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

Will Ernst, Computer Science & Economics, '17

Although both traditional econometrics and machine learning techniques are used for data analysis, there is a clear distinction between the two fields. The former is best suited for small data causal inference, while the latter is best suited for big data prediction. Additionally, econometricians build their models based on intuition and theories, while machine learning techniques begin with a fully data-driven approach. Because econometricians' goal is to find causality, they generally tend to avoid ML techniques, which may result in spurious relationships between covariates. But is there room for ML techniques to improve the capabilities of traditional econometric tools?

One area of improvement for econometricians is the selection of variables to use when building a model. The more independent variables included in a model, the fewer degrees of freedom that model has, and thus the model's variance rises due to overfitting. In general, econometricians use theoretical underpinnings to find the best causal links between variables, but they may be missing some. In cases like this, regularization methods (such as Ridge Regresison or LASSO) could be helpful in determining which estimators have the most significant impact on the dependent variable and should be included in the model. This is especially helpful when one has large number of possible predictors to choose from. Cross validation techniques could be another helpful ML tool used by econometricians. This helps avoid overfitting issues when using other ML tools. Additionally, Principal Components Analysis or other clustering tools can be used to find statistically similar variables and help improve a model's variance with data driven selection.

Prediction can also be helpful when you need to control for a counterfactual – for something that *would have* happened if there was or wasn't some change. One may create a synthetic control group, which is a prediction of the counterfactual based on variables with similar difference of differences. This prediction allows the model to take into account something that realistically cannot be observed. As previously stated, ML tools are particularly good at prediction and selecting variables based on similarities in data. ML tools then could be used in this sort of analysis to select proper variables for a synthetic control group and generate the data for the group with prediction methods.

Econometrics and ML are distinct fields in their purposes (causal inference and prediction, respectively). But, econometrics methods are a somewhat limited due to the difficulty of establishing causal inference in complex systems. ML methods may be useful in conjunction with econometrics in building models in complex systems with large amounts of data and predictors, but should not be used for the causal inference itself.