Mathematics, Statistics, and Computer Science **@ UIC**

Statistics Seminar

Understanding the Effects of Predictor Variables in Black Box Supervised Learning Models

Daniel W. Apley (Northwestern University)

Abstract: For many supervised learning applications, understanding and visualizing the effects of the predictor variables on the predicted response is of paramount importance. A shortcoming of black box supervised learning models (e.g., complex trees, neural networks, boosted trees, random forests, nearest neighbors, local kernel-weighted methods, support vector regression, etc.) in this regard is their lack of interpretability or transparency. Partial dependence (PD) plots, which are the most popular general approach for visualizing the effects of the predictors with black box supervised learning models, can produce erroneous results if the predictors are strongly correlated, because they require extrapolation of the response at predictor values that are far outside the multivariate envelope of the training data. Functional ANOVA for correlated inputs can avoid this extrapolation but involves prohibitive computational expense and subjective choice of additive surrogate model to fit to the supervised learning model. We present a new visualization approach that we term accumulated local effects (ALE) plots, which have a number of advantages over existing methods. First, ALE plots do not require unreliable extrapolation with correlated predictors. Second, they are orders of magnitude less computationally expensive than PD plots, and many orders of magnitude less expensive than functional ANOVA. Third, they yield convenient variable importance/sensitivity measures that possess a number of desirable properties for quantifying the impact of each predictor.

Wednesday, April 4 at 4:00 PM in SEO 636