CS 860: Algorithms for Private Data Analysis Fall 2020

Lecture 12 — Private ML and Stats: What is Privacy?

Prof. Gautam Kamath *Scribe: Gautam Kamath*

For the next several lectures, we will cover results in differentially private machine learning and statistics. Before we go into the technical details, today's lecture will be a bit more philosophical in nature. Via a series of examples, we will couch our expectations and try to understand what types of violations differential privacy can and can not protect against. Misunderstanding this point is a common pitfall, even in many published works. Misleadingly, the casual term "privacy" is often not exactly aligned with what differential privacy guarantees.And indeed, under casual use of this term, this type of privacy violation might not be the data analyst's responsibility to prevent, and instead the individual should be more careful with other data they reveal.

Informally, differential privacy implies that an algorithm's output will be similar when a single datapoint is modified. Nothing more and nothing less. In particular, it does *not* prevent statistics and machine learning from working, and anything that they imply. When you are in doubt whether or not differential privacy can prevent some alleged privacy violation in an analysis, ask yourself the following questions:

1. Would the answer truly be fundamentally different if a single datapoint was changed?
2. Is this just machine learning, working as intended?
3. If so, but things still "feel" wrong, at which point was privacy given up?

Another way of saying point 2: if the supposed privacy violation is synonymous with the learned predictor itself being accurate, then it is not a privacy violation. Indeed, the only way that one could possibly avoid this type of situation would be to output an inaccurate predictor, which is easy to do but entirely useless. To quote a movie older than anyone in this class (myself included), "the only winning move is not to play."

Without getting into technical details, we recall what a machine learning algorithm is. A machine learning algorithm takes as input a dataset, consisting of several datapoints, each of which has a feature vector and a label. In the differentially private setting, this dataset is usually the sensitive information. The algorithm outputs a classifier. The classifier is an algorithm which takes in a feature vector, and outputs a prediction of its label. Under differential privacy, this is generally the object which we wish to be differentially private (though there is some work in which we only require the *preidctions* to be private). Its accuracy on this point is measured via some distance between the prediction and the true label, which is usually averaged over an entire test dataset.

Let's start with a very simple but contrived example to demonstrate what "machine learning" can do, and differential privacy does not preclude. Suppose you had the following training data of feature and label pairs: $(0, 1), (11, 12), (\pi, \pi + 1), (-6, -5), (0.5, 1.5), \ldots$. Just by inspecting this data, you would quickly be able to deduce that the pattern is $y = x+1$. Suppose we were tasked with predicting the label of a point $x = 4$, and we output the (correct) prediction $y = 5$. I think everyone would be comfortable with characterizing this as "not a privacy violation" for the training set. But one might instead worry that given the point $x = \pi$ (which appeared in the sensitive training

set), we could correctly predict $y = \pi + 1$! Is *this* a privacy violation? Though instantiations of this phenomenon might seem creepy as we proceed with less contrived examples, this is just science working as intended. In some sense, we are not learning anything private about the points in the dataset themselves, but we are learning something about *nature*. Indeed, going down our checklist, the answer would not have been different if any datapoint was excluded, including the point $(\pi, \pi + 1)$ itself: we still would have learned $y = x + 1$. Correspondingly, there exists a simple differentially private algorithm which would discover precisely this relationship.[1] This is indeed just (a rudimentary form) of machine learning in action. Nonetheless, if this still feels wrong to you: convince yourself that privacy was given up when it was revealed that the point had a value of $x = \pi$. While this example may seem silly, if you fully understand all the implications, you are prepared to judge whether or not something is a privacy violation.

Let's instantiate this with an example we have already seen before. Suppose you are an individual who smokes, and is weighing whether or not to participate in a study testing the hypothesis that smoking causes cancer. You are averse to this information becoming public, as this would lead to your insurance premiums rising. However, this is very similar to the above example. Instead of the classifier being $y = x + 1$, it is "If an individual smokes, they are likely to get cancer," and instead of the point being $(\pi, \pi + 1)$, the point is (smokes, likely to get cancer). Indeed, whether or not one participated in this study, it would be likely to arrive at the same outcome, and the only way to prevent this would be to have not run the study at all. In this case, privacy was relinquished when you disclosed your status as a smoker.

Let's take this to its logical extreme, borrowing an example from Frank McSherry [McS16], stripping out the need for an experiment which simply obfuscates the point. Suppose you see someone wearing a cast on their foot. From this, you can infer that they probably have a broken foot. Even if this person considered their broken foot to be sensitive information, they are publicly providing the information that they are wearing a cast. One can infer the former from the latter using common knowledge, meaning that "wearing a cast" was when privacy was given up.

Here are a couple of examples where machine learning is creepy. See if you can spot when exactly the privacy was surrendered.

One example has to do with Facebook. By default, the set of pages which a user "likes" is public information. In 2013, Kosinski, Stillwell, and Graepel showed that based on an individual's set of likes, it was possible to infer sensitive information, including the user's sexuality or political affiliation [KSG13]. While their analysis was not run with differential privacy, presumably the results would still hold – but regardless, this is a bit of a red herring. Most of the data they used (the set of Facebook likes) was public, and even though they did collect non-public survey information (including sexuality and political affiliation), presumably there would be enough labeled data publicly available on Facebook for anyone to reach the same conclusions that they did. Therefore, in this case, one could consider that privacy was given up when you publicly revealed the pages you liked. Even if you do not consider this to be sensitive, it may serve as a proxy for information you do. One important note is that which of your data is and is not revealing may only be discovered at a later date. For example, if you liked Facebook pages prior to the publication of this paper (at which point they might be considered "safe" to share), it might only be revealed afterwards that they are incriminating.

We provide another example, though this one is apocryphal, and there are doubts regarding its

[1] One could use the stability-based histogram mentioned in Lecture 10.

authenticity. The store Target is known for sending targeted mailers to its customers: based on past purchases, it predicts what the customer might be interested in purchasing next. One young woman in high school is said to have been sent an unusual amount of advertising pertaining to babies and motherhood. This upset her parents, thinking she was far too young to be sent irrelevant material of this sort. However, it was later revealed that the woman was in fact pregnant, and Target managed to infer this prior to her parents based on her past purchases (which were seemingly unrelated to pregnancy). Target's data analysis was not when privacy was lost, it was when the woman provided her purchase information (linked to her identity) to Target. In short, this story is creepy, but is not an instance of Target performing a privacy violation.

We've seen so many things at this point which are *not* privacy violations, you might be wondering what a privacy violation looks like at this point. We will re-inspect some of the instances we've already discussed. First, consider the Netflix Prize, when Netflix released an "anonymized" version of their users' view history. Narayanan and Shmatikov [NS08] linked this with public IMDb data, allowing them to discover movies the users had watched which were not previously public. The privacy violation here was in Netflix's faulty anonymization of the data, which revealed sensitive information about individual users. In particular, if a specific user's data was withheld, then the same reidentification would *not* be possible. This is a fundamental difference between this setting and all the previous ones: the privacy attack was contingent on the individual's sensitive data (i.e., which movies they watched but didn't review on IMDb) being in the dataset. However, consider a similar setting, where the Netflix data was never released and instead, one could infer what movies an individual watched based on what else they reviewed on IMDb. For example, if someone reviewed Rush Hour 1 and Rush Hour 3, you may be able to infer that they watched Rush Hour 2. Even if it happened to be perfectly accurate, this would not be a privacy violation. There is no sensitive dataset, and all the information is public. If anything, privacy was given up when users chose to reveal parts of their viewing history.

As a final example, let's revisit The Secret Sharer [CLE+19]. Recall in this setting, a language model is trained on a sensitive dataset. The model can be given a sequence of characters as input, and will output a log-perplexity score for the sequence (roughly speaking, a lower score corresponds to a higher probability of the sequence). In this setting, we saw that if an individual repeatedly uses a single phrase that is unique to them, the model would be prone to memorizing and could be coerced into revealing this secret phrase. This is again a privacy violation, as it appears the model has overfit to a particular individual's data, and the answer would change significantly if their data were left out. Recalling their example where an individual's social security number was memorized: a perfectly accurate model of English (which is admittedly underdefined) would have no reason to assign significantly higher probability to an individual's SSN versus any other number of the same format.

We have discussed in detail several examples of what is and what is not a privacy violation. This can be a tricky topic to get right, but you should be in good shape if you just ask yourself the questions outlined at the start of the lecture. In the following lectures, we will see how to prevent privacy attacks using differential privacy.

References

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, USENIX Security '19, pages 267–284. USENIX Association, 2019.

[KSG13] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[McS16] Frank McSherry. Statistical inference considered harmful. https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md, June 2016.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 29th IEEE Symposium on Security and Privacy*, SP '08, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.