Explainable Artificial Intelligence

Seminar Kick Off
Introduction

- Artificial Intelligence / Machine Learning in the early days:
  - Explainable by nature

Introduction

- Artificial Intelligence / Machine Learning nowadays:
  - Not very transparent

The Principle of Explicability: “Operate transparently”

Transparency is key to building and maintaining citizen’s trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable, comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems.

Explicability is a precondition for achieving informed consent from individuals interacting with AI systems and in order to ensure that the principle of explicability and non-maleficence are achieved the requirement of informed consent should be sought. Explicability also requires accountability measures be put in place. Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organisations and developers of an AI system, the technology implementers, or another party in the supply chain.
Why Bother?

10. Transparency

Transparency concerns the reduction of information asymmetry. Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system. Being explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required from all models that use human data or affect human beings or can have other morally significant impact.
Introduction

• Who is the target for Explainable AI?
  – End users
  – Developers
  – …

Introduction

• Intrinsic vs. post-hoc
  – e.g., decision trees are intrinsically interpretable
  – post-hoc: take a trained model (e.g., neural net) and create explanations

• Global vs. local
  – global: explain the entire model at once
  – local: explain a decision made by the model

• Model-specific vs. model-agnostic
  – some approaches work for a particular model type (e.g., CNN)
  – others are applicable to any ML model

Organization

• Requirements
  – Familiarize yourself with a particular approach to XAI
  – Present it in the seminar
  – Write a seminar paper
  – Review others’ seminar papers
    • it is a good idea to also read the main papers for the topics you review

• First step
  – Pick topics of interest
  – Send a ranked list to Ms. Bianca Lermer by the end of the week
Organization

• We will use a process called “peer review”
  – widely used (and discussed) in science
  – you will review your fellow students’ seminar papers

• Timeline
  – Prepare a draft until March 30th
  – You will get two seminar papers to review
  – Submit your reviews until April 14th

• Seminar (i.e., presentations, discussions)
  – April 28th, May 5th, 12th, 19th
  – We are planning an on campus seminar

• Final seminar paper submission: June 18th
Preparing Your Seminar Paper

• Conduct initial literature research
  – get your hands on some core papers on the topic
• Read follow up papers about the topic
  – which follow-up approaches/refinements have been developed?
• Find papers citing the original paper
  – who says what about it?
  – who uses it, and for which purposes?
• Anything else you want to discuss
Preparing a Review

• 1st rule: be constructive!

• What you should point at
  – can you follow easily? is there information missing at any point?
  – are all claims well supported?
  – do you have any questions not answered?
  – aspects underrepresented

• What you should not do
  – provide general criticism (“don’t like the paper”)
  – correct every spelling mistake
  – rewrite the seminar paper
Preparing the Presentation

• You don’t have to start at zero
  – everybody should be familiar with XAI

• Focus on key aspects
  – which approach is taken in the study you present?
  – how was it evaluated?
  – what are its strengths and weaknesses?

• Be illustrative
  – use examples
  – try it out yourself, if possible

• Be entertaining
  – it’s (partly) up to you whether we are having fun here ;-)
Let’s Talk about Topics

• Note: these are suggestions
  – you are free to set your own focus
  – and formulate questions which you find interesting
### TABLE 2. Summary of explainability techniques.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>References</th>
<th>Intrinsic/Post-hoc</th>
<th>Global/Local</th>
<th>Model-specific/Model-agnostic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision trees</td>
<td>[139], [140], [141], [142], [143]</td>
<td>I</td>
<td>G</td>
<td>SP</td>
</tr>
<tr>
<td>Rule lists</td>
<td>[66], [143], [144], [145], [146]</td>
<td>I</td>
<td>G</td>
<td>SP</td>
</tr>
<tr>
<td>LIME</td>
<td>[84], [85], [102], [147]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Shapely explanations</td>
<td>[101]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Saliency map</td>
<td>[87], [88], [89], [90], [91], [96], [97]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Activation maximization</td>
<td>[82], [83]</td>
<td>H</td>
<td>G</td>
<td>AG</td>
</tr>
<tr>
<td>Surrogate models</td>
<td>[106], [107], [84]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Partial Dependence Plot (PDP)</td>
<td>[108], [51], [110]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Individual Conditional Expectation (ACE)</td>
<td>[112], [113]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Rule extraction</td>
<td>[74], [114], [115], [116], [117], [118]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Decomposition</td>
<td>[93], [94], [95]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Model distillation</td>
<td>[49], [123], [124], [125], [126], [127]</td>
<td>H</td>
<td>G</td>
<td>AG</td>
</tr>
<tr>
<td>Sensitive analysis</td>
<td>[129], [130]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Layer-wise Relevance Propagation (LRP)</td>
<td>[131]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Feature importance</td>
<td>[113], [132], [86]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Prototype and criticism</td>
<td>[133], [134], [135], [136]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Counterfactuals explanations</td>
<td>[137]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
</tbody>
</table>

†: Intrinsic, H: Post-hoc, G: Global, L: Local, SP: Model-specific, AG: Model-agnostic

Topics

- Local interpretable model-agnostic explanations (LIME)

Figure 3: Toy example to present intuition for LIME. The black-box model’s complex decision function $f$ (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using $f$, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Topics

- Feature importance

Topics

• SHAP (Shapley Values)

Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1

Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1

https://github.com/slundberg/shap
Topics

- Saliency Maps

https://www.geeksforgeeks.org/what-is-saliency-map/
Topics

• Activation Maximization

Topics

• Surrogate Models
Topics

- Rule Extraction from Neural Networks

**Rule 1:**
\[\text{Exp(CFD)} \in [1.80, 7.20] \land \text{Exp(AMACR)} > 4.04 \rightarrow \text{class 1}\]

**Rule 2:**
\[\text{Exp(NELL2)} > 4.32 \land \text{Exp(ATP8A1)} > 3.22 \rightarrow \text{class 2}\]

**Rule 3:**
\[\text{Exp(HBB)} \in [5.57, 8.88] \rightarrow \text{class 1}\]

**Rule 4:**
\[\text{Exp(SUMO3)} > 2.09 \rightarrow \text{class 2}\]

**Rule 5:**
Otherwise \(\rightarrow\) **class 1**
Topics

• Partial Dependency Plots

https://www.researchgate.net/publication/283071368_Local-scale_topoclimate_effects_on_treeline_elevations_A_country-wide_investigation_of_New_Zealand%27s_southern_beech_treelines
Topics

• Individual Conditional Expectation Plots

Figure 3: c-ICE plot for age with $x^*$ set to the minimum value of age. The right vertical axis displays changes in $f$ over the baseline as a fraction of $y$‘s observed range.

https://www.researchgate.net/publication/257028373_Peeking_Inside_the_Black_Box_Visualizing_Statistical_Learning_With_Plots_of_Individual_Conditional_Expectation
Topics

• Decomposition

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4407709
Topics

- Model Distillation

Figure 1: Auditing a loan risk scoring model by training transparent models on data labeled with the risk scores and with ground-truth outcomes for loan defaults.

Topics

- Sensitivity Analysis

Topics

- Layer-wise Relevance Propagation

Topics

• Prototypes and criticisms

![Graph showing classification error vs. number of prototypes](https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf)

Figure 1: Classification error vs. number of prototypes $m = |S|$. MMD-critic shows comparable (or improved) performance as compared to other models (left). Random subset of prototypes and criticism from the USPS dataset (right).

Topics

• Layer-wise Relevance Propagation

Topics

• Counterfactual explanations

https://pureai.com/articles/2020/03/13/open-source-counterfactuals.aspx
Now it’s up to you...

- ...pick your favorites
- Maybe you have other ideas? They are welcome!