Explainable Machine Learning

Presented by:
Manuel Capel

Fablab Winti
10. June 2024

Table of Content

What is Machine Learning

AI: Program that can learn how to perform a task requiring typical “human intelligence”

ML: AI algorithm learning to perform a precise task given training data.
Tasks:

- Classification: objects on images, disease…
- Regression: predicting stock exchange, conversion rate on a website…
- Reinforcement learning: games like chess or go, robots in an environment…
- Generative: text (ChatGPT), images (DALL-E)...

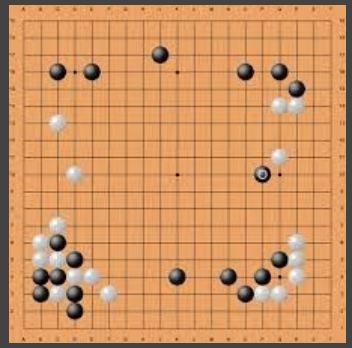
Deep Learning: ML algorithms using *neural networks*
Main Architectures:

- CNN (convolutional): especially for images
- RNN (recurrent): for sequence predictions (time series…)
- Transformers: for generative models

Black Box Effect

Alpha Go, Game 2 vs. Lee Segol, move 37 (article)

- Absolutely brilliant
- Admittedly no human would have thought of it
- Where does it come from??

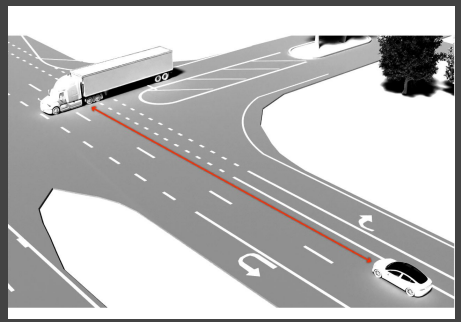


Tesla crash (article):

- Truck on the lane not evaluated as obstacle
- Why??

ChatGPT Hallucinations (article)

- Leonardo da Vinci painted the Mona Lisa in 1815
- False quotes, facts, article citations…
- Where are they from?



Why Explainability Matters

Legal
- Even if way less error-prone than humans
- Who is to blame if an error occurs?
- Especially when human lives involved
- E.g. medical decisions, self-driving…

Moral biases
- Are the reasons behind a decision acceptable vs. learned biases?
- Learned discrimination
- E.g. credit attribution, CV selection…

=> More trust, fairness, possibility to improve..

What is explainability

“Explain the behavior in human terms” - AWS documentation

“Set of processes and methods to provide a clear and human-understandable explanation”

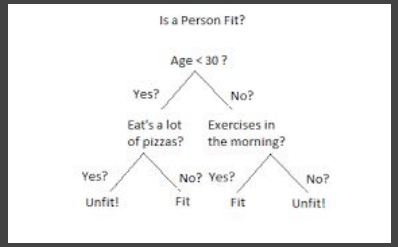
Etc.

Cyclical definition => “I know it when I see it” (Stewart test)

Topology of Explainability

Three main types of explainability

1. By design:
   a. Decision tree
   b. polynomial model
   c. Bayesian rules…

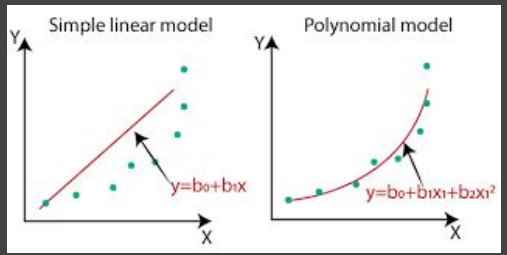


2. Surrogate: Approximate black-box by model explainable by design

3. Ex-post: Feature importance, Shapley values…

Surrogate and ex-post can be:

- Local
- Global



Future: Interpretation / explanation with LLMs like ChatGPT ?
Encoding the situation and the decision in a way the LLM can understand so it's provide an human understandable explanation, see paper

Conclusion

- ★ “*Computers are incredibly fast, accurate and stupid, humans are incredibly slow, inaccurate and brilliant, together they are powerful beyond imagination*” - Einstein

- ★ Darpa Project (archived)
- ★ From ML to Explainable AI - Holzinger

Demo: see notebook

Questions?