Seminar: 9/16/24

Yanyuan Ma
Penn State University

Title: Doubly Flexible Estimation under Label Shift

Abstract: In studies ranging from clinical medicine to policy research, complete data are usually available from a population P, but the quantity of interest is often sought for a related but different population Q which only has partial data. We consider the setting when both outcome Y and covariate X are available from P but only X is available from Q, under the label shift assumption; i.e., the conditional distribution of X given Y is the same in the two populations. To estimate the parameter of interest in Q by leveraging information from P, three ingredients are essential: (a) the common conditional distribution of X given Y, (b) the regression model of Y given X in P, and (c) the density ratio of the outcome Y between the two populations. We propose an estimation procedure that only needs some standard nonparametric technique to approximate the conditional expectations with respect to (a), while by no means needs an estimate or model for (b) or (c); i.e., doubly flexible to the model misspecifications of both (b) and (c). This is conceptually different from the well-known doubly robust estimation in that, double robustness allows at most one model to be misspecified whereas our proposal can allow both (b) and (c) to be misspecified. This is of particular interest in label shift because estimating (c) is difficult, if not impossible, by virtue of the absence of the Y-data from Q. While estimating (b) is occasionally off-the-shelf, it may encounter issues related to the curse of dimensionality or computational challenges. We develop the large sample theory for the proposed estimator, and examine its finite-sample performance through simulation studies as well as an application to the MIMIC-III database.