Inference on Deep Neural Network Estimators

Yi LI

Department of Biostatistics, University of Michigan, Ann Arbor, USA

ABSTRACT

While deep neural networks (DNNs) are used for prediction, inference on DNN-estimated subject-specific means for categorical or exponential family outcomes remains underexplored. We address this by proposing a DNN estimator under generalized nonparametric regression models (GNRMs) and developing a rigorous inference framework. Unlike existing approaches that assume independence between estimation errors and inputs to establish the error bound, a condition often violated in GNRMs, we allow for dependence and our theoretical analysis demonstrates the feasibility of drawing inference under GNRMs. To implement inference, we consider an Ensemble Subsampling Method (ESM) that leverages U-statistics and the Hoeffding decomposition to construct reliable confidence intervals for DNN estimates. We show that, under GNRM settings, ESM enables model-free variance estimation and accounts for heterogeneity among individuals in the population.

Through simulations under nonparametric logistic, Poisson, and binomial regression models, we demonstrate the effectiveness and efficiency of our method. We further apply the method to the electronic Intensive Care Unit (eICU) dataset, a large scale collection of anonymized health records from ICU patients, to predict ICU readmission risk and offer patient-centric insights for clinical decision making.

Keywords: deep neural network; ensemble estimation; nonparametric regression.

Causal Learning with Label Noise: A Classification Approach for Paired Vectors

Grace Y. Yi*

University of Western Ontario

ABSTRACT

Causal inference involves determining whether a cause-effect relationship exists between two sets of interest, a task that can be framed as a binary classification problem. When dealing with a sequence of independent and identically distributed paired vectors, the kernel mean embedding of the probability distribution can be utilized to map the empirical distribution to a feature space. Subsequently, a classifier is trained in this feature space to predict causation for future vector pairs. However, this approach is susceptible to mislabeling of causal relationships, a common challenge in causation studies. In this talk, I will discuss the impact of mislabeled outputs on the training results. Moreover, I will present a learning method that takes into account the mislabeling effects and offer theoretical justifications for the validity of the proposed method.

Keywords: Binary classification, Causal inference, Label noise

In-Sample Evaluation of Subgroups Identified by Generic Machine Learning

Shuoxun Xu, **Xinzhou Guo***

Department of Mathematics, Hong Kong University of Science and Technology

ABSTRACT

When a subgroup is identified from the data, we must evaluate the post-hoc identified subgroup in a replicable way. The usual in-sample approach, which evaluates the post-hoc identified subgroup as predefined, might suffer from selection bias, and the issue can be exacerbated by generic machine learning-based subgroup identification and nonregularity; i.e. the boundary of the subgroup is non-smooth. The out-of-sample approach, which splits data into two parts—one for subgroup identification and the other for evaluation, can help address selection bias but might suffer from efficiency loss and instability issue, as the subgroup is identified using only part of the data. In this paper, we propose a conditional m-out-of-n perturbation approach to remove selection bias in in-sample subgroup evaluation and deliver valid inference on post-hoc identified subgroups when the subgroup is identified from the whole dataset of an observational study by generic machine learning. The proposed method is easy-to-compute and model-free, and remains valid regardless of whether regularity is satisfied. Through a novel theoretical framework of triple robustness linking rates of subgroups identification and nuisance estimation, we show that the proposed method, with an adaptive selection of the subsample size, achieves full efficiency across broad scenarios in generic machine learning for subgroup analysis. The merits of the proposed method are demonstrated by a re-analysis of the ACTG 175 trial.

Keywords: Asymptotically efficient; Conditional m-out-of-n perturbation; Model-free; Nonregularity; Selection bias; Triple robustness.

*The work was partially supported by a grant from Research Grants Council of the Hong Kong Special Administrative Region, China (HKUST 26308323), the Seed fund of the Big Data for Bio-Intelligence Laboratory (Z0428) and the grant L0438 from the Hong Kong University of Science and Technology

A Unified Framework of Analyzing Missing Data and Variable Selection Using Regularized Likelihood

Yuan Bian, Grace Yi, **Wenqing He***

Department of Statistical and Actuarial Sciences, University of Western Ontario

ABSTRACT

Missing data arise commonly in applications, and research on this topic has received extensive attention in the past few decades. Various inference methods have been developed under different missing data mechanisms, including missing at random and missing not at random. The assessment of a feasible missing data mechanism is, however, difficult due to the lack of validation data. The problem is further complicated by the presence of spurious variables in covariates. Focusing on missingness in the response variable, a unified modeling scheme is proposed by utilizing the parametric generalized additive model to characterize various types of missing data processes. Taking the generalized linear model to facilitate the dependence of the response on the associated covariates, the concurrent estimation and variable selection procedures are developed using regularized likelihood, and the asymptotic properties for the resultant estimators are rigorously established. The proposed methods are appealing in their flexibility and generality; they circumvent the need of assuming a particular missing data mechanism that is required by most available methods. Empirical studies demonstrate that the proposed methods result in satisfactory performance in finite sample settings. Extensions to accommodating missingness in both the response and covariates are also discussed.

Keywords: Missing data, Missing mechanism, Regularized likelihood