THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

DISSERTATION PRESENTATION AND DEFENSE

ANDREW GOLDSTEIN

Department of Statistics
The University of Chicago

A Variational Bayesian Approach for Combining Weak Learners into a Strong Learner in Regression Problems

WEDNESDAY, September 14, 2022, at 12:00 PM
Jones 304, 5747 S. Ellis Avenue

ABSTRACT

One of the pillars of machine learning is that of non-linear regression on tabular data.

For the last few decades, the performance of ensemble methods based on a sum-of-trees model (gradient boosting and random forest methods in particular) has been start-of-theart (Shwartz-Ziv and Armon [2022], Chen and Guestrin [2016], Fernández-Delgado et al. [2014]). However, such methods can suffer from a few weaknesses; in particular, they often require time-consuming cross-validation procedures to tune a slew of hyper-parameters, and they provide no level of uncertainty about their predictions. Bayesian methods can address both through the use of hierarchical modelling. But many methods rely on Markov chain Monte Carlo (MCMC) methods that can be slow and scale poorly, not to mention the further complications that can arise due to poor mixing of the Markov chain.

In this dissertation, we introduce a new Bayesian framework, VEB-Boost, that aims to address these challenges (implemented in our R package VEB.Boost). In particular, it relies on empirical Bayes and variational inference, allowing us to bypass hyper-parameter tuning while being able to scale well. In the VEB-Boost framework, we combine weak learners (à la boosting) by adding and/or multiplying them together in an arbitrary order. Doing so yields a modular fitting procedure that reduces to iteratively fitting a single weak learner at a time.

We demonstrate the potential of VEB-Boost with a simulation study and real-dataset benchmarking analysis. We also show how to extend the VEB-Boost model to non-Gaussian response data. We derive extensions for: logistic regression, multinomial logistic regression, negative binomial regression, accelerated failure time models, ordinal logistic regression, Bradley-Terry pairwise ranking models, Plackett-Luce listwise ranking models, Cox proportional hazards models, and multivariate Gaussian regression. Many of these approximations are new, and we note some interesting connections between them. Lastly, we demonstrate the logistic and multinomial logistic models in a simulation study and real-data benchmarking analysis.