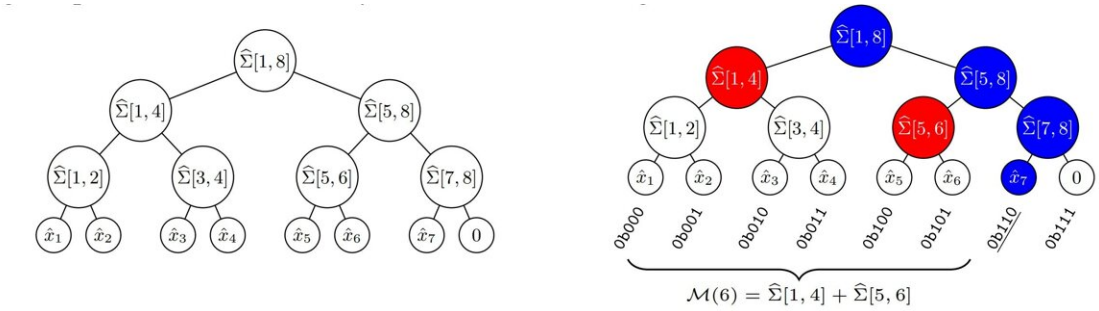
Computer scientists makes noisy data: Can it improve treatments in health care?

January 16 2024



(a) Binary mechanism tree for $T = 7$. (b) Computation of $\mathcal{M}(6)$ using Algorithm 1.

Binary trees for a sequence of length T = 7. In Figure 1(b) each leaf is labeled by bin(t−1), and it illustrates how the prefix sum up to t = 6 can be computed from bin(t). Blue nodes describe the path taken by Algorithm 1, and the sum of red nodes form the desired output *M*(6). Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2306.09666

University of Copenhagen researchers have developed software able to disguise sensitive data such as those used for machine learning in health care applications. The method protects privacy while making datasets available for the development of better treatments.

A key element in modern health care is collecting and analyzing data for a large group of patients to discover patterns. Which patients benefit from a given treatment? And which patients are likely to experience side

effects?

Such data must be protected, or else the privacy of individuals is broken. Furthermore, breaches will harm general trust, leading to fewer people giving their consent to take part. Researchers at the Department of Computer Science, University of Copenhagen, have found a clever solution.

"We have seen several cases in which data was anonymized and then released to the public, and yet researchers managed to retrieve the identities of participants. Since many other sources of information exist in the public domain, an adversary with a good computer will often be able to deduct the identities even without names or citizen codes."

"We have developed a practical and economical way to protect datasets when used to train machine learning models," says Ph.D. student Joel Daniel Andersson.

The level of interest in the new algorithm can be illustrated by the fact that Joel was invited to give a Google Tech Talk on it. Also, he recently held a presentation at NeurIPS conference on machine learning.

Deliberately polluting your output

The key idea is to mask your dataset by adding "noise" to any output derived from it. Unlike encryption, where noise is added and later removed, in this case, the noise stays. Once the noise is added, it cannot be distinguished from the "true" output.

Obviously, the owner of a dataset should not be happy about noising outputs derived from it.

"A lower utility of the dataset is the necessary price you pay for ensuring

the privacy of participants," says Joel Daniel Andersson.

The key task is to add an amount of noise sufficient to hide the original data points, but still maintain the fundamental value of the dataset, he notes:

"If the output is sufficiently noisy, then it becomes impossible to infer the value of an individual data point in the input, even if you know every other data point. By noising the output, we are in effect adding safety rails to the interaction between the analyst and the dataset."

"The analysts never access the raw data, they only ask queries about it and get noisy answers. Thereby, they never learn any information about individuals in the dataset. This protects against information leaks, inadvertent or otherwise, stemming from analysis of the data."

Privacy comes with a price tag

There is no universal optimal trade-off. Joel Daniel Andersson says, "You can pick the trade-off which fits your purpose. For applications where privacy is highly critical—for instance, health care data—you can choose a very high level of privacy. This means adding a large amount of noise."

"Notably, this will sometimes imply that you will need to increase your number of data points—so include more persons in your survey, for instance—to maintain the value of your dataset. In applications where privacy is less critical, you can choose a lower level. Thereby, you will maintain the utility of your dataset and reduce the costs involved in providing privacy."

Reducing costs is exactly the prime argument behind the method developed by the research group, he adds. "The crux is how much noise

you must add to achieve a given level of privacy, and this is where our smooth mechanism offers an improvement over existing methods. We manage to add less noise and do so with fewer computational resources. In short, we reduce the costs associated with providing privacy."

Interest from industry

Machine learning involves large datasets. For instance, in many health care disciplines, a computer can find patterns that human experts cannot see. This all starts with training the computer on a dataset with real patient cases. Such training sets must be protected.

"Many disciplines depend increasingly on machine learning. Further, we see machine learning spreading beyond professionals like medical doctors to various private applications. These developments open a wealth of new opportunities, but also increases the need for protecting the privacy of the participants who provided the original data," explains Joel Daniel Andersson, noting that interest in the groups' new software is far from just academic:

"Besides the health care sector plus Google and other large tech companies, industry like consultants, auditing firms, and law firms need to be able to protect the privacy of their clients and participants in surveys."

Public regulation is called for

The field is known as differential privacy. The term is derived from the privacy guarantee for datasets differing in a single data point: output based on two datasets differing only in one data point will look similar. This makes it impossible for the analyst to identify a single data point.

The research group advocates for public bodies to take a larger interest in the field.

"Since better privacy protection comes with a higher price tag due to the loss of utility, it easily becomes a race to the bottom for market actors. Regulation should be in place, stating that a given sensitive application needs a certain minimum level of privacy. This is the real beauty of differential privacy."

"You can pick the level of privacy you need, and the framework will tell you exactly how much noise you will need to achieve that level," says Joel Daniel Andersson. He hopes that differential privacy may serve to facilitate the use of machine learning.

"If we again take medical surveys as an example, they require patients to give consent to participate. For various reasons, you will always have some patients refusing—or just forgetting—to give consent, leading to a lower value of the dataset. However, since it is possible to provide a strong probabilistic guarantee that the privacy of participants will not be violated, it could be morally defensible to not require consent and achieve 100 % participation to the benefit of the medical research."

"If the increase in participation is large enough, the loss in utility from providing privacy could be more than offset by the increased utility from the additional data. As such, differential privacy could become a win-win for society."

The work is published on the *arXiv* preprint server.

More information: Joel Daniel Andersson et al, A Smooth Binary Mechanism for Efficient Private Continual Observation, *arXiv* (2023). DOI: 10.48550/arxiv.2306.09666

Provided by University of Copenhagen