THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PHD DISSERTATION PROPOSAL PRESENTATION

IRINA CRISTALI

Department of Statistics
The University of Chicago

"Two Applications of Machine Learning: Causally Estimating Peer Influence on Networks and Learning Concepts from Text and Images"

WEDNESDAY, May 17, 2023, at 2:30 PM
Jones 304, 5747 S. Ellis Avenue

ABSTRACT

Part 1 of this talk addresses the problem of using observational data to estimate peer contagion effects in networked data. The main challenge is that homophily - the tendency of connected units to share similar latent traits - acts as an unobserved confounder for peer influence. Informally, it's hard to tell whether your friends have similar outcomes because they were influenced by your treatment, or whether it's due to some common trait that caused you to be friends in the first place. Because these common causes are not usually directly observed, they cannot be simply adjusted for. We describe an approach to perform the required adjustment using node embeddings learned from the network itself. The goal is to perform this adjustment nonparametrically, without functional form assumptions on either the process that generated the network or the treatment assignment and outcome processes. The key contributions are nonparametrically formalizing the causal effect to account for homophily, and showing how embedding methods can be used to identify and estimate this effect.

Part 2 of this talk highlights challenges in understanding the "structure" of representations used in state-of-the-art machine learning systems. We focus on the influential model CLIP (Contrastive Language-Image Pretraining), a neural network which jointly learns vector representations from image and text data pairs. Despite CLIP's impressive zero-shot prediction capabilities, the structural properties of its representations, i.e. the way in which semantically meaningful aspects of the input space are mapped into representations with structural properties (e.g., algebraic), are not that well understood. The talk presents a formalization of semantically meaningful information in terms of concepts, states, and measurements, and discusses several desiderata for the structure of the learned representations.