Global Explainability (XAI) Techniques

Quantifying the uncertainty of the explanations

Vasilis Gkolemis[1]

[1]ATHENA Research and Innovation Center

November 2021

Program

Hypothetical (?) scenarios

- The computer vision subsystem of an autonomous vehicle leads the vehicle to take a left turn, in front of a car moving in the opposite direction[1]

[1] https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco

[2] https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-l

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Hypothetical (?) scenarios

- The computer vision subsystem of an autonomous vehicle leads the vehicle to take a left turn, in front of a car moving in the opposite direction[1]
- The credit assessment system leads to the rejection of an application for a loan - the client suspects racial bias[2]

[1] https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco

[2] https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-l

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Hypothetical (?) scenarios

- The computer vision subsystem of an autonomous vehicle leads the vehicle to take a left turn, in front of a car moving in the opposite direction[1]
- The credit assessment system leads to the rejection of an application for a loan - the client suspects racial bias[2]
- A model that assesses the risk of future criminal offenses (and used for decisions on parole sentences) is biased against black prisoners[3]

[1] https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco

[2] https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-l

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Questions

- Why did the model make a specific decision? local XAI
- What could we change so that the model will make a different decision? counterfactual
- Can we summarize the model's behavior? global XAI
- Models as knowledge extractors, what hat the model learnt global XAI

Interpretability of Machine Learning Models

Qualitative definitions:

- "Interpretability is the degree to which a human can understand the cause of a decision" [4]

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

[5] 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).

[6] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences, 116(44), 22071-22080. (2019)

Interpretability of Machine Learning Models

Qualitative definitions:

- "Interpretability is the degree to which a human can understand the cause of a decision" [4]
- "Interpretability is the degree to which a human can consistently predict the model's result"[5]

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

[5] 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).

[6] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences, 116(44), 22071-22080. (2019)

Interpretability of Machine Learning Models

Qualitative definitions:

- "Interpretability is the degree to which a human can understand the cause of a decision" [4]
- "Interpretability is the degree to which a human can consistently predict the model's result" [5]
- "Extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model" [6]

---

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

[5] 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).

[6] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences, 116(44), 22071-22080. (2019)

My understanding

Interpretability is the degree to which a human can understand the reasoning process for a (specific) prediction

- interpretability: either by-design or assisted by a post-hoc XAI technique
- degree: non binary, interpretability is a spectrum
- human: interpretability is a human-centric procedure
- reasoning process: mechanism for predicting

Global vs Local

- Local
  - Interpret the model's output for a particular input
  - Extract interpretable quantity that holds for x close to $x^{(i)}$
- Global
  - Provide a general interpretation of the model's behavior
  - Extract interpretable quantity that holds for $x \in \mathcal{X}$

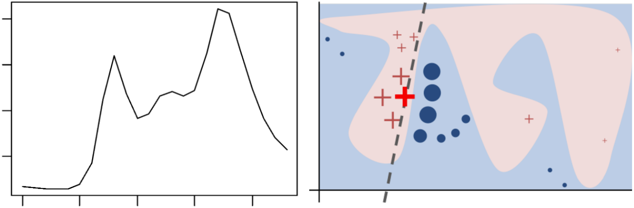


Figure: (Left) Global vs (Right) Local

Challenges on global methods

Extract an interpretable quantity that holds for $x \in \mathcal{X}$

- Fidelity: does the interpretable quantity mimics the model's behavior?
- Interpretability: is the extracted quantity interpretable enough?
- Can we have both?
  - if yes, why not replacing the original model with an interpretable one?
  - if no, how to deal with the trade-off?

Spoiler: Maybe uncertainty helps...

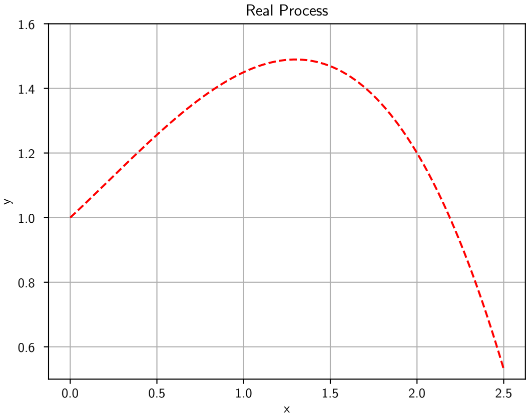
Methods we will discuss

- Feature Effect
- Feature Interaction
- Feature Importance

Program

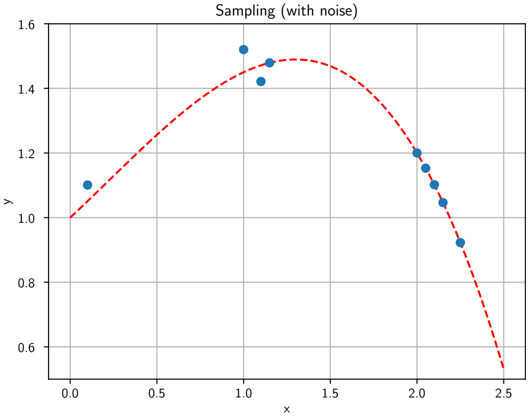
Example

Consider the following mapping $x \rightarrow y$



Example

Process unknown $\rightarrow$ we only have samples



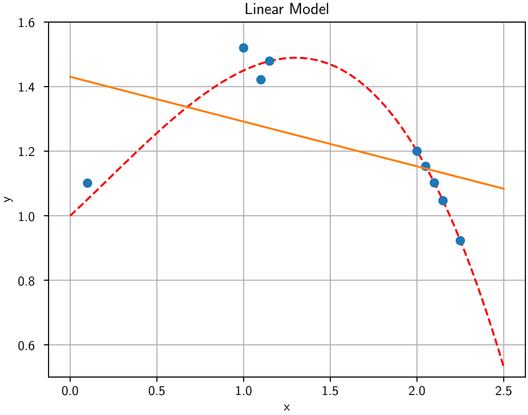
Example

Our goal is to model the process using the available samples (regression)

Example

Linear model $\rightarrow$ Underfiting!

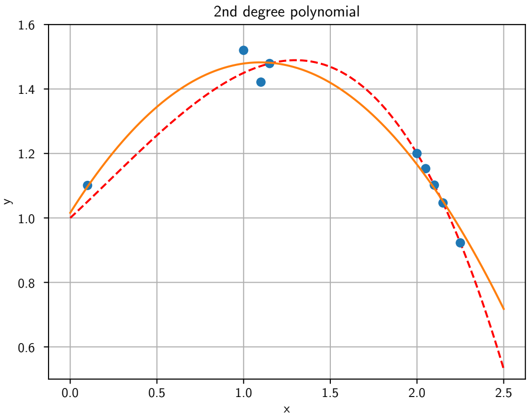
$$y = w_1 \cdot x + w_0$$



Example

2^{nd} degree polynomial $\rightarrow$ Decent Fit!

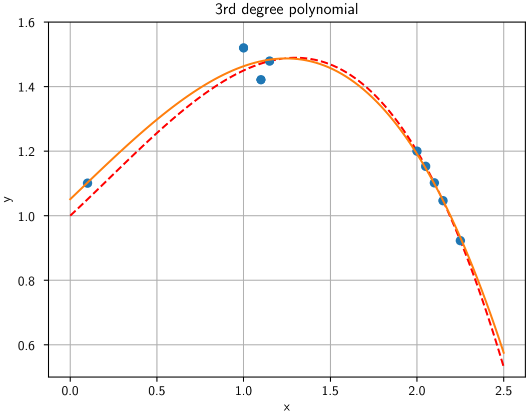
$$y = w_2 \cdot x^2 + w_1 \cdot x + w_0$$



Example

3^{rd} degree polynomial $\rightarrow$ Good Fit!

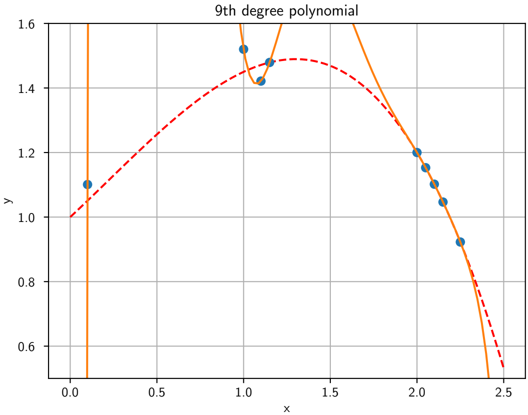
$$y = w_3 \cdot x^3 + w_2 \cdot x^2 + w_1 \cdot x + w_0$$



Example

9^{th} degree polynomial $\rightarrow$ Overfitting!

$$y = \sum_{i=0}^{9} w_i \cdot x^i$$



Problem diagnosis

- Model behavior is *explained* by the shape of the function
- Overfitting, Underfitting are easily diagnosed
- If the input has multiple dimensions D?
  - We often have tens or hundreds of features
  - Images and signals: Several thousands of input dimensions

Bike Sharing Problem

- Predict Bike rentals per hour in California
- We have 11 features
  - e.g., month, hour, temperature, humidity, windspeed
- We fit a Neural Network $y = \hat{f}(\boldsymbol{x})$
- How to make a plot like before?
  - Feature Effect methods

Feature effect methods

- High-dimensional input space $\boldsymbol{x} \in \mathbb{R}^D$
  - $x_s \rightarrow$ feature of interest
  - $\mathbf{x}_c \rightarrow$ other features
- How do we isolate the effect of x_s?

Program

Program

Program