Finding patterns in corrupted data: New model-fitting technique efficient even for data sets with hundreds of variables

October 26 2016, by Larry Hardesty

A team, including researchers from MIT's Computer Science and Artificial Intelligence Laboratory, has created a new set of algorithms that can efficiently fit probability distributions to high-dimensional data. Credit: MIT News

Data analysis—and particularly big-data analysis—is often a matter of fitting data to some sort of mathematical model. The most familiar example of this might be linear regression, which finds a line that approximates a distribution of data points. But fitting data to probability distributions, such as the familiar bell curve, is just as common.

If, however, a data set has just a few corrupted entries—say, outlandishly improbable measurements—standard data-fitting techniques can break down. This problem becomes much more acute with high-dimensional data, or data with many variables, which is ubiquitous in the digital age.

Since the early 1960s, it's been known that there are algorithms for weeding corruptions out of high-dimensional data, but none of the algorithms proposed in the past 50 years are practical when the variable count gets above, say, 12.

That's about to change. Earlier this month, at the IEEE Symposium on Foundations of Computer Science, a team of researchers from MIT's Computer Science and Artificial Intelligence Laboratory, the University of Southern California, and the University of California at San Diego presented a new set of algorithms that can efficiently fit probability distributions to high-dimensional data.

Remarkably, at the same conference, researchers from Georgia Tech presented a very similar algorithm.

The pioneering work on "robust statistics," or statistical methods that can tolerate corrupted data, was done by statisticians, but both new papers come from groups of computer scientists. That probably reflects a shift of attention within the field, toward the computational efficiency of model-fitting techniques.

"From the vantage point of theoretical computer science, it's much more apparent how rare it is for a problem to be efficiently solvable," says Ankur Moitra, the Rockwell International Career Development Assistant Professor of Mathematics at MIT and one of the leaders of the MIT-USC-UCSD project. "If you start off with some hypothetical thing—'Man, I wish I could do this. If I could, it would be robust'—you're going to have a bad time, because it will be inefficient. You should start off with the things that you know that you can efficiently do, and figure out how to piece them together to get robustness."

Resisting corruption

To understand the principle behind robust statistics, Moitra explains, consider the normal distribution—the bell curve, or in mathematical parlance, the one-dimensional Gaussian distribution. The one-dimensional Gaussian is completely described by two parameters: the mean, or average, value of the data, and the variance, which is a measure of how quickly the data spreads out around the mean.

If the data in a data set—say, people's heights in a given population—is well-described by a Gaussian distribution, then the mean is just the arithmetic average. But suppose you have a data set consisting of height measurements of 100 women, and while most of them cluster around 64 inches—some a little higher, some a little lower—one of them, for some reason, is 1,000 inches. Taking the arithmetic average will peg a woman's mean height at 6 feet 4 inches, not 5 feet 4 inches.

One way to avoid such a nonsensical result is to estimate the mean, not by taking the numerical average of the data, but by finding its median value. This would involve listing all the 100 measurements in order, from smallest to highest, and taking the 50th or 51st. An algorithm that uses the median to estimate the mean is thus more robust, meaning it's

less responsive to corrupted data, than one that uses the average.

The median is just an approximation of the mean, however, and the accuracy of the approximation decreases rapidly with more variables. Big-data analysis might require examining thousands or even millions of variables; in such cases, approximating the mean with the median would often yield unusable results.

Identifying outliers

One way to weed corrupted data out of a high-dimensional data set is to take 2-D cross sections of the graph of the data and see whether they look like Gaussian distributions. If they don't, you may have located a cluster of spurious data points, such as that 80-foot-tall woman, which can simply be excised.

The problem is that, with all previously known algorithms that adopted this approach, the number of cross sections required to find corrupted data was an exponential function of the number of dimensions. By contrast, Moitra and his coauthors—Gautam Kamath and Jerry Li, both MIT graduate students in electrical engineering and computer science; Ilias Diakonikolas and Alistair Stewart of USC; and Daniel Kane of USCD—found an algorithm whose running time increases with the number of data dimensions at a much more reasonable rate (or, polynomially, in computer science jargon).

Their algorithm relies on two insights. The first is what metric to use when measuring how far away a data set is from a range of distributions with approximately the same shape. That allows them to tell when they've winnowed out enough corrupted data to permit a good fit.

The other is how to identify the regions of data in which to begin taking cross sections. For that, the researchers rely on something called the

kurtosis of a distribution, which measures the size of its tails, or the rate at which the concentration of data decreases far from the mean. Again, there are multiple ways to infer kurtosis from data samples, and selecting the right one is central to the algorithm's efficiency.

The researchers' approach works with Gaussian distributions, certain combinations of Gaussian distributions, another common distribution called the product distribution, and certain combinations of product distributions. Although they believe that their approach can be extended to other types of distributions, in ongoing work, their chief focus is on applying their techniques to real-world data.

More information: Robust estimators in high dimensions without the computational intractability. arxiv.org/pdf/1604.06443v1.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Finding patterns in corrupted data: New model-fitting technique efficient even for data sets with hundreds of variables (2016, October 26) retrieved 1 May 2024 from https://techxplore.com/news/2016-10-patterns-corrupted-model-fitting-technique-efficient.html