MICHIGAN STATE UNIVERSITY
Department of Statistics and Probability

COLLOQUIUM

Lucas Mentch
University of Pittsburgh

Random Forests: Why They Work and Why That's a Problem

Tuesday, October 19, 2021
10:20 AM - 11:10 AM Eastern Time
Zoom

Abstract

Random forests remain among the most popular off-the-shelf supervised machine learning tools with a well-established track record of predictive accuracy in both regression and classification settings. Despite their empirical success, a full and satisfying explanation for their success has yet to be put forth. In this talk, we will show that the additional randomness injected into individual trees serves as a form of implicit regularization, making random forests an ideal model in low signal-to-noise ratio (SNR) settings. From a model-complexity perspective, this means that the mtry parameter in random forests serves much the same purpose as the shrinkage penalty in explicit regularization procedures like the lasso. Realizing this, we demonstrate that alternative forms of randomness can provide similarly beneficial stabilization. In particular, we show that augmenting the feature space with additional features consisting of only random noise can substantially improve the predictive accuracy of the model. This surprising fact has been largely overlooked within the statistics community, but has crucial implications for thinking about how best to define and measure variable importance. Numerous demonstrations on both real and synthetic data are provided.

Zoom details can be found at: https://stt.natsci.msu.edu/stt-colloquium-zoom-info/

To request an interpreter or other accommodations for people with disabilities, please call the Department of Statistics and Probability at 517-355-9589.