Why you shouldn't trust me: A survey on Adversarial Model Interpretation Manipulations

Verena Heusser

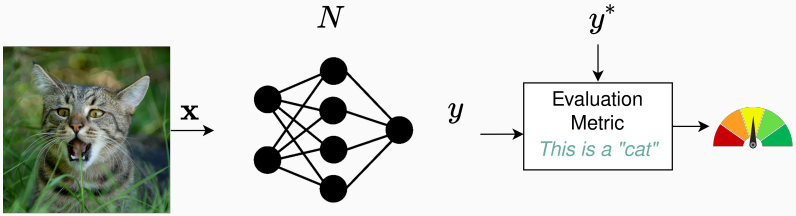
02 Februar 2021

KIT, Intelligent System Security Research Group, Seminar Explainable Machine Learning

Motivation

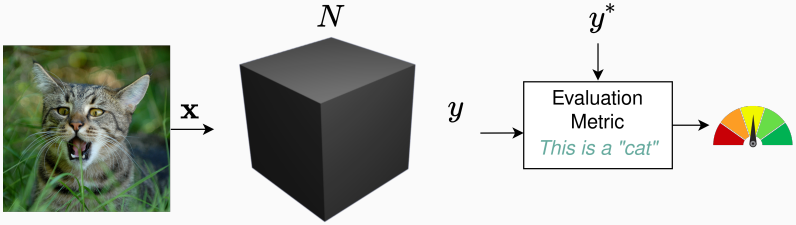
Omnipresent Machine Learning

- Machine learning algorithms are moving our of the lab into the real world
- Performance comes at the cost of complexity



Omnipresent Machine Learning

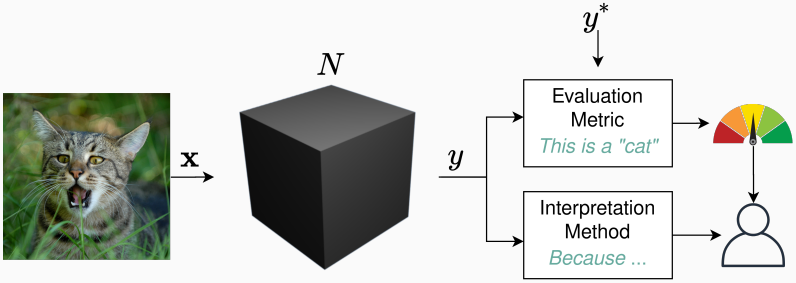
- Machine learning algorithms are moving our of the lab into the real world
- Performance comes at the cost of complexity



Omnipresent Machine Learning

- Machine learning algorithms are moving our of the lab into the real world
- Performance comes at the cost of complexity

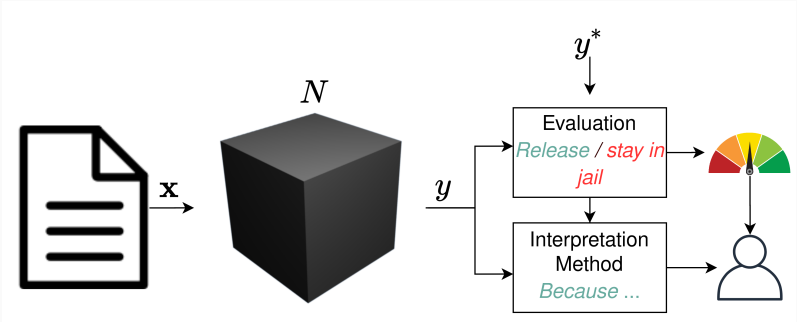
- so far: *what* is the most likely label
- now also: *why* does the model choose this label and which features were important for the decision → Explainable ML



Omnipresent Machine Learning

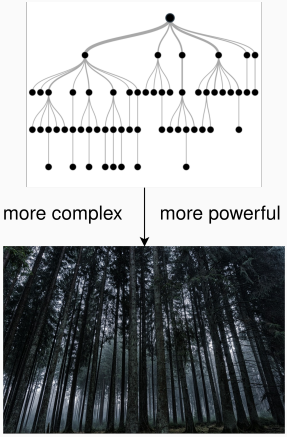
- Machine learning algorithms are moving our of the lab into the real world
- Performance comes at the cost of complexity

- so far: *what* is the most likely label
- now also: *why* does the model choose this label and which features were important for the decision → Explainable ML



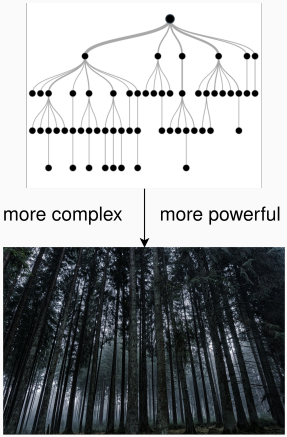
Explainable Machine Learning

- Goal: Make stakeholders and customers comfortable
- Problem: Tradeoff between complexity and interpretability



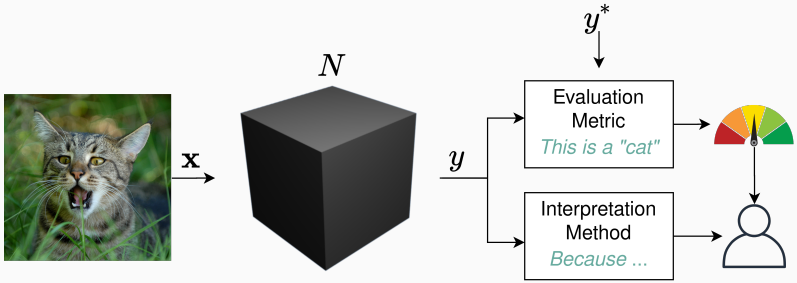
Explainable Machine Learning

- Goal: Make stakeholders and customers comfortable
- Problem: Tradeoff between complexity and interpretability
- $\rightarrow$ Need for methods helping with uncovering the *why*



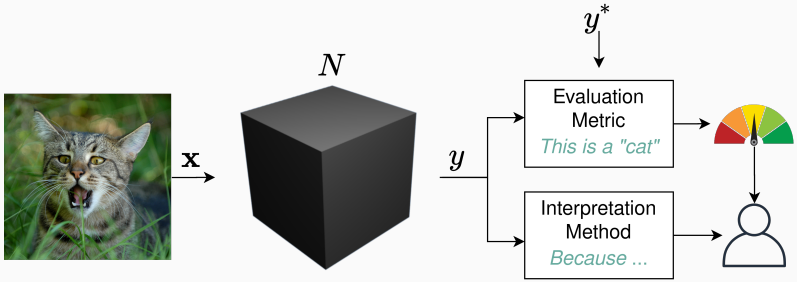
Terminology

- *Interpretability*: Observation of cause and effect ~ uncover the *why*
- *Explainability*: Observation of inner workings ~ uncover the *how*



Terminology

- ***Interpretability***: Observation of cause and effect ~ uncover the *why*
  - *Post-hoc* interpretability: interpretations are computed by applying methods that analyze the model after training
- *Explainability*: Observation of inner workings ~ uncover the *how*



Interpretation Methods

Types of Interpretation Methods

Local Methods

- why you think this image is a cat?

Model agnostic methods

- the model is seen as black box (no access to the parameters, we can only query the model)
- the underlying model is approximated with a surrogate model

Global Methods

- what does a cat look like?

Model transparent methods

- the model is known and the parameters can be accessed

Methods

Interpreter Fooling

Adversarial Setting

- why are interpretations so important and why is model fooling dangerous?

Adversarial Setting:

Implications of Interpreter Fooling

Methods

Methods

now, we want to take a look it specific findings of researchers from the past years