Distribution Regression: Computational vs. Statistical Trade-off

BHARATH SRIPERUMBUDUR*, ANANT RAJ
Department of Statistics, The Pennsylvania State University, USA
Email: bks18@psu.edu

Distribution regression is a novel paradigm of regressing vector-valued response on probability measures where the probability measures are not fully observed but only through finite number N of samples drawn from them. This paradigm has many applications in forensic science, climate science, ecological inference and natural language processing. In our work, we investigate this paradigm in a risk minimization framework involving reproducing kernel Hilbert spaces and propose a ridge regressor based on kernel mean embeddings. We investigate the computational vs. statistical tradeoff involving the training sample size ℓ and the number of samples N drawn from each probability measure and show the minimax optimality of the regressor for certain growth behavior of N with respect to ℓ, with the growth rate being dependent on the smoothness of the true regressor. In particular, we construct minimax optimal estimators such that N has different regimes of dependence on ℓ: sub-square-root, super-square-root to sub-linear and super-linear to sub-quadratic and investigate their computational requirement.