The Mythos of Model Interpretablility

Zachary C. Lipton

https://arxiv.org/abs/1606.03490
Outline

• What is *interpretability*?

• What are its desiderata?

• What model properties confer interpretability?

• Caveats, pitfalls, and takeaways
What is Interpretability?

- Many papers make axiomatic claims. 
  *This model is {interpretable, explainable, intelligible, transparent, understandable}*

- But what is interpretability? & why is it desirable?

- Does it hold consistent meaning across papers?
We want *good* models
We also want \textit{interpretable} models
The Human *Wants* Something the Metric Doesn’t
So What’s Up?

It seems either:

• Metric captures everything and people are crazy
• The metric mismatched from real objectives

We hope to refine the discourse on interpretability

In dialogue with the literature, we create a taxonomy of both objectives & methods
Outline

• What is interpretability?

• What are its desiderata?

• What model properties confer interpretability?

• Caveats, pitfalls, and takeaways
Trust

- Does the model know when it’s uncertain?
- Does the model make same mistakes as humans?
- Are we comfortable with the model?
Causality

• Tell us something about the natural world

• Predictions vs actions

• Caruana (2015) shows a mortality predictor (for use in triage) that assigns lower risk to asthma patients
Transferability

- Training setups differ from the wild
- Reality may be non-stationary, noisy
- Don’t want model to depend on weak setup
Informativeness

• We may train a model to make a *decision*

• But its real purpose is to be a feature

• Thus an interpretation may simply be valuable for the extra bits it carries
Outline

• What is interpretability?

• What are its desiderata?

• What model properties confer interpretability?

• Caveats, pitfalls, and takeaways
Transparency

• Proposed solutions conferring interpretability tend to fall into two categories

• **Transparency** addresses understanding how the model works

• **Explainability** concerns the model’s ability to offer some (potentially post-hoc) explanation
Simulatability

- One notion of transparency is simplicity
- Small decision trees, sparse linear models, rules
- A model is simulatable if a person can *run* it
Decomposability

- A relaxed notion requires understanding individual components of a model
- Such as: weights of a linear model or the nodes of a decision tree
Transparent Algorithms

• We understand the behavior algorithm (but maybe not output)

• E.g. convergence of convex optimizations, generalization bounds
Post-Hoc Interpretability

Ah yes, something cool is happening in node 750,345,167… maybe it sees a cat?

Try jiggling the inputs?
Verbal Explanations

• Just as people generate explanations (absent transparency), we might train a (possibly separate) model to generate explanations.

• Could think of captions as interpretations of classification model.

(Image: Karpathy et al 2015)
Saliency Maps

- Mapping b/w input & output might be impossible to describe succinctly, local explanations are potentially useful.

(Image: Wang et al 2016)
Case-Based Explanations

- Retrieve labeled items that look similar to the model
- Doctors employ this technique to explain treatments

(Image: Mikolov et al 2014)
Outline

• What is interpretability?

• What are its desiderata?

• What model properties confer interpretability?

• Caveats, pitfalls, and takeaways
Discussion Points

• Linear models not strictly more interpretable than deep learning
• Claims about interpretability must be qualified
• Transparency may be at odds with the goals of AI
• Post-hoc interpretations may potentially mislead
Thanks!

Acknowledgments:
Zachary C. Lipton was supported by the Division of Biomedical Informatics at UCSD, via training grant (T15LM011271) from the NIH/NLM. Thanks to Charles Elkan, Julian McAuley, David Kale, Maggie Makar, Been Kim, Lihong Li, Rich Caruana, Daniel Fried, Jack Berkowitz, & Sepp Hochreiter

References:
The Mythos of Model Interpretability (ICML Workshop on Human Interpretability 2016) - ZC Lipton
http://arxiv.org/abs/1511.03677