THE UNIVERSITY OF CHICAGO

Department of Statistics

DISSERTATION PROPOSAL

YONGHOON LEE

Department of Statistics
The University of Chicago

Binary Classification with Corrupted Labels and Distribution-Free Inference with Atoms

THURSDAY, November 19, 2020, at 12:30 PM
ZOOM Meeting

ABSTRACT

We discuss two problems: binary classification with corrupted labels and distribution-free inference with atoms. Various approaches have been proposed for the problem of classification under the presence of label noise, including the adjustment of the loss function and finding a robust algorithm for estimating the classifier. In our work, we focus on the behavior of the corrupted estimator itself under the assumption of homogeneous noise, while keeping the method intact. We show that the corrupted estimator mimics regularization and therefore may reduce the risk for finite sample size. We conjecture that a vanishing fraction of corruption, namely (d/n)^{1/3}, provides sufficient regularization and leads to consistency.

Distribution-free prediction aims to find a confidence region that provides coverage under any distribution, with only the assumption of i.i.d sample. Recent work has shown that it is impossible to obtain a meaningful coverage for conditional predictive inference and binary regression for nonatomic distributions. We investigate whether it is possible to obtain such coverage for distributions with atoms. We categorize the problem into three cases: If the test point is likely to have an X value we've already observed, the distribution-free inference is trivial. In the second case where there are no repeated points in the data set with high probability, we prove that the inference is as hard as the nonatomic case. Finally, for the 'in between' setting where with high probability the test point was never seen before, but there are some repeats in

---

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@statistics.uchicago.edu. If you wish to subscribe to our email list, please visit the following web site: https://lists.uchicago.edu/web/subscribe/statseminars.

the training data set, we show that the distribution-free inference is, surprisingly, possible and propose an algorithm that provides the distribution-free coverage for binary regression with a meaningful power.

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@statistics.uchicago.edu. If you wish to subscribe to our email list, please visit the following web site: https://lists.uchicago.edu/web/subscribe/statseminars.