香港中文大學統計學系
Department of Statistics
THE CHINESE UNIVERSITY OF HONG KONG

SEMINAR

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG

Towards modern datasets: laying mathematical foundations to streamline machine learning

INVITED SPEAKER

Chen Cheng
Department of Statistics
Stanford University

TIME

December 10, 2024 (Tue) · 10:30 am - 11:30 am

VENUE

LSB LT2 (1/F) · Lady Shaw Building LT2 · CUHK

ABSTRACT

Datasets are central to the development of statistical learning theory, and the evolution of models. The burgeoning success of modern machine learning in sophisticated tasks crucially relies on the vast growth of massive datasets, such as ImageNet, SuperGLUE and Laion-5b. However, such evolution breaks standard statistical learning assumptions and tools. In this talk, I will present two stories tackling challenges modern datasets present, and leverage statistical theory to shed insight into how should we streamline modern machine learning.

In the first part, we study multilabeling—a curious aspect of modern human-labeled datasets that is often missing in statistical machine learning literature. We develop a stylized theoretical model to capture uncertainties in the labeling process, allowing us to understand the contrasts, limitations and possible improvements of using aggregated or non-aggregated data in a statistical learning pipeline.

In the second part, I will present novel theoretical tools that are not simply convenient from classical literature, such as random matrix theory under proportional regime. Theoretical tools for proportional regime are crucially helpful in understanding "benign-overfitting" and "memorization". This is not always the most natural setting in statistics where columns correspond to covariates and rows to samples.With the objective to move beyond the proportional asymptotics, we revisit ridge regression (ℓ_2-penalized least squares) on i.i.d. data $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$. We allow the feature vector to be infinite-dimensional ($d = \infty$), in which case it belongs to a separable Hilbert space.