X-AI

Explainable Artificial Intelligence

Florence d’Alché-Buc
AI today

- significant advances in Statistical Machine Learning vs Symbolic Machine Learning
- spectacular results of Deep Neural Networks
- data-driven AI embedded in decision-making processes

Explainability in AI

« to describe the purpose, rationale and decision-making process of the AI tool in a way that can be understood by the average person »

Data scientist

Expert of the field (finance for instance)

User/customer

Regulator / lawyer /…
The lack of explainability in data-driven AI

Linked to the nature of statistical machine learning algorithms

Learning is a complex optimization process that takes a training dataset and produces a predictive model.

Visualization of a loss function, Li et al. NeurIPS 2018.

Bias in your data?
Local Minima?
Model selection?

(© Matthieu Ferrand)
The lack of explainability in data-driven AI

Linked to the objectives of machine learning algorithms

A learning algorithm attempts to define a predictive model by searching for input patterns correlated with the output variable based on a strong assumption about data: the i.i.d. assumption

Animal in the snow?

Wolf?

(© Matthieu Ferrand)

Guestrin et al. 2016

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted the bad model</td>
<td>10 out of 27</td>
</tr>
<tr>
<td>Snow as a potential feature</td>
<td>12 out of 27</td>
</tr>
</tbody>
</table>

Table 2: “Husky vs Wolf” experiment results.
The lack of explainability in data-driven AI

Due to the nature of the predictive models

- Some models are more explainable than others: sparse linear models, decision trees, probabilistic graphical models, random forests, …

- deep neural networks exhibit a very high level of complexity (millions of parameters)

Pb : performance is often associated to very complex models & ability to tackle massive training datasets
The need for Explainability

- Human-readable justification of a decision
- Compliance to legislation “Right to explanation”
- Identification of systems flaws
- Information extraction
- Explain to build trust
- Explain to control
- Explain to improve
- Explain to discover
XAI, a compound of trustworthy AI

[Diagram showing interrelated concepts such as Human agency and Oversight, Technical robustness and Safety, Accountability, Societal and Environmental wellbeing, Privacy and Data Governance, Diversity, Non-Discrimination and Fairness, To be continuously evaluated and addressed throughout the AI system’s life cycle, and Transparency.]
Explainability in data-driven AI

Focus on local explainability: provide an “explanation” of the predictive model’s decision

What is an « explanation » ? For whom ? a data scientist, an expert of the field a user, the regulator ?

Main factors that led to that prediction

High level concepts that are activated when the prediction is given

Counterfactual reasoning: if I change this feature value, does the prediction change ?
Explanations also depend on the nature of data

Multivariate hand-defined features

Image / Audio / Video

Natural Language processing
Post-hoc Approaches: local linear proxy

- **LIME (Ribeiro et al. 2016)**

- Model-agnostic approach that builds a sparse linear **proxy model** to get insights on a local decision once the whole model is learned.

- **Perturbation-based approach**
Post-hoc Approaches: saliency maps

\[ f \] is the predictive model, \( x \) is the input:

\[ \frac{df}{dx} \]


Fig. Adebayo et al. NeurIPS 2018.
Post-hoc approaches: Tree explainer
(Lundberg et al. 2019)

(A) "Black box" model prediction

TreeExplainer

"White box" local explanation

(B) Combining local explanations from many samples...

...can lead to global model insights
Explainability By Design

- Identify a (specific) neural network to a set of logical rules (hybrid networks)

- Modify the architecture of a network to make it interpretable (Self-explainable Networks, Alvarez-Melis & Jaakola 2018)

(SENN)
Explainability by design

  - (logical) consistency: non contradictory rules
  - Completeness
  - Fidelity of the « explanations » to the model’s output
  - Sparsity of high level concepts
  - Stability of explanations

- Learn jointly two models: one for prediction, the other for explanation (Hendricks et al. 2016, Dong et al. 2017, Parekh et al. 2020)
Explainability by design


This is a pine grosbeak because this bird has a red head and breast with a gray wing and white wing.

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

This is a cardinal because ...

Deep Finegrained Classifier

Recurrent explanation generator model

- VI
- LSTM
- EOS
Tools on the shelves: mainly post-hoc approaches – good for existing black box models currently in production, but with some flaws: provide an explanation but may be not the one « used » by the model

Formal work on interpretations/explanations in ML, re-think machine learning/AI at the lense of explainability for a next generation AI tools

« Can biologist fix a radio ? » (Lazebnik, 2002) the celebrated paper in quantitative biology in 2000’s applies somehow here. Can a statistician provides an explanation ?

- Explanations are currently more interpretations than explanations: what link with reasoning ? What link with logics ? What link with knowledge ?

- Making a predictive model explainable belongs more to symbolic AI and calls for automated reasoning, knowledge representation etc… a lot to borrow from years of AI.

Other ways of thinking: counterfactual reasoning, intervention, Bayesian approaches, probabilistic programming, knowledge graph and automated reasoning
FLINT: a framework for learning interpretable network

Usage 1:
Joint learning of f and g,
Mutual benefits,
g can even the final predictor

Usage 2:
Post-hoc/reverse engineering of a pre-defined network f

Parekh et al. 2020.
References

- Christoph Molnar’s online book on interpretable machine learning (link)
- Adebayo et al. Sanity Checks for Saliency Maps, NeurIPS 2018
- Hendricks, Akata, …, Darrell, Generating visual explanations, ECCV 2016 (link)
- Platform, common tools:
  - What If tools (Google), Captum (Pytorch/Facebook), 360xAI (IBM) https://aix360.mybluemix.net/, iml R package