Phil Hoxie, Economics, ‘17.5
Problem Set 10
Math 218—Machine Learning
May 5, 2017

Using Machine Learning to aid Economic Theory

Theory and measurement are the two tools that define empirical economics. Economists are concerned with understanding causal relationship between two variables. In order to do this, economists use theory to justify statistical tests used to measure the marginal effects of one variable on another. The classic problem with this is how the outcome variable is measured and whether or not the statistical test matches its given interpretation. This means that economists must be deliberate in their research to avoid running too many statistical tests and therefor misinterpreting a random result for a causal one. This process of over testing is called p-hacking. Machine leaning can aid economists in solving measurement problems if it used in a way that does not result in p-hacking.

Economists often times cannot ethically create randomized control trials, which means they must find alternative means for finding causal relationships. This fact of economics creates the need for alternative specifications that closely approximate a randomized trial. Difference-in-difference, regression discontinuity, and instrumental variables are all techniques that have been developed to approximate treatment and control groups. However, when using real world data, oftentimes specifications between the treatment and control groups are not balanced. This created a problem of how to control for these real world inconveniences. Theory is useful for identifying important controls without running into the problem of p-hacking.

Theory is useful to economists in identifying controls for models without running additional statistical tests. There are times in which there may be so many available variables that theory alone cannot give insights into all the possible controls for a given model. Machine learning and resampling methods can help solve this problem. Shrinkage, or regularization, methods can be used to identify the variables that are most important for predicting a given outcome variable. However, the nature of cross validation is that it runs additional tests to minimize error in predicting the outcome variable. By resampling the data into two pieces, economics can use shrinkage methods on one piece of data to identify important controls from hundreds of choices and then use those controls in a model on the remaining data. This method of resampling and shrinkage maintains the causal relationship for the coefficient of interest in the final model but sacrifices some of the sample. This is a clear tradeoff, but in circumstances that have hundreds of variables to choose from and ample observations, this method would greatly improve accurate measurement of the true relationship and remove some researcher bias.

If a researcher is trying to understand the effect of age on unemployment, there are dozens of possible controls to choose from. Shrinkage can help identify which variables are important to control for. It is important to note that in the final model, the controls themselves will not have a causal interpretation, but the coefficient of interest will. Theory is mostly supported by empirical work, but using machine learning to improve model specification could remove some of the art from economic research. A more objective process for identifying controls could greatly aid in improving the external validity of models when used correctly.