Aayam Poudel, Computer Science, '03

How can Machine Learning algorithms help take the 'con' out of 'econometrics'?

Econometrics relies on economists studying economic systems using mathematical models. To do so, economists look at a bunch of variables that are present in these economic systems, and try to determine causal relationships between these variables to gather information about the system. For example, using econometrics, economists would look at how a college education might affect the income of people. They would look at the existing data of people who went to college and who did not go to college and figure out the relationship between a college education and income. They could come up with findings that say, "People with college education tend to have higher income than people without college education". But the fact might actually be that other variables like socio-economic status might affect whether a person will attend college or not, and even though it looks like college education might have a causal relationship with income, it might actually not exactly be so.

To look at true causal relationships we have to look at both the factual and counterfactual cases, i.e. "the true scenario" and "what might have happened". In our case, we have to look at what the income of people is because they attended college, and what the income might have been if they had never attended college to see how much of higher earnings is due to college education and how much of it is due to high socio economic status. Getting the counterfactual is not quite possible without Machine Learning (ML).

ML is a way of teaching computers how to make predictions on new data based on past datasets. Economists generally have disregarded ML in the past because it treats its models like a black box. They want to know how exactly variables affect models; they want interpretably in their models, and ML does not really provide this. When we are training a model using ML and making predictions on a dataset, we do not really care about what variable causes what changes, and how much each variable affects other variables; we only care about getting good predictions. Although this might seem like a very unscientific way of making predictions – it seems unscientific because we have no theoretical understanding of what is going on in the model – it does work really well. We could use ML algorithms to perform regularization on predictors when we have too many variables compared to the number of rows in the dataset. We could also use these algorithms to perform cross validation on the dataset and see which instance of the model (for example in case of splines, what is the optimal degree of freedom) makes best predictions. This way, even though we do not really know how the ML black box treats the modeling process, we still get good predictions. For our purposes, ML could be a great tool to help us get counterfactuals, even though we might not understand what exactly is going on inside the black box.

We should combine ML with econometrics to make better interpretations. We could get the factual and counterfactual information needed for our model through ML so that we could infer proper causal relationships in our data and make better interpretations. Rather than treating ML's black box approach like unscientific way of making predictions, we could think of them as valuable resource in econometrics, and use them as a tool to take the 'con' out of 'econometrics'.