2.8 Interpretability techniques

Practical guidance – cross-domain

Authors: Rhys Ward

One way to provide assurance is to make the ML system being used interpretable. Interpretability may help us to:

- Understand the system retrospectively: to understand, with respect to a harm-causing action or decision, what went wrong, and why
- Understand the system prospectively: to predict, mitigate, and prevent future harm-causing actions or decisions.

In some sense an algorithm is interpretable if we can understand how it works and/or why it makes the decisions that it does make. [17] defines interpretability in the context of ML as ‘the ability to explain or to present in understandable terms to a human’ but notes that what constitutes an explanation is not well-defined. In practice, the term interpretability is used to refer to a number of distinct concepts. We want to answer the question ‘to what extent does machine learning need to be interpretable to provide assurance?’ To answer this question we must decide who needs to understand the system, what they need to understand, what types of explanations are appropriate, and when do these explanations need to be provided.

Types of interpretability

[41] seeks to clarify the myriad different notions of interpretability of ML models in the literature – what interpretability means and why it is important. It is noted that interpretability is not a monolithic concept and relates to a number of distinct ideas. The distinction is often made between methods which are intrinsically transparent and post-hoc methods which attempt to explain a model. We identify the following types of interpretability. A model/system is:

- Transparent if we understand how it works (mechanistically, at some level, for some part of the process). A transparent model is one which is simple enough for humans to understand. We may have transparency at the level of the:
  - Learning algorithm
  - Learned model
  - System logic
  - Parameters or model structures (do they relate to human-understandable concepts?)
- Explainable if we understand why it makes the decisions/predictions that it does make.
  - Global explainability techniques approximate the model with a simpler more transparent one. This simple approximate model is an explanation.
Local explainability techniques map inputs to outputs and identify important inputs. Other methods locally approximate the model. These methods help us to answer the question ‘what were the important factors in this decision?’.

We can also categorise some of the features of these different types of interpretability. Are they faithful representations of the model, or approximations? Do they interpret the whole model (global) or individual decisions (local)? Transparency is an intrinsic property of a model (it is either easy to understand or not, or some degree in between), whereas explainability techniques are post-hoc methods which require some extra effort to implement. Table 1 summarises this.

<table>
<thead>
<tr>
<th>Transparency</th>
<th>Faithful/Approximate</th>
<th>Global/Local</th>
<th>Intrinsic/Post-hoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Ex.</td>
<td>F</td>
<td>G/L</td>
<td>I</td>
</tr>
<tr>
<td>Global Ex.</td>
<td>A</td>
<td>G</td>
<td>PH</td>
</tr>
</tbody>
</table>

Table 1: Features of different types of interpretability

**Interpretability techniques**

There is extensive literature surrounding different techniques to interpret or explain ML models or systems. [1] provides a thorough review of current interpretability techniques as summarised in Table 2.

<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>[4] [30] 27 3 [61]</td>
<td>I</td>
<td>G/L</td>
<td>SP</td>
</tr>
<tr>
<td>Rule lists</td>
<td>[44] [61] 74 19</td>
<td>I</td>
<td>G/L</td>
<td>SP</td>
</tr>
<tr>
<td>LIME</td>
<td>[57] [66] 56 67</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Shapely explanations</td>
<td>[42]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Saliency map</td>
<td>[29] [36] 77 53 43 15</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Surrogate models</td>
<td>[51] [64] 57</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Partial Dependency Plot (PDP)</td>
<td>[26] [7] 35</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Individual Conditional Expectation (ACI)</td>
<td>[10] 22</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Rule extraction</td>
<td>[23] [71] 2 28 55 25</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Decomposition</td>
<td>[58] 50</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
<tr>
<td>Model distillation</td>
<td>[27] [38] 75 68 34 33</td>
<td>H</td>
<td>G</td>
<td>AG</td>
</tr>
<tr>
<td>Sensitive analysis</td>
<td>[13] 12</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Layer-wise Relevance Propagation (LRP)</td>
<td>[49]</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Feature Importance</td>
<td>[22] [59] 69</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Prototype and criticism</td>
<td>[8] [60] 16 37</td>
<td>H</td>
<td>G/L</td>
<td>AG</td>
</tr>
<tr>
<td>Counterfactual explanations</td>
<td>[45]</td>
<td>H</td>
<td>L</td>
<td>AG</td>
</tr>
</tbody>
</table>

Table 2: Summary of interpretability techniques
Table 2 differentiates between local and global explainability i.e. the interpretability of a single decision vs the interpretability of the whole logic of a model. In [41] they also differentiate between intrinsic explainability (e.g. transparency) - simple models which are inherently easy to understand, and post-hoc explainability - methods that analyse the model after training. Post-hoc techniques refer to the global and local techniques described earlier.

Note that each intrinsic technique is also global. This is because this survey considers models which are intrinsically transparent to be transparent globally. In this sense, models cannot be transparent (simple enough for us to understand) for some decisions but not others. The distinction is also made between techniques which are model specific vs model agnostic. A number of other surveys have also been conducted:

- [5] surveys ML methods as they relate to assurance at each stage of the ML life-cycle (Data Management, Model Learning, Model Verification, and Model Deployment).
- [9] surveys interpretable models differentiating between intrinsically explainable and justifiable models/decisions (i.e. transparency vs local explainability).
- [46] surveys different interpretability techniques and compares them on their effectiveness to different user-groups. We will discuss how to evaluate explanations in section 3.
- [18] summarises recent developments in explainable supervised learning.
- [78] reviews recent studies in understanding neural-network representations and learning neural networks with interpretable/disentangled middle-layer representations.
- [48] Seeks to investigate “What makes for a good explanation?” with reference to AI systems and takes a psychological approach. It discusses some explanation methods (e.g. visualisation, text based).
- [24] focuses on explainable methods in deep neural architectures, and briefly highlights review papers from other subfields.
- [47] describes some model-agnostic interpretability methods, their pros and cons, and how/when to implement them.

Comparing the interpretability of different machine learning models

Table 3 summarises some of the multitude of interpretability techniques in the literature. We classify these by the type of interpretability which they capture (see section 1.1) and by the type of model which they can be used to interpret. Some techniques can be used on multiple models and are referred to as ‘Model Agnostic’. Some methods provide some transparency to the system logic without necessarily interpreting the ML model(s) being used (e.g. [20]).

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Interpretability Type</th>
<th>Transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local Ex.</td>
<td>Global Ex.</td>
</tr>
<tr>
<td>System Level</td>
<td>[70]</td>
<td></td>
</tr>
<tr>
<td>Model Agnostic</td>
<td>[39] [42] [73]</td>
<td>[32] [40]</td>
</tr>
<tr>
<td>Supervised Learning</td>
<td>[6] [18] [41]</td>
<td>[18] [41]</td>
</tr>
<tr>
<td>Unsupervised Learning</td>
<td>[42] [57]</td>
<td></td>
</tr>
<tr>
<td>Reinforcement Learning (RL)</td>
<td>[31]</td>
<td></td>
</tr>
<tr>
<td>Classifiers</td>
<td>[31] [54] [57]</td>
<td></td>
</tr>
<tr>
<td>Neural Networks</td>
<td>[6] [24] [78] [65] [76]</td>
<td>[24] [52] [78] [63]</td>
</tr>
</tbody>
</table>
Table 3: Interpretability techniques for different ML methods

Some of these methods offer very technical “explanations” which would not be suitable for most stakeholders (e.g. doctors, lay-users) and different stakeholders require different types of explanation [46]. When explaining algorithmic decisions the format of the explanation is key.

Summary of approach

1. Define the extent to which the ML system needs to be interpretable and define a set of interpretability requirements (e.g. 'Local decisions can be explained to identify the cause of accidents after they occur’) – see guidance on interpretability requirements.
2. Define the types of interpretability needed to meet requirements (e.g. 'local explainability methods implemented to map inputs to outputs and identify important inputs') and consider model selection trade-offs.
3. Implement suitable interpretability techniques (this may result in choosing a transparent model).
4. Ensure explanations are provided in a suitable format for, and are made available to, the audience.
5. Evaluate explanations (see guidance on interpretability evaluation):
   a. Are they suitable for the audience?
   b. Are they faithful to the system process?

References


