RUTGERS
School of Arts and Sciences

Department of Statistics

DEPARTMENT OF STATISTICS

Samory Kpotufe

Department of Statistics
Columbia University

Measuring Transferability: some recent insights

March 13, 2019
3:20 – 4:20pm
Light refreshments will be served
110 Frelinghuysen Road
Hill Center, Room 552

Abstract: Training data is often not fully representative of the target population due to bias in the sampling mechanism; in such situations, we aim to 'transfer' relevant information from the training data (a.k.a. source data) to the target application.
How much information is in the source data? How much target data should we collect if any? These are all practical questions that depend crucially on 'how far' the source domain is from the target. However, it remains generally unclear how to properly measure 'distance' between source and target.
In this talk we will argue that much of the traditional notions of 'distance' (e.g. KL-divergence, extensions of TV such as D_A discrepancy, and even density-ratios) can yield an over-pessimistic picture of transferability.
In fact, much of these measures are ill-defined or too large in common situations where, intuitively, transfer should be possible (e.g. situations with structured data of differing dimensions, or situations where the target distribution puts significant mass in regions of low source mass). Instead, we show that a notion of 'relative dimension' between source and target (which we simply term the 'transfer-exponent') captures the continuum from easy to hard transfer. The transfer-exponent uncovers a rich set of situations where transfer is possible even at fast rates, helps answer questions such as the benefit of unlabeled data, and has interesting implications for related problems such as multi-task learning. Finally, the transfer-exponent yields sharp guidance as to when and how to sample target data and guarantee fast improvement over source data alone. We illustrate these new insights through various simulations on controlled data, and on the popular CIFAR-10 image dataset.

The talk is based on work with Guillaume Martinet, and ongoing work with Steve Hanneke.

Bio: I graduated (Sept 2010) from Computer Science at the University of California, San Diego, advised by Sanjoy Dasgupta. I then was a researcher at the Max Planck Institute for Intelligent Systems. At the MPI I worked in the department of Bernhard Schoelkopf, in the learning theory group of Ulrike von Luxburg. Following this, I spent a couple years as an Assistant Research Professor at the Toyota Technological Institute at Chicago. I then spent some fun 4.5 years at ORFE, Princeton University as Assistant Professor.