Transparency reports make AI decision-making accountable

26 May 2016

Machine-learning algorithms increasingly make decisions about credit, medical diagnoses, personalized recommendations, advertising and job opportunities, among other things, but exactly how usually remains a mystery. Now, new measurement methods developed by Carnegie Mellon University researchers could provide important insights to this process.

Was it a person's age, gender or education level that had the most influence on a decision? Was it a particular combination of factors? CMU's Quantitative Input Influence (QII) measures can provide the relative weight of each factor in the final decision, said Anupam Datta, associate professor of computer science and electrical and computer engineering.

"Demands for algorithmic transparency are increasing as the use of algorithmic decision-making systems grows and as people realize the potential of these systems to introduce or perpetuate racial or sex discrimination or other social harms," Datta said.

"Some companies are already beginning to provide transparency reports, but work on the computational foundations for these reports has been limited," he continued. "Our goal was to develop measures of the degree of influence of each factor considered by a system, which could be used to generate transparency reports."

These reports might be generated in response to a particular incident—why an individual's loan application was rejected, or why police targeted an individual for scrutiny or what prompted a particular medical diagnosis or treatment. Or they might be used proactively by an organization to see if an artificial intelligence system is working as desired, or by a regulatory agency to see whether a decision-making system inappropriately discriminated between groups of people.

Datta, along with Shayak Sen, a Ph.D. student in computer science, and Yair Zick, a post-doctoral researcher in the Computer Science Department, will present their report on QII at the IEEE Symposium on Security and Privacy, May 23-25, in San Jose, Calif.

Generating these QII measures requires access to the system, but doesn't necessitate analyzing the code or other inner workings of the system, Datta said. It also requires some knowledge of the input dataset that was initially used to train the machine-learning system.

A distinctive feature of QII measures is that they can explain decisions of a large class of existing machine-learning systems. A significant body of prior work takes a complementary approach, redesigning machine-learning systems to make their decisions more interpretable and sometimes losing prediction accuracy in the process.

QII measures carefully account for correlated inputs while measuring influence. For example, consider a system that assists in hiring decisions for a moving company. Two inputs, gender and the ability to lift heavy weights, are positively correlated with each other and with hiring decisions. Yet transparency into whether the system uses weight-lifting ability or gender in making its decisions has substantive implications for determining if it is engaging in discrimination.

"That's why we incorporate ideas for causal measurement in defining QII," Sen said. "Roughly, to measure the influence of gender for a specific individual in the example above, we keep the weight-lifting ability fixed, vary gender and check whether there is a difference in the decision."

Observing that single inputs may not always have high influence, the QII measures also quantify the joint influence of a set of inputs, such as age and income, on outcomes and the marginal influence of

each input within the set. Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using principled game-theoretic aggregation measures previously applied to measure influence in revenue division and voting.

"To get a sense of these influence measures, consider the U.S. presidential election," Zick said. "California and Texas have influence because they have many voters, whereas Pennsylvania and Ohio have power because they are often swing states. The influence aggregation measures we employ account for both kinds of power."

The researchers tested their approach against some standard machine-learning algorithms that they used to train decision-making systems on real data sets. They found that the QII provided better explanations than standard associative measures for a host of scenarios they considered, including sample applications for predictive policing and income prediction.

Now, they are seeking collaboration with industrial partners so that they can employ QII at scale on operational machine-learning systems.

Provided by Carnegie Mellon University