THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PhD Dissertation Presentation

Xiaohan Zhu
Department of Statistics
The University of Chicago

"Overfitting and Generalizing with MDL and (PAC) Bayesian Learning in Supervised Classification"

April 16, 2025, at 11:00 AM
DSI Room 103, 5460 S University Ave.

Abstract

This thesis studies overfitting, regularization, and generalization in information-theoretic learning rules for supervised binary classification, focusing on Minimum Description Length (MDL) and PAC-Bayesian prediction. First, it provides a complete characterization of the regularization path of a modified two-part-code MDL learning rule in the agnostic setting, precisely quantifying the worst-case limiting error as a function of the regularization parameter and noise level. Second, it extends this analysis to the PAC-Bayes learning rule with continuous priors and randomized predictions, showing that these rules admit analogous regularization behavior and establishing explicit connections to empirical Bayes, profile posterior, and Bayesian prediction. In particular, the PAC-Bayes rule with $\lambda=1$ corresponds to an empirical Bayes procedure. Finally, the thesis analyzes the well-specified setting of MDL, where labels are generated by a true predictor with random label noise, and shows that in contrast to the agnostic case, $\lambda=1$ yields asymptotic consistency while other regularization regimes still exhibit distinct forms of overfitting or underfitting. More broadly, the thesis suggests a general perspective on learning rules that balance empirical loss and model complexity, and points toward characterizing worst-case limiting error for more general bi-criteria objectives.
