Explainability and Interpretability of Black-box Models

Supervisor
Arthur Charpentier
Department of Mathematics, UQAM, Montréal
charpentier.arthur@uqam.ca

Co-Supervisor
Marie-Pier Côté
School of Actuarial Sciences, Université Laval, Québec
marie-pier.cote@act.ulaval.ca

Abstract
Explainability and Interpretability are crucial for predictive models because they serve as a bridge between advanced machine learning techniques and their real-world applications. While black box models like deep neural networks often exhibit remarkable predictive performance, their inner workings can remain inscrutable, making their decisions challenging to trust, explain, and debug. In high-stakes decision contexts such as healthcare, finance and insurance, or autonomous vehicles, the ability to understand these models is paramount for accountability, fairness, and regulatory compliance. Explainability not only helps uncover potential biases and errors but also enables domain experts to gain valuable insights from the models, fostering collaboration between human expertise and machine intelligence. Ultimately, it is the key to unlocking the full potential of black box models while ensuring their alignment with human values and societal needs. Interpreting black box models becomes significantly more challenging when predictive features are correlated. Correlation among features can lead to multicollinearity, in which case the unique contribution of each variable to the model's predictions is difficult to discern. This complicates the attribution of decision-making to specific features, making it nearly impossible to determine which features are driving the model's output. Consequently, when features are highly correlated, the model's behavior becomes opaque, hindering efforts to extract meaningful insights, identify causality, or diagnose potential fairness issues. In such cases, interpretability techniques must grapple with disentangling the intricate web of feature interactions, further underscoring the importance of robust methods for understanding these complex, real-world machine learning systems.

Interdisciplinary/Applied experience
For applications, public domain databases will be used as much as possible. Training in statistical methodology and actuarial science will be provided. This experience will mainly occur at Université du Québec à Montréal (home of supervisor), with visits to Université

Laval (home of co-supervisor). Prof. Arthur Charpentier is a member of OBVIA (Observatoire international sur les impacts sociétaux de l'IA et du numérique - International observatory on the societal impacts of AI and digital technology) and Prof. Marie-Pier Côté is a member of IID (Institut intelligence et données - Intelligence and Data Institute), both groups having experience and expertise on explainability and interpretability of statistical models, on legal and philosophical perspectives. Prof. Marie-Pier Côté will also provide application to real insurance cases.

Teaching/Training/Education

The postdoctoral fellow (PDF) will be given rich opportunities to acquire important teaching and mentoring skills, including a three-credit course at Université Laval, and a graduate course at UQAM (jointly with Arthur Charpentier). The PDF will be offered to mentor students at the professional Master's in Actuarial Science program (Université Laval) for their essay. Also, the PDF will be participating in a CANSSI summer training program, and helping on the organizing committee of workshops, in Montréal or Québec.

Mentoring

The PDF will have committed co-supervisors who will jointly meet twice per month and will provide additional support individually or jointly as needed. Prof. Arthur Charpentier has experience in training PDFs (former PDFs are now at ETH Zurich, Aix-Marseille School of Economics and Oxford University, and one is currently working on fairness of predictive models). The successful candidate will have many opportunities for EDI training and career preparation. Collaborations with OBVIA and IID will expose the PDF to real-world problems as well as assist them in acquiring clear communications skills and practice disseminating knowledge to non-experts. The PDF will present their research in national and international conferences. These conferences will also provide networking opportunities.

Schedule

Year 1: The PDF will attend one or two international conferences (e.g., Neurips, ICML, AIStats, AAAI, JSM) as well as the SSC annual conference. During that first year, the PDF will work on classical interpretability techniques, and challenge them in the context of correlated features, with various types of dependence.
Year 2: The PDF will present at one or two international conferences (Neurips, ICML, AIStats, AAAI, JSM) as well as the SSC annual conference.

List of qualifications

- A doctoral degree in Statistics or a closely related subject.
- Knowledge of statistical and predictive modeling, ideally with some insights about interpretability, dependence, optimal transport or functional analysis.

- Knowledge of cooperative game theory (with a strong emphasis on the applied / statistical perspective) would be appreciated.
- Knowledge of numerical methods such as convex optimization would be appreciated.
- Proficiency with either R or Python, ability to develop packages. Knowledge of parallel computing implementation would be a plus.
- Ability and willingness to teach Statistics at the undergraduate and postgraduate levels.
- A good level of written and verbal communication and presentations skills, and ideally some basic knowledge of French.