Machine learning and causal inference

Univ. Bretagne Sud, IRISA / team OBELIX, Vannes

PhD position to be filled in September 2020

1 Ph.D description

Context In artificial intelligence, in many fields of application, statistical learning methods have demonstrated their high level of performance. One of the tasks often performed by this type of method consists in studying the dependence / statistical associations between variables in order to understand the relationship between these explanatory variables and a variable of interest, and to predict this variable from the observation of new individuals. The aim of this research project is to evaluate the performance of machine learning methods through the angle of causality. Indeed, if the inference approach has a formal and theoretical framework tested for many contexts, statistically significant associations are not necessarily linked to causal structures between the variables. These questions arise in various fields of application, notably when analyzing clinical research data or social science data for example. The development of new technologies generating complex, large data and where the experience plan is not controlled (observation data), provides a new framework for inference by causal approach. In addition, the characteristics of the available data (in terms of volume, representativeness, quality, format, temporality) complicate the operational development of this type of procedure for extracting and predicting information from real data. This is particularly the case for Earth observation data in remote sensing, where the causal relationships between variables are crucial for understanding the underlying environmental phenomena. A second application concerns epidemiological data, in the case of an observational study with two groups of patients receiving treatment or not. The objective in a causal analysis is to know whether the difference in the observed values of the variable of interest (death) between the "treated" and "untreated" individuals can legitimately be attributed to the intervention.

Objectives Motivated by the methodological problems which arise for the analysis of data in this type of contexts, the methodological developments of the research project aim to propose innovative machine learning approaches for problems of causal inference, in the case where the availability of data is constraint. This type of analysis is based on the possibility of being able to compare probability distributions in a consistent manner. One of the difficulties stems from the difficulty associated with the large dimensions involved (both in terms of the number of data

or the number of variables available). From a technical point of view, we will notably use the team's recognised expertise in the field of optimal transport theory to offer innovative models in a modern machine learning framework. Causality implies a deeper analysis than the simple correlations study. We will focus more specifically on counterfactual approaches, which answer the question 'what would have happened if ...?', that is to say, being able to explain an output from the learning model by exhibiting patterns. This analysis will therefore rely on the ability of AI to 'imagine' alternatives. It is precisely the role of generator models, which will be considered in this thesis. The recent theoretical and applied developments of optimal transport, and in particular regarding computational aspects, have recently opened the way to multiple applications in machine learning. In particular, our research team was a pioneer in the use of optimal transport in domain adaptation. This research field contains many similarities with causality modelling, in particular in its ability to work directly on empirical probability distributions, associated with observation data, and without the need for parametric models beforehand.

Expectations Causality in machine learning is clearly described as the next step to be explored to improve AI algorithms nowadays and therefore constitute the next methodological challenge for machine learning. Widely explored in the field of epidemiology, we wish to evaluate, transfer or develop methods to deal with causality in the case of complex data such as Earth observation data. This project fully keeps up with current issues in multidisciplinary research between IT, applied mathematics and statistics by seeking to develop machine learning methods to explore and analyze information from complex data in a timely manner, adapting to contexts of unavailability of part of the data. It fits into a context of craze for artificial intelligence (AI) globally and more locally, while constituting a flagship research axis of the laboratory and the hosting team.

2 Research environment/Location

The research project of this PhD. will be conducted in the context of the IRISA laboratory (http://www.irisa.fr/), which is a joint research unit between CNRS, INRIA and several Universities and Engineering schools in Brittany (West of France). IRISA conducts research in computer science, applied mathematics and signal and image processing. More specifically, supervisors belong to the OBELIX team (Environment Observation by Complex Imagery, https://www-obelix.irisa.fr/), which main research area is machine learning for environmental data and remote sensing. The PhD topic is at the interface of several research themes of OBELIX team and will be in collaboration with IRMAR / Rennes. It will be supervised by Nicolas Courty (Professor Computer Sciences, Univ. Bretagne-Sud, IRISA, Vannes) and co-supervised by Chloé Friguet (Associate professor Statistics, Univ. Bretagne-Sud, IRISA, Vannes) and Valérie Garès (Associate professor Statistics, INSA, IRMAR, Rennes). PhD candidate will be host in IRISA-Vannes (South of Brittany), Campus de Tohannic, during the thesis. This work will be supported by Université Bretagne-Sud and Brittany Region (ARED).

3 Candidate profile and application procedure

Applicants are expected to be graduated in computer science and/or machine learning and/or applied mathematics/statistics, and show an excellent academic profile. Beyond, good programming skills are expected. To apply, please send **a resume and a motivation letter BEFORE MAY, 25** to nicolas.courty@irisa.fr, chloe.friguet@irisa.fr and valerie.gares@insa-rennes.fr

4 References

- Judea Pearl (2009) Causal inference in statistics: An overview, tech. report in Statistics Surveys,Vol. 3, 96–146 DOI: 10.1214/09-SS057
- Marc Höfler (2005) Causal inference based on counterfactuals, BMC medical research methodology, Vol. 5-28, doi:10.1186/1471-2288-5-28
- Nicolas Courty, Rémi Flamary, Devis Tuia and Alain Rakotomamonjy (2017) Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (9): 1853–1865.
- Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard and Nicolas Courty (2019) Optimal transport for structured data with application on graphs. ICML, pp.6275–6284
- Bernhard Scholkopf (2019), Causality for Machine Learning, arXiv:1911.10500
- Adrián Pérez-Suay and Gustau Camps-Valls (2019), Causal Inference in Geoscience and Remote Sensing From Observational Data, IEEE Transactions on geoscience and remote sensing, Vol 57 (3)
- Gabriel Peyré and Marco Cuturi (2019). Computational optimal transport. Foundations and Trends in Machine Learning, 11(5-6), 355-607.
- Dumas, O., Siroux, V., Le Moual, N. and Varraso, R. (2014). Approches d'analyse causale en épidémiologie. Revue d'épidémiologie et de santé publique, Vol 62(1), 53-63.