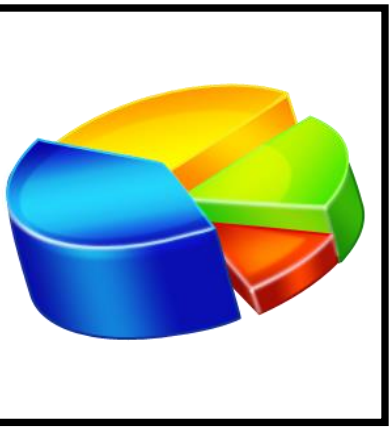
STATISTICS SEMINAR

Lucas Mentch
Cornell University

Monday, February 9, 2015
Leasure Hall, Room 13, 4:00-5:00 pm
Refreshments: Dickens 108, 3:30-4:00 pm



Ensemble Trees and CLTs: Using Subsamples to Peek Inside the Black Box

Abstract: Modern learning algorithms are typically seen as prediction-only tools, meaning that the interpretability and intuition provided by a more traditional modeling approach are sacrificed in order to achieve superior predictions. In this talk, we argue that this black-box perspective need not always be the case. We demonstrate that predictions from ensemble learners like bagged trees and random forests, when built with subsamples in lieu of full bootstrap samples, can be viewed as incomplete, infinite-order U-statistics and as such, are asymptotically normal. Furthermore, we show that the limiting variance depends only on the size of the ensemble relative to the size of the training set and by enforcing a structure on the subsamples used in the ensemble, we can form a consistent estimate of variance at no additional computational cost. This allows for statistical inference to be carried out in practice and in particular, we can produce confidence intervals to accompany predictions and define formal hypothesis tests for both additivity and feature significance. When a large test set is required, we extend our testing procedures and utilize random projections to accommodate the potential p>>n setting. These tools are illustrated on data provided by Cornell University's Lab of Ornithology.