INRIA Saclay, Équipe Parietal http://team.inria.fr/parietal bat 145, CEA Saclay

Causal inference with machine learning: application to population studies

Research theme: machine learning, life sciences
Keywords: Causality, .
Duration & salary: 4 to 6 months, 500 € monthly
Research team: Parietal (INRIA Saclay and CEA)
Adviser: Bertrand Thirion, Gael Varoquaux
Contact: bertrand.thirion@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: Modern health datasets present population characteristics with many variables and in multiple modalities. They can ground prediction and understanding of individual outcomes. On the one hand, machine learning has made it possible to leverage the rich description of each individual to characterize inter-individual differences; on the other hand, the variables have complex relationships, making it hard to tease out precisely each factor independently of others. High-dimensional data makes causal understanding more challenging. The main roadblock to proper causal inference is the presence of interaction between variables that prevent to compute precisely the impact of each variable in isolation [1]. To address this, heterogeneous treatments effects models have been devised. However, their behavior in high-dimensional settings, with both the number of features and the number of samples are large, are still poorly understood. The statistical behavior (consistency and efficiency) under non-parametric models is also unknown [2].
The objective of this internship is thus to understand the theory underlying these concepts, validate it on simulations and apply it to actual population studies.
It is important to note that this question is actually pervasive across applied statistics, whenever causal conclusions are needed [1].

Proposed work: We will consider the following model:

$$\mathbf{y} = f_1(\mathbf{X}) + f_2(\mathbf{W}) + \varepsilon,$$

where $\mathbf{y}$ is the outcome variable, $\mathbf{X}$ are high-dimensional variables, providing some background, and $\mathbf{W}$ is a possibly multi-dimensional factor, whose effect on $\mathbf{y}$ is of interest. The problem is that $\mathbf{W}$ is in general not independent from $\mathbf{X}$; a standard approach consists in capturing the effect of $\mathbf{X}$ on both $\mathbf{y}$ and $\mathbf{W}$, and then to study the residual or partial effect of $\mathbf{W} - \mathbb{E}(\mathbf{W}|\mathbf{X})$ on $\mathbf{y} - \mathbb{E}(\mathbf{y}|\mathbf{X})$. to avoid overfit, the estimators are cross-validated in a nested fashion.
We propose to study the convergence properties of weakly parametric estimators [3], in particular, study the concentration on the expected value when the number of samples goes to ∞. We will then proceed with simulations, then population studies on very large-scale data ($O(10^6)$ samples in the UKbiobank dataset), to study the effects of sociological factors in conjunction with neurological factors. Future extensions of these works include handling of noise sources and missing data. The work will be done in Python and will be made available openly upon publication.
Required skills: The successful candidate will be interested in applications of machine learning and in the understanding of human cognition. Knowledge of scientific computing in Python (Numpy, Scipy, scikit image, Pandas) is encouraged. All the work will be done in Python based on the Nilearn library http://nilearn.github.io.

[1] Susan Athey and Guido Imbens. The State of Applied Econometrics - Causality and Policy Evaluation. *ArXiv e-prints*, page arXiv:1607.00699, July 2016.

[2] Xinkun Nie and Stefan Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *ArXiv e-prints*, page arXiv:1712.04912, December 2017.

[3] Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *ArXiv e-prints*, page arXiv:1510.04342, October 2015.