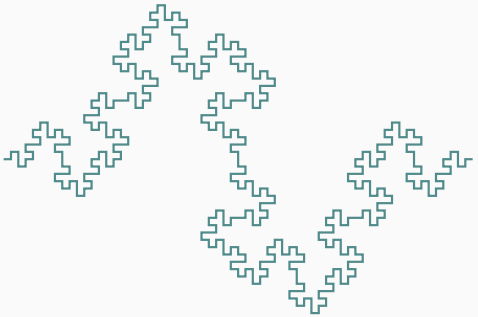
Explicabilité des algorithmes
Quelques questions éthiques

N. Maudet, LIP6, Sorbonne Université

Nov. 2021



Journée RO & Ethique – GdR RO – Univ. Paris Dauphine

Explanation, ethics, AI, OR...

Explanation is often cited as a prerequisite for an ethical use of AI

One of the (sub-)requirements for trustworthy AI (under transparency), and also the additional principle identified by Floridini and Cowls:

> *a new principle is needed in addition: explicability, understood as incorporating both the epistemological sense of intelligibility (as an answer to the question 'how does it work?') and in the ethical sense of accountability (as an answer to the question: 'who is responsible for the way it works?')*

Floridini et al. *A Unified Framework of Five Principles for AI in Society*. 2019.

High level expert group on AI. *Requirements for trustworthy AI*. 2019.

Explanation in two slides

Scope of explanations:

- Global explanation: explain how the system works in general
- Local explanation: explain a specific outcome/decision

Types of explanation:

- basic (why)
- contrastive (why ... instead of ...)
- counter-factual (what-if)

Explanation in two slides

Explanation is a social process, and has a purpose

- Model debugging/system development, auditing, user benefit, society trust

Context of use:

- high-stake decisions or not (more generally, see also the risk-based approach to regulation of the EU), autonomous systems vs. decision-aiding, time to process the explanation, stakeholders

Bhatt et al. *Explainable Machine Learning in Deployment*. ArXiv-2020.

Ecole IA2. *Gdr-IA*. https://ia2.gdria.fr/.

Ethical issues with explanations

But there are also ethical issues coming with the production of explanations, eg:

- may reveal information: sensitive data, may give rise to attacks
- may create inequalities if users have not the same ability/resources to process the explanation
- may create unjustified trust
- may hinder other aspects of the modelling process

Some principles of explanation

Pushed by lesglislation in many countries (eg. GDPR in EU), see also Four Principles of explanation (NIST report, US):

- Explanation: Systems deliver accompanying evidence or reason(s) for all outputs.
- Meaningful: Systems provide explanations that are understandable to individual users.
- Explanation Accuracy: The explanation correctly reflects the system's process for generating the output.
- Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.

Four Principles of explanation. *NIST report*. 2020.

Explaining all outputs

Which means in particular that a system must be able to explain:

- the lack of output: strong tradition in OR to explain conflicts (QuickXplain, MUS, ISS...). Crucial component when one wants to respond to contrastive questions. (By showing infeasibility of the target outcome and the current theory).
- outputs selected by tie-breaking: often those most in need of an explanation.

Juncker. *Preferred Explanations and Relaxations for Over-Constrained Problems*. AAAI-2004.

Example: kidney exchange

Kidney exchange. Patient may have willing but incompatible donor. Find cycles of compatible patient-donor pairs. Main objective: max number of patient receiving a kidney transplant.
Example of a methodology used in (Freedman et al.) in

1. elicit list of attributes deemed acceptable to used as priority
2. comparison queries on patient profile to asses weights
3. maximize number of patients receiving a kidney but use weights as a tie-breaking

Freedman et al.. *Adapting a kidney exchange algorithm to align with human values*. AIJ-2020.

Meaningful

- general principle of sparsity: can be integrated as an objective eg. learn optimal rules, scoring systems, or decision trees with sparsity constraints.
- language: highly dependent of the context of use: graphical, statistical information, natural language, logical language (which vocabulary?), etc.

Example: explaining with simple and meaningful languages

Suppose an underlying additive model, with binary criteria:

a	b	c	d	e
0.10	0.15	0.20	0.25	0.30

We want to explain why some outcome is preferred over another outcome (contrastive), eg. $(10110) \succ (01001)$.

Weights cannot be revealed / do not make sense to the end-user. Meaningful statements for explanation from end-user perspective :

All other things equal, obtaining a high quality computer is better than getting a cheap one from a supplier with a bad reputation

Suppose you have (cognitively founded) bounds on the number of criteria to use in the langage: what can be explained (and how)?

Accuracy

Many of the current approaches are heuristic in nature:

- some are mostly based on intuitions and subject to pitfalls (eg. surrogate models)
- some are backed by axiomatic properties—certainly a good direction in general (but see next slide)

Procaccia. *Axioms should explain solutions*. The Future of Economic Design, 2019.

Accuracy

Many of the current approaches are heuristic in nature:

- some are mostly based on intuitions and subject to pitfalls (eg. surrogate models)
- some are backed by axiomatic properties—certainly a good direction in general (but see next slide)

Procaccia. *Axioms should explain solutions*. The Future of Economic Design, 2019.

Note: there are also exact methods, based on MIP or logic encodings of various types of classifiers, and seeking (subset-minimal/min cardinality) prime-implicant/abductive explanations (sufficient reasons to guarantee the outcome).

Ignatiev etl et al. *Abduction-Based Explanations for Machine Learning Models*. AAAI-19.

Audemard et al.. *On the Computational Intelligibility of Boolean Classifiers*. KR-21.

Example: Shapley values for feature attribution

Basic idea: use power indices for feature attribution problem, ie. explain which features are important in the prediction $f(x_1, \ldots, x_n)$ of a (typically, ML) model.

Power index based on the evaluation of marginal contributions:

$$M_i(S) = v(S \cup \{i\}) - v(S)$$

then (weighted) averaged (Shapley: all permutations equally likely)

Claim: Use the Shapley value to explain because it is is the only method guaranteeing : Dummy, Symmetry, Efficiency, Additivity

Example: Shapley values for feature attribution

Game formulation:

Intuitively:

- Players = Features
- Payoff = Prediction of the model
- Characteristic function = payoff for all possible coalitions

Express axioms as properties of the model function f:

Eg. Dummy: for any pair of values x_i and x_i' and any values $x_{N\setminus i}$:

$$f(x_i; x_{N\setminus i}) = f(x_i'; x_{N\setminus i})$$

that is, x is never considered by the model.

Example: Shapley values for feature attribution

Problem: how do we get v from f? what does it mean for a feature to be absent? (The model was trained with all features...)

- SHAP: sample absent features cond. to contributing features from the input distribution

But then some axioms fail to hold! (eg. Dummy)

x_1	x_2	$P[X = x]$	f
0	0	0.1	0
0	1	0.0	0
1	0	0.4	1
1	1	0.5	1

baseline prediction of the model: 0.9
when computing attribution for $x_1 = 1$ and $x_2 = 1$
SHAP: attribution of $x_2 = x_1 = 0.05$!
because $v(x_2) = 1$
even though x_2 is a dummy feature

Other approaches make different choices regarding the distribution

The Many Shapley Values for Model Explanation. *Sundararajan et al.* 2020.

Knowledge limits

Unpredicted (or not?) evolution of the system, may be used in a completely different context. Absence of the “moral patient”...

Is it fine if the explanation feature is used for another purpose?

Eg. use explanation techniques to design persuasion technologies

And do properties still hold in the new context?

Example: Shapley values for feature selection

Basic idea:

1. compute feature importance
2. pick top-k as the selected feature

But consider the example:
$C(ABC) = 10, C(AB) = C(AC) = 10, C(BC) = 7$
$C(B) = C(C) = 7, C(A) = 0$

Shapley values: $(A : 2, B : 4, C : 4)$.
Any optimal model of size 2 would have to include A.

❝ *over-reliance on axiomatic "guarantees" (e.g., of "fairness") when appropriating Shapley based feature attribution methods for feature selection*

Fryer et al. *Shapley values for feature selection: The good, the bad, and the axioms*. arXiV, 2020.

Final remarks (1)

Are there trade-off between explainability and other objectives, in particular accuracy of prediction in ML (often assumed)

Final remarks (1)

Are there trade-off between explainability and other objectives, in particular accuracy of prediction in ML (often assumed)

Note sure. This one the main point made by C. Rudin:

> *It is a myth that there is necessarily a trade-off between accuracy and interpretability*

Famously showed that the COMPAS model (predictive justice in the US, black box, 100+ features) reproduced with only three rules, obtained via COREL (Certifiably Optimal Rule List), involving only two features.

Rudin. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Mach. Intell, 2019.

Final remarks (2)

Most approaches of explainability remain static, and do not allow actual contestability of the results.

However, explanation is inherently a dialectical process.

Without proper means of contesting and challenging the outcome it faces inherent limitations. Still, formal approaches exist to address these issues (argumentation theory, etc.)

Klutzz et al. *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*. 2020.