

Data Science Seminar

Hosted by the Department of Mathematics and Statistics

- Date: Tuesday, October 3, 2023
- Time: 12:00pm - 1:00pm
- Room: Whitney Hall 100E
- Speaker: Dr. HaiYing Wang (University of Connecticut)
- Title: Rare Events Data and Maximum Sampled Conditional Likelihood.

Abstract

In this talk, we show that the available information about unknown parameters in rare events data is only tied to the relatively small number of cases, which justifies the usage of negative sampling. However, if the negative instances are subsampled to the same level of the positive cases, there is information loss. We derive an optimal sampling probability for the inverse probability weighted (IPW) estimator to minimize the information loss. We further propose a likelihood-based estimator to further improve the estimation efficiency, and show that the improved estimator has the smallest asymptotic variance among a large class of estimators. It is also more robust to pilot misspecification. The likelihood-based estimator is also generalized to a class of models beyond binary response models. We validate our approach on simulated data, the MNIST data, and a real click-through rate dataset with more than 0.3 trillion instances.

Biography of the speaker: Dr. Wang is an Associate Professor in the Department of Statistics at the University of Connecticut. He obtained his Ph.D. from the Department of Statistics at the University of Missouri in 2013, and his M.S. from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences in 2006. His research interests include informative subdata selection for big data, model selection, model averaging, measurement error models, and semi-parametric regression.

From:

<http://www2.math.binghamton.edu/> - **Department of Mathematics and Statistics, Binghamton University**

Permanent link:

<http://www2.math.binghamton.edu/p/seminars/datasci/100323>

Last update: **2023/09/20 23:54**

