Statistics Seminar

Department of Mathematics and Statistics

DATE:	Thursday, October 17, 2024
TIME:	1:15pm – 2:15pm
LOCATION:	WH 100E
SPEAKER:	Samruddhi Thakar, Binghamton University
TITLE:	Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests

Abstract

This work develops formal statistical inference procedures for predictions generated by supervised learning ensembles. Ensemble methods based on bootstrapping, such as bagging and random forests, have improved the predictive accuracy of individual trees, but fail to provide a framework in which distributional results can be easily determined. Instead of aggregating full bootstrap samples, we consider predicting by averaging over trees built on subsamples of the training set and demonstrate that the resulting estimator takes the form of a U-statistic. As such, predictions for individual feature vectors are asymptotically normal, allowing for confidence intervals to accompany predictions. In practice, a subset of subsamples is used for computational speed; here our estimators take the form of incomplete U-statistics and equivalent results are derived. We further demonstrate that this setup provides a framework for testing the significance of features. Moreover, the internal estimation method we develop allows us to estimate the variance parameters and perform these inference procedures at no additional computational cost. Simulations and illustrations on a real data set are provided.

Reference:

Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. Journal of Machine Learning Research, 17(26), 1-41.