University of California, Irvine
Statistics Seminar

Learning from a Biased Sample

Stefan Wager
Associate Professor
Stanford University

4 p.m., Thursday, Nov. 17, 2022
6011 Donald Bren Hall

The empirical risk minimization approach to data-driven decision making assumes that we can learn a decision rule from training data drawn under the same conditions as the ones we want to deploy it under. However, in a number of settings, we may be concerned that our training sample is biased, and that some groups (characterized by either observable or unobservable attributes) may be under- or over-represented relative to the general population; and in this setting empirical risk minimization over the training set may fail to yield rules that perform well at deployment. Building on concepts from distributionally robust optimization and sensitivity analysis, we propose a method for learning a decision rule that minimizes the worst-case risk incurred under a family of test distributions whose conditional distributions of outcomes Y given covariates X differ from the conditional training distribution by at most a constant factor, and whose covariate distributions are absolutely continuous with respect to the covariate distribution of the training data. We apply a result of Rockafellar and Uryasev to show that this problem is equivalent to an augmented convex risk minimization problem. We give statistical guarantees for learning a robust model using the method of sieves and propose a deep learning algorithm whose loss function captures our robustness target. We empirically validate our proposed method in simulations and a case study with the MIMIC-III dataset.

Roshni Sahoo, Lihua Lei, Stefan Wager