"L'explicabilité via le prisme de la décision algorithmique et des jeux" des groupes de travail TADJ et Explicon

Stefano Moretti, Patrice Perny et Anaëlle Wilczynski
Sébastien Destercke et Wassila Ouerdane

Programme

9h30-10h30 : Interprétation et contrôle en planification

- Salomé Lepers (LORIA, Université de Lorraine) : "Planification probabiliste consciente d'un observateur sous observabilité partielle"
- Yoann Poupart (LIP6, Sorbonne Université) : "lczerolens: Interpreting Chess-Playing Agents"

10h30-11h : *Pause café*

11h-12h30 : Explications et recommendation en optimisation combinatoire ou multi-critère

- Sébastien Martin (Huawei) : "L'explicabilité des boucles dans les réseaux de télécommunication à l'aide des plus courts chemins"
- Manuel Amoussou (LIP6, Sorbonne Université) : "Explication de recommandations issues d'un modèle additif : application au cas général"
- Brice Mayag (LAMSADE, Université Paris-Dauphine) : "L'algorithme de l'application Yuka est-il transparent ?"

12h30-14h : *Pause déjeuner*

14h-15h30 : Explicabilité pour et par le choix social

- Clément Contet (IRIT, Université de Toulouse Capitole) : "Explaining tournament solutions: Copeland and Borda"
- Anaëlle Wilczynski (MICS, CentraleSupélec) : "Explaining the Lack of Locally Envy-Free Allocations"
- Stefano Moretti (LAMSADE, Université Paris-Dauphine) : "Social Ranking for Feature Selection"

15h30-16h : *Pause café*

16h-17h : Explicabilité en apprentissage automatique

- David Cortés (AI-vidence) : "Interpréter entre local et global (régional) : illustration de l'apport de la prise en compte des interactions de premiers ordres (via valeurs de Shapley)"
- Alessio Ragno (LIRIS, INSA Lyon) : "Distilling Rules from GIN: Logic-based Explanations for Graph Neural Networks"

17h : Rump session

Programme détaillé

1. **Salomé Lepers : Planification probabiliste consciente d'un observateur sous observabilité partielle**
   Dans des situations de collaboration homme-robot, certaines propriétés du comportement du robot peuvent être appréciées de l'humain, voire permettre une meilleure collaboration. Divers travaux ont porté sur l'obtention automatique de tels comportements. En particulier, dans une situation où un agent est conscient d'être observé par un observateur passif, l'agent peut adapter son comportement pour transmettre des informations à cet observateur. Chakraborti et al. (2019) proposent une taxonomie des différents concepts rencontrés dans ces travaux, 1. certains cherchant à transmettre de l'information, tels que ceux portant sur (a) la lisibilité, quand le comportement permet de facilement comprendre quelle tˆache il cherche 'a accomplir, (b) l'explicabilité, quand le comportement semble conduire à accomplir une tâche plutôt qu'être aléatoire, ou (c) la prédictibilité, quand la fin de la trajectoire en cours est facile à deviner ; et 2. d'autres cherchant à cacher de l'information, par exemple l'obscurcissement, quand le comportement vise à cacher la tâche réelle de l'agent. Miura et Zilberstein (2021) proposent un cadre général qui permet de planifier le comportement d'un agent conscient de la présence d'un observateur (OAMDP pour Observer-Aware Markov Decision Process) dans un cadre stochastique mais sous observabilité complète.

   Notre objectif est d'étendre ce cadre au cas d'une observabilité seulement partielle et potentiellement bruitée (pour l'observateur). Nous proposons ainsi le modèle PO-OAMDP (pour Partially Observable Observer-Aware Markov Decision Process). On pourra présenter différentes propriétés théoriques des PO-OAMDP (Lepers, Thomas et Buffet 2024), une approche de résolution et les comportements obtenus avec différents critères tels que la lisibilité ou l'explicabilité sur différents exemples.

   Travail commun avec Vincent Thomas et Olivier Buffet

2. **Yoann Poupart: lczerolens: Interpreting Chess-Playing Agents**
   AlphaZero has demonstrated exceptional efficiency in mastering chess and other games (Silver et al., 2018). While the design of these superhuman programs is intended to gain performances, e.g., by optimising the tree search, the node evaluation or the training procedure, there is still much to uncover the inner representations and the underlying mechanisms that drive the superhuman capabilities. In this respect, Leela Chess Zero (Pascutto, Gian-Carlo and Linscott, Gary, 2019), an open-source reimplementation of AlphaZero (Silver et al., 2018), provides access to numerous trained deep neural networks (DNNs). This myriad of models is truly an opportunity for research in interpretability. Indeed, this variety of models enables the study of different architectures at different stages of the training process, opening a way to gain valuable insights. Previous studies have begun to analyze the inner workings of chess models. For example, (McGrath et al., 2022) demonstrated that traditional chess concepts such as "attacks" and "material advantage" are linearly encoded in the latent representations of these models. In subsequent works, these findings were extended to dynamic concepts (Schut et al., 2023), showing the transferability of certain superhuman chess strategies. Recent research has also explored the capacity of these models for multi-step reasoning (Jenner et al., 2024; Poupart, 2024). In this context, we propose the development of an interpretability library aimed at dissecting the neural network heuristics and the associated tree search algorithms used by these models. The key objectives of the library include:

   - Conducting large-scale analyses of chess models utilizing DNNs with varying architectures.
   - Reproducing classical interpretability methods and applying theme to Leela Chess Zero chess models.
   - Investigating how planning is represented and executed in such agents.

- Comparing and exploring interpretability methods within the controlled environment of a simulable game.
- Benchmarking interpretability techniques using tailored puzzles designed to test specific reasoning capabilities.

The library is publicly available at `https://github.com/Xmaster6y/lczerolens`, and we welcome any contribution through issues or pull-requests.

3. **Sébastien Martin : L'explicabilité des boucles dans les réseaux de télécommunication à l'aide des plus courts chemins**
Dans les réseaux de télécommunication, plusieurs protocoles différents co- existent et sont interconnectés. Par exemple, dans les protocoles IGP (Interior Gateway Protocol) comme ISIS et OSPF, qui sont basés sur la notion de plus court chemin où les poids sont partagés par les routeurs afin que chaque routeur puisse calculer l'arbre des plus courts chemins et définir le prochain saut pour chaque destination. Les protocoles ISIS et OSPF assurent qu'il n'y a pas de boucle dans les réseaux quand ils les gèrent de bout en bout. Quand plusieurs instances, avec des caractéristiques différentes, doivent s'interconnecter, il est nécessaire de configurer les interconnexions. Dans ce cas, il est nécessaire de configurer les routeurs de bordure appartenant à plusieurs instances. La majorité des boucles dans ces grands réseaux est souvent due à une mauvaise configuration manuelle des routeurs de bordure. Aujourd'hui, la détection de boucles dans ces réseaux est faite à l'aide de simulateurs comme Batfish [Pedrosa et al. 2015], où les paquets sont envoyés un par un dans un réseau dont les configurations sont dupliquées. Nous proposons une méthode algorithmique plus rapide et permettant d'expliquer l'origine de la boucle.

   Plus formellement, le problème peut se voir comme un graphe $G = (V, A)$ où les sommets de V représentent les routeurs et les liens sont représentés par les arcs de A. Considérons $\mathcal{A}$ une partition des liens. Chaque ensemble de la partition représente une instance du réseau, c-à-d, une partie du réseau gérée par un protocole. Dans chaque instance, les paquets suivent le plus court chemin. Ce qui n'est pas forcément le cas quand un paquet doit traverser plusieurs instances ; cela revient à une concaténation de plus courts chemins locaux à chaque instance. Cela est dû à la configuration des routeurs de bordure. Les routeurs de bordure sont ceux ayant des liens incidents dans au moins deux ensembles de la partition $\mathcal{A}$. Nous considérons deux causes de boucles dans les réseaux IGP. Pour chaque cause, nous proposons une méthode basée sur le calcul des plus courts chemins permettant d'analyser les causes des boucles et de proposer un ensemble de solutions possibles. Certaines configurations permettent au routeur de bordure de préférer un chemin même si celui-ci est plus cher.

   Nous présenterons notre solution permettant de détecter, localiser et expliquer les boucles dans les réseaux. Nous expliquerons comment étendre ces travaux afin de gérer les réseaux gérés à l'aide de protocoles de type BGP.

   Travail commun avec Youcef Magnouche et Jérémie Leguay

4. **Manuel Amoussou : Explication de recommandations issues d'un modèle additif: application au cas général.**
Le cadre d'étude est celui de l'Aide MultiCritère à la Décision (AMCD) où un analyste fournit à un décideur une explication de la recommandation qu'il vient de lui fournir. La recommandation se fonde le plus souvent sur les préférences préalablement collectées auprès du décideur. La littérature scientifique est riche de techniques permettant cette collecte. Ces techniques se distinguent généralement les unes des autres par le modèle mathématique servant à l'élicitation des préférences. Dans ce travail, nous nous concentrons sur le modèle additif où l'utilité globale d'une alternative est obtenue en sommant les utilités marginales associées à chaque niveau d'évaluation réalisé par l'alternative sur chacun des critères considérés: dans une comparaison par paire, l'alternative préférée est celle ayant l'utilité globale la plus élevée. Notre proposition d'explication s'applique à la comparaison par paire d'alternatives; ceci rend notre approche

uniquement adaptée aux problèmes de choix et de rangement (où les alternatives sont évaluées les unes par rapport aux autres). L'une de ses spécificités, c'est que l'explication produite est compatible avec le modèle de préférences utilisé. À ce titre, elle est une preuve de la déduction de la préférence exprimée à travers la comparaison par paire à expliquer. Pour être intelligible, l'explication décompose cette dernière en sous-comparaisons explicatives atomiques dont la combinaison induit nécessairement celle-ci.

Ce travail présente une approche d'explication qui vise à jutifier des comparaison par paire voire un choix d'une alternative parmi un ensemble fini d'alternatives, en considérant le cas général d'alternatives décrites non plus seulement sur des critères binaires. Ce faire peut s'envisager dès lors que l'on constate que, dans une comparaison par paire d'alternatives (2 alternatives), on observe au plus 2 niveaux d'évaluation différents de chaque critère. L'explication produite dans ce cadre peut donc être qualifiée de "locale" à la comparaison par paire expliquée: (i) elle ne mobilise aucun autre niveau d'évaluation de critère que ceux observés dans la comparaison par paire à expliquer (ii) la fonction de score utilisée pour garantir à l'explication son caractère de preuve est spécifique aux niveaux d'évaluation de critères observés et calculée à partir de la fonction d'utilité globale considérée. Forts de ce constat, nous avons mené deux types d'expérimentations numériques: la première qui est exhaustive (en générant toutes les relations d'ordre linéaire additives définies sur des ensembles de 4, 5 et 6 critères) et la seconde, qui porte sur des fonctions d'utilité additives élicitées auprès de décideurs réels ou qui correspondent plus ou moins aux préférences réelles des auteurs des articles de journaux et de conférences consultés (une douzaine à peu près).

5. **Brice Mayag : L'algorithme de l'application Yuka est-il transparent ?**
   L'application mobile Yuka est devenue un outil incontournable pour les consommateurs souhaitant évaluer la qualité des produits alimentaires. Grâce à une interface graphique soignée et à partir d'un score global, elle fournit des informations et recommandations sur chaque produit alimentaire. L'algorithme de Yuka se veut, d'après ses concepteurs, transparent et explicable car ce score global découle de trois critères pondérés : la valeur nutritionnelle (60% du score), la présence d'additifs (30%) et la dimension biologique ou écologique (10%). Sous le prisme de l'Aide Multicritère à la Décision, nous analyserons le modèle mathématique utilisé pour le calcul du score Yuka. Ce décryptage mettra en lumière, en termes d'explicabilité et de transparence des résultats fournis aux consommateurs, les avantages et insuffisances de cet algorithme.

6. **Clément Contet : Explaining tournament solutions: Copeland and Borda**
   Tournaments are widely used to represent pairwise notions of dominance between candidates, alternatives, or teams. Tournament solutions associate a winning alternative to a tournament, notable examples including the Copeland and the Borda rule (on unweighted and weighted tournaments respectively). In this paper we study abductive explanations for tournament solutions defined as minimal subtournaments that support the winning alternative selected by the tournament solution. We develop polynomial algorithms to compute abductive explanations for the Copeland rule on unweighted tournaments and we examine how these results can be partially extended to the Borda rule on weighted tournaments. Finally, we also characterize and provide bounds on the size of the shortest abductive explanation showing that our approach can provide human-readable explanations supporting winning alternatives in tournaments.

   Joint work with Umberto Grandi and Jérôme Mengin

7. **Anaëlle Wilczynski : Explaining the Lack of Locally Envy-Free Allocations**
   In fair division, local envy-freeness is a desirable property which has been thoroughly studied in recent years. In this work, we study explanations which can be given to explain that no allocation of items can satisfy this criterion, in the house allocation setting where agents receive a single item. While Minimal Unsatisfiable Subsets (MUSes) are key concepts to extract explanations, they cannot be used as such: (i) they highly depend on the initial encoding of the problem; (ii)

they are flat structures which fall short of capturing the dynamics of explanations; (iii) they typically come in large number and exhibit great diversity. In this work, we provide two SAT encodings of the problem which allow us to extract MUS when instances are unsatisfiable. We build a dynamic graph structure which allows to follow step-by-step the explanation. Finally, we propose several criteria to select MUSes, some of them being based on the MUS structure, while others rely on this original graphical explanation structure. We give theoretical bounds on these metrics, showing that they can vary significantly for some instances. Experimental results on synthetic data complement these results and illustrate the impact of the encodings and the relevance of our metrics to select among the many MUSes.

Joint work with Aurélie Beynier, Jean-Guy Mailly, and Nicolas Maudet

8. **Stefano Moretti : Social Ranking for Feature Selection**
Various methods based on the Shapley value have enjoyed notable success in recent years within the field of Explainable AI (XAI), in particular as feature selection mechanisms and for providing feature attributions for explaining machine learning models. Nevertheless, recent studies have raised concerns regarding the use of the Shapley value in this framework. In this paper, we delve deeper into these limitations through the lens of the axiomatic analysis of the Shapley value and its implications in the realm of machine learning. Leveraging on specific examples of classification models, we compare the effects of axioms for the Shapley value with other axioms for ranking methods based on a coalitional framework, where features are the "players" and the worth of a coalition of features corresponds to their predictive capacity. As an alternative feature selection method we pay particular attention to the lex-cel, a social ranking solution introduced in the recent literature at the intersection between coalitional games and social choice theory. Our analysis suggests that axioms characterizing the lex-cel, under certain circumstances, are more suitable for ranking features in machine learning models, compared to axioms satisfied by the Shapley value. Furthermore, through experiments conducted on public datasets, we show that the lex-cel outperforms some commonly employed feature selection algorithms based on the Shapley value, in particular with respect to the capacity of selecting less redundant features. An approximated version of the lex-cel, showing a satisfactory compromise between scalability of the approach and selection performance, is also presented and discussed.

   Joint work with Laurent Gourvès and Satya Tamby

9. **David Cortés : Interpréter entre local et global (régional) : illustration de l'apport de la prise en compte des interactions de premiers ordres (via valeurs de Shapley)**
L'interprétation en conditions industrielles des algorithmes d'apprentissage statistique se heurte à 2 problèmes principaux : l'adoption par les utilisateurs finaux, et l'acceptation par des régulateurs. Les outils d'interprétation existant s'avèrent insuffisant car traitent isolément chaque décision, ou présentent une version trop synthétique/monolithique du problème. Une approche par sous-ensembles, et à plusieurs échelles s'avère indispensable pour interpréter à des niveaux de complexité différents en fonction des enjeux, permettre une critique experte et le cas échéant de remplacer le modèle "boîte noire" initial par des modèles spécifiques, plus adaptés à ces sous-ensembles, et plus robustes. Or, ce sont souvent les combinaisons de plusieurs variables, dans plusieurs plages de valeur qui permettent de réaliser cette découpe, puis de modéliser efficacement. Récemment, une bibliothèque permettant de calculer aisément l'importance d'interactions d'ordre > 1 rend l'implémentation d'approches d'interprétations régionales plus rapides. Nous présenterons rapidement les apports de l'approche régionale, une méthodologie (développée avec Confiance.AI) ainsi que de premières illustrations sur des jeux de données simples, et les pistes de nos futurs développements (formalisation causale notamment).

10. **Alessio Ragno : Distilling Rules from GIN: Logic-based Explanations for Graph Neural Networks**
Distilling Rules from GIN: Logic-based Explanations for Graph Neural Networks Explainable AI

has made strides in improving the interpretability of deep learning models, yet the explainability of Graph Neural Networks (GNNs) remains an open challenge due to their complex message-passing mechanisms. In this work, we extend the Transparent Explainable Logic Layer (TELL) to GNN architectures, introducing TELL-GIN, a GNN layer that allows extracting rule-based explanations from GNNs while maintaining high classification performance. Specifically, we adapt TELL to the Graph Isomorphism Network (GIN) by constraining its message-passing layers with positive weights and integrating activation-based logic rules that can directly interpret node and graph-level predictions. To address the vanishing gradient problem caused by the sigmoid activation, we propose to train the self-explainable model by means of distillation of a regularized GIN model with the Gumbel-sigmoid. Experiments on benchmark graph classification tasks demonstrate that TELL-GIN replicates the performance of the black-box model while providing faithful explanations for its predictions.

Joint work with Marc Plantevit and Celine Robardet