IBM Differential Privacy Library: The single line of code that can protect your data

Naoise Holohan

IBM published a new release of its IBM Differential Privacy Library, which boasts a suite of tools for machine learning and data analytics tasks, all with built-in privacy guarantees. It's not unlike the the differential privacy the US Census will use to keep the responses of its citizens confidential when the data is made available.

This year for the first time in its 230-year history the US Census will use differential privacy to keep the responses of its citizens confidential when the data is made available. But how does it work?

Differential privacy uses mathematical noise to preserve individuals' privacy and confidentiality while allowing population statistics to be observed. This concept has a natural extension to machine learning, where we can protect models against privacy attacks, while maintaining overall accuracy.

For example, if you want to know my age (32) I can pick a random number out of a hat, say ±7 – you will only learn that I could be between 25 and 39. I've added a little bit of noise to the data to protect my age and the US Census will do something similar.

While the US government built its own differential privacy tool, IBM has been working on its own open source version and today we are publishing our latest release v0.3 . The IBM Differential Privacy Library boasts a suite of tools for machine learning and data analytics tasks, all with built-in privacy guarantees.

Our library is unique to others in giving scientists and developers access to lightweight, user-friendly tools for data analytics and machine learning in a familiar environment – in fact, most tasks can be run with only a single line of code.

What also sets our library apart is our machine learning functionality enables organizations to publish and share their data with rigorous guarantees on user privacy like never before.[1]

Technical details

With v0.3, the library now comes with a budget accountant to track privacy budget spend across different operations. Using advanced composition techniques, the budget accountant allows users to extract even more insight than simpler accounting methods and while it's hard to quantify, under typical workloads, privacy budget savings in excess of 50 percent are not uncommon.

Our library includes an array of functionality to extract insight and knowledge from data with robust privacy guarantees. We have focused on developing solutions for the most popular algorithms, including histograms, logistic regression, k-means clustering and principal component analysis (PCA), as well as giving developers the basic building blocks of differential privacy to allow them to develop their own custom solutions.

The library includes the following key components which don't exist in similar libraries currently

available:

- **Accountant:** Track and limit privacy spend across multiple operations;
- **Mechanisms:** A comprehensive collection of the basic building blocks of differential privacy, used to build new tools and applications;
- **Machine learning:** Machine learning algorithms for pre-processing, classification, regression and clustering. Also included is a collection of fundamental tools for data exploration and analytics. All the details for getting started with the library can be found at IBM's Github repository .