Kyra Gray ‘17
Statistical Learning
Professor Kim
5/5/17

How can machine learning methods can help take the ‘con’ out of ‘econometrics’?

While machine learning won’t address the important econometric question of correlation versus causation, it can be very useful in the realm of econometrics; especially when the nature of the data is that you are working with many more variables to control for than observations, as is often the case. One method that can be implemented, albeit very carefully, is data mining and the corresponding concept of distinguishing between causal variables and everything else, i.e applying different treatment to outcome variables and predictors. However, it is important to acknowledge and understand that if you’re using machine learning algorithms to determine which covariates should go into your predictive model, you cannot in good faith assign a causal interpretation to those decisions, because it could have been the case that a lot of variables were high correlated and the mechanism of the machine learning regularization process essentially picks some variables and not others and the variables that are ultimately picked are because they can act as a proxy for all of the other variables. However, this inability to assign a causal inference interpretation to the choice of model covariates shouldn't be a problem in econometrics, because you in theory are only looking to apply a causal interpretation to whatever change you are interested in the effect of (e.g. a minimum wage policy change or taking a drug). Another useful machine learning method that can be applied to econometrics to protect against model over fitting is sample splitting by using half of your data to figure out what the right model is and then another half of the data which is “clean” (not included in the model) to estimate your effects.