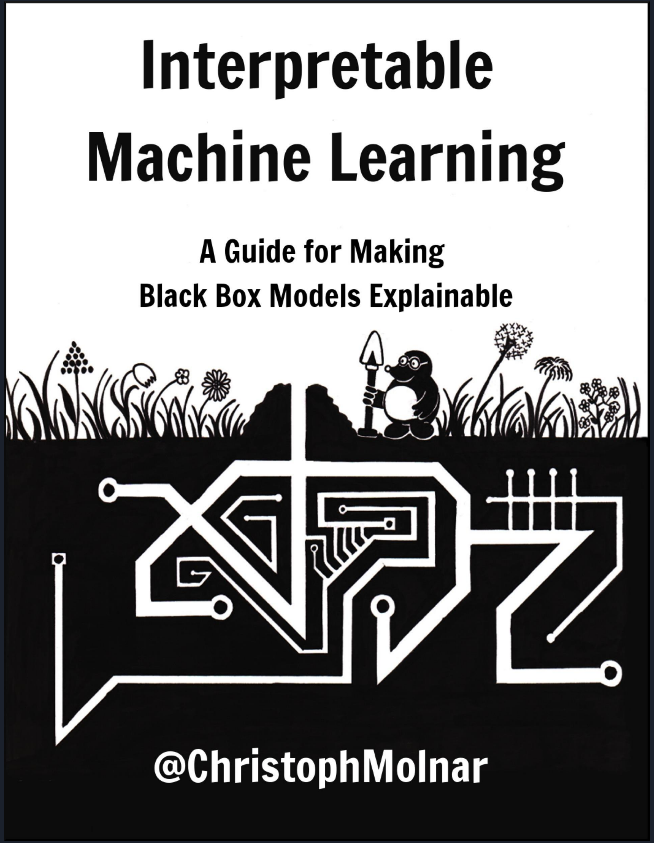
Interpretable Machine Learning

Chapter 2 Review

Benjamin Etheredge
Benjamin.Etheredge+IML@gmail.com
HSV-AI Meetup
04/15/2020



Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

@ChristophMolnar

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Terminology

*"An **Algorithm** is a set of rules that a machine follows to achieve a particular goal"*

*"**Machine Learning** is a set of methods that allow computers to learn from data to make and improve predictions"*

*"A **Learner** or **Machine Learning Algorithm** is the program used to learn a machine learning model from data."*

*"A **Machine Learning Model** is the learned program that maps inputs to predictions."*

*"A **Black Box Model** is a system that does not reveal its internal mechanisms."*

*"The opposite of a black box is sometimes referred to as **White Box**, and is referred to in this book as interpretable model."*

Interpretable Machine Learning: A Guide for Making Black Box Models Explained

Terminology

*“**Interpretable Machine Learning** refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans.”*

*“A **Dataset** is a table with the data from which the machine learns.”*

*“An **Instance*** [or **Record**] *is a row in the dataset.”*

*“The **Features** are the inputs used for prediction or classification.”*

*“The **Target*** [or **Label**] *is the information the machine learns to predict.”*

*“A **Machine Learning Task** is the combination of a dataset with features and a target.”*

*“The **Prediction** is what the machine learning model “guesses” what the target value should be based on the given features.”*

Interpretable Machine Learning: A Guide for Making Black Box Models Explained

Interpretability

"Interpretability is the degree to which a human can understand the cause of a decision." (Miller)

"Interpretability is the degree to which a human can consistently predict the model's result." (Kim)

Photo by bruce mars on Unsplash

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

Picture property of Mister Mittens

*“If a machine learning model performs well, **why do not we just trust the model** and ignore why it made a certain decision?”*

“The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez)”

Goodhart’s Law: *When a measure becomes a target, it ceases to be a good measure.*

Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).

Trade-offs

Do you just want to know what is predicted?

Or do you want to know why the prediction was made and possibly pay for the interpretability with a drop in predictive performance?

In some cases, you do not care why a decision was made, it is enough to know that the predictive performance on a test dataset was good.

But in other cases, knowing the 'why' can help you learn more about the problem, the data and the reason why a model might fail.

The need for interpretability arises from an incompleteness in problem formalization (Doshi- Velez 2017).

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).

Quest for Knowledge

“Closely related to learning is the human desire to find meaning in the world. We want to harmonize contradictions or inconsistencies between elements of our knowledge structures.”

“When opaque machine learning models are used in research, scientific findings remain completely hidden if the model only gives predictions without explanations.”

When do you want interpretability?

The goal of science is to gain knowledge, but many problems are solved with big datasets and black box machine learning models. The model itself becomes the source of knowledge instead of the data. Interpretability makes it possible to extract this additional knowledge captured by the model.

The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior.

Machine learning models take on real-world tasks that require safety measures and testing.

When do you want interpretability?

BIAS!!!

Social Acceptance

The process of integrating machines and algorithms into our daily lives requires interpretability to increase social acceptance.

A machine or algorithm that explains its predictions will find more acceptance.

Social Acceptance

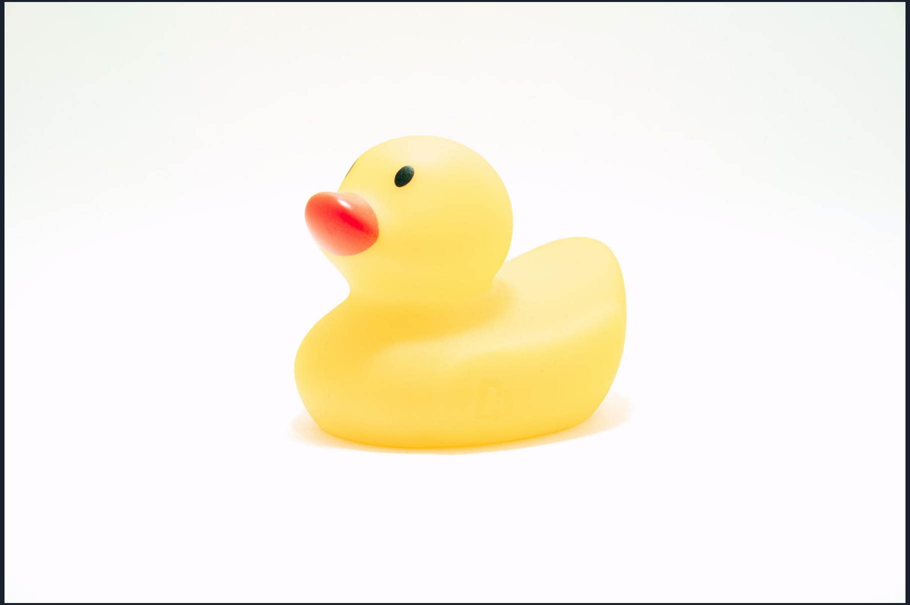
Explanations are used to manage social interactions.

By creating a shared meaning of something, the explainer influences the actions, emotions and beliefs of the recipient of the explanation.

For a machine to interact with us, it may need to shape our emotions and beliefs. Machines have to “persuade” us, so that they can achieve their intended goal.

Debugging

Machine learning models can only be debugged and audited when they can be interpreted.



When is interpretability **not** needed?

Interpretability is not required if the model has no significant impact.

Interpretability is not required when the problem is well studied.

System manipulation

Interpretability might enable people or programs to manipulate the system.

This mismatch between the goals introduces incentives for applicants to game the system to increase their chances of getting a loan.

The system can only be gamed if the inputs are proxies for a causal feature, but do not actually cause the outcome. Whenever possible, proxy features should be avoided as they make models gameable.

Desirable Traits to Look For

Fairness - Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.

Privacy - Ensuring that sensitive information in the data is protected.

Reliability or **Robustness** - Ensuring that small changes in the input do not lead to large changes in the prediction.

Causality - Check that only causal relationships are picked up.

Trust - It is easier for humans to trust a system that explains its decisions compared to a black box.

END.

Interpretable Machine Learning

Chapter 2 Review

Benjamin Etheredge
Benjamin.Etheredge+IML@gmail.com
HSV-AI Meetup
04/15/2020

BACKUPS

Model-agnostic methods for interpretability treat machine learning models as black boxes, even if they are not.