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

Verena Heusser

29 Januar 2021

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

Omnipresent Machine Learning

- Verification of machine learning (ML) algorithms mostly w.r.t. accuracy and efficiency
- Also:
  - Regulations: Right to explanation (GDPR 2018)
  -

Motivation

Interpretation Methods

Adversarial Setting

Adversarials: How to fool a model

- Adversarial examples []

Interpreter Manipulation Methods

Fooling Examples

References

Sources