Jewel Chen
MATH0218
May 5, 2017

In fields such as economics and psychology, where robust statistics are critical to deriving causal inferences from research, machine learning offers a great opportunity to improve upon techniques that are currently limited by processing power. In Professor Susan Athey's interview conducted by Russ Roberts, she discusses the advantages of incorporating machine learning into econometrics while cautioning against a blind application of the technique.

Counterfactuals are commonly used to infer causality in economics. Specifically, in order to prove that a certain factor causes – as opposed to correlates with – the outcome, it is important to note what the outcome would have been if the factor had not been present, all other variables constant. Unfortunately, few perfect counterfactual set-ups exist in real life. Machine learning can help offset this dearth by making it easier to generate synthetic counterfactuals – or a hypothetical counterfactual based on aggregating elements and trends from various real-life sources. In the future, it may be easy to create an ideal counterfactual town with demographics from one city and the education spending from another city when a natural counterfactual for studying the effect of education stimulus packages on different demographics does not exist.

Machine learning is also adept at processing large quantities of data, and may be particularly useful when the data set presents with more variables than observations. Algorithms within machine learning are able to determine which variables are significant and which clusters of covariates are spurious over the entire data set, and can therefore quickly eliminate noise. The danger in this approach, however, lies in the interpretation of the remaining variables. Machine learning can hone in on important predictive elements, but these elements may not necessarily be causal in nature. If machine learning is applied blindly without careful consideration to the model and its implications, the results may be non-robust. Athey thus, cautions the economist to distinguish between the causal interest and the rest of the variables, and to treat the two different differently. Additionally, she warns against giving causal interpretation to the non-causal variables. Lastly, she suggests the use of sample splitting or cross validation to protect against overfitting the model.

There are many ways in which machine learning can contribute to econometrics. Traditionally, economists have been able to make causal inferences using small data sets, but are limited by their samples and also by the need for counterfactuals. Machine learning, on the other hand, cannot derive causal inferences on its own, but is able to determine both trends and significant predictive factors over large sets of data. The combination of the two, if applied mindfully, could provide economists with even more powerful tools to understand the relationships in the world around us.