Advances in Latent Variable and Causal Models

Paul Kishan Rubenstein

This thesis considers three different areas of machine learning concerned with the modelling of data, extending theoretical understanding in each of them. First, the estimation of f-divergences is considered in a setting that is naturally satisfied in the context of autoencoders. By exploiting structural assumptions on the distributions of concern, the proposed estimator is shown to exhibit fast rates of concentration and bias-decay. In contrast, in much of the existing f-divergence estimation literature, fast rates are only obtainable under strong conditions that are difficult to verify in practice. Next, novel identifiability results are presented for nonlinear Independent Component Analysis (ICA) in a multi-view setting, extending the scarce literature of known identifiability results for nonlinear ICA. A result of particular note is that if one noiseless view of the sources is supplemented by a second view that is appropriately corrupted by source-level noise, the sources can be fully reconstructed from the observations up to tolerable ambiguities. This setting is applicable to areas such as neuroimaging, where multiple data modalities may be available. Finally, a framework is introduced to evaluate when two causal models are consistent with one another, meaning that a correspondence can be established between them such that reasoning about the effects of interventions in both models agree. This can be used to understand when two models of the same system at different levels of detail are consistent, and has application to the problem of causal variable definition. This work has broad implications to the causal modelling process in general, as there is often a mismatch between the level at which measurements are made and the level at which the underlying 'true' causal structure exists, yet causal inference algorithms generally seek to discover causal structure at the level of measurements.