $H$ with low VC-dimension only tells us coarse information about the underlying database.
So how many people do I need in my database?

Suppose you want to answer a set of statistical queries from a class $H$ with VC dimension $d$ and you have $k$ types of people in your database. If you have a database with more than

$$C \left( \frac{d \log k + \log \frac{1}{\beta}}{\epsilon \alpha^3} \right)$$

individuals (for some constant $C$) then there exists a $\epsilon$-differentially private algorithm that with probability $1 - \beta$ will output an answer to every query that is within $\alpha$ of the real answer.
So how many people do I need in my database?

Suppose you want to answer a set of statistical queries from a class $H$ with VC dimension $d$ and you have $k$ types of people in your database. If you have a database with more than

$$C \left( \frac{d \log k + \log \frac{1}{\beta}}{\epsilon \alpha^3} \right)$$

individuals (for some constant $C$) then there exists a $\epsilon$-differentially private algorithm that with probability $1 - \beta$ will output an answer to every query that is within $\alpha$ of the real answer.

How do you do this? You actually carefully (privately) choose a synthetic database then compute all the queries on the synthetic database.
Differential privacy is a young field and there are still loads of things we don’t know how to do or even if we can do them!

There is lots of new research on connections to broader fields like statistics and learning theory. We don’t really understand these connections yet.

What type of math is used? Probability theory, information theory, combinatorics, who knows what could come up?