Combining computer science, statistics creates machines that can learn

July 17 2013

Learning a subject well means moving beyond the recitation of facts to a deeper knowledge that can be applied to new problems. Designing computers that can transcend rote calculations to more nuanced understanding has challenged scientists for years. Only in the past decade have researchers' flexible, evolving algorithms—known as machine learning—matured from theory to everyday practice, underlying search and language-translation websites and the automated trading strategies used by Wall Street firms.

These applications only hint at machine learning's potential to affect daily life, according to John Lafferty, the Louis Block Professor in Statistics and Computer Science. With his two appointments, Lafferty bridges these disciplines to develop theories and methods that expand the horizon of machine learning to make predictions and extract meaning from data.

"Computer science is becoming more focused on data rather than computation, and modern statistics requires more computational sophistication to work with large data sets," Lafferty says. "Machine learning draws on and pushes forward both of these disciplines."

Lafferty's work focuses on the theories and algorithms that power machine learning. The goal is to develop computer programs that, with little or no human input, can extract knowledge from large amounts of numbers, text, audio or video and make predictions and decisions about events that haven't been coded in its instructions.

"The classical areas of applied mathematics, including partial differential equations, developed from the study of physical processes such as fluid flow," Lafferty says. "What we're seeing now is that entirely new directions in applied mathematics are opening up from the study of modern large data sets."

As big data becomes more common in fields including astronomy, biology, and the humanities, researchers need new statistical techniques to reveal meaningful signals amid the noise. Machine learning powers advanced technologies from face and speech recognition to cars that drive themselves, and scientists hope to apply it to personalized medical treatments—all problems where computer programs must make decisions based on a flood of data, much of it previously unseen.

Lafferty joined the University in 2011 as part of the Physical Science Division's computational and applied mathematics initiative, launched in 2008 to recruit faculty and students in these areas. Before that he helped to found the world's first machine-learning department at Carnegie Mellon University in 2002 and spent several years at the IBM Thomas J. Watson Research Center, working on early machine-learning projects in natural speech and text processing.

At Carnegie Mellon, Lafferty's projects included building a topic model, with Dave Blei of Princeton University, from the complete database of articles in the journal *Science* since its 1880 founding. By connecting semantically related groups of words that appear together in papers more frequently than random chance, the topic model carves a structure from this enormous text database. Researchers can then explore the journal by following terms relevant to their field, finding overlooked papers that simple search engines might not spot.

Today Lafferty studies the balancing act between the competing demands of computational efficiency and statistical accuracy, finding

new ways to handle high-dimensional data sets. While traditional statistics is focused on "tall and thin" data, with many records and few variables, modern data sets, such as those in genomics, are often "short and wide," featuring few subjects and tens of thousands of variables. Researchers need ways to sort the relevant factors from the irrelevant to make statistical analysis possible.

Lafferty also studies semi-supervised learning, a machine-learning technique where a human trains the computer to categorize inputs, such as speech or images, and then turns it loose on new, unseen data. At Carnegie Mellon, Lafferty and colleagues developed a computer-vision program to recognize people on a webcam that graduate students had set up in the computer-science department lounge to watch for seminar leftovers. After an observer entered labels the first day of data collection, the program was more than 80 percent accurate in identifying ten individuals who appeared regularly over four months.

This spring Lafferty is teaching machine learning to a new generation of scientists. In the course Machine Learning and Large-Scale Data Analysis, undergraduates develop algorithms to predict Chicago crime, search for exoplanets, find New Year's Day wishes on Twitter, and study the language in State of the Union addresses, running their analyses on virtual computer clusters built in Amazon Web Services. The students enrolled in the course include majors in physics, mathematics, computer science, linguistics, economics, neuroscience and political science, reflecting the wide relevance of machine learning to today's research world.

While many machine-learning experts are drawn to companies such as Google to work on specific applications, Lafferty sees broader and more surprising work coming from research universities. "The potential impact is very large," Lafferty says, "and the ideas that we're developing will be applied in ways that we can't even anticipate."

Provided by University of Chicago