Interpretable Machine Learning

Global explainability techniques

Vasilis Gkolemis[1,2]

[1]ATHENA Research and Innovation Center

[2]Harokopio University of Athens

February 2022

Motivation

- The credit assessment system leads to the rejection of an application for a loan - the client suspects racial bias[1]

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

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

Motivation

- The credit assessment system leads to the rejection of an application for a loan - the client suspects racial bias[1]
- A model that assesses the risk of future criminal offenses (and used for decisions on parole sentences) is biased against black prisoners[2]

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

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

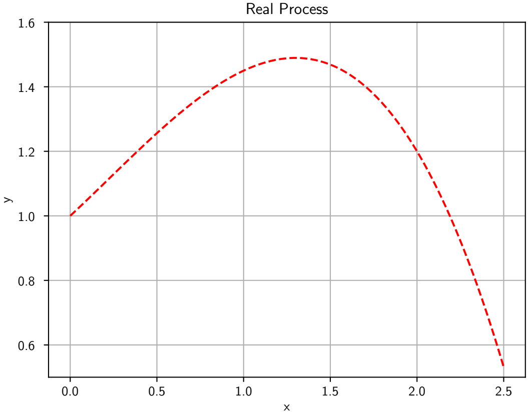
Interpretability

- Questions:
  - Why did a model make a specific decision?
  - Can we summarize the model's behavior?
- Answer:
  - Interpretable Machine Learning (IML)
  - "Extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model"[3]

[3] 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)

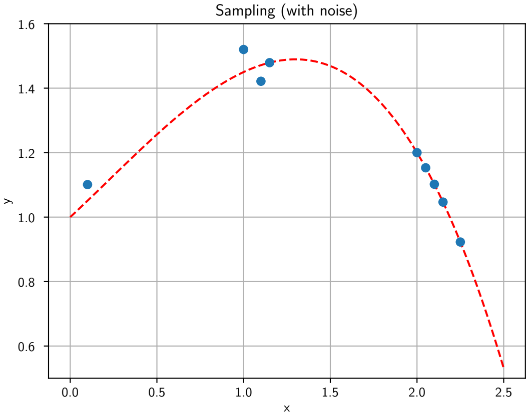
Example

Consider the following mapping $x \rightarrow y$



Example

Process unknown → we only have samples



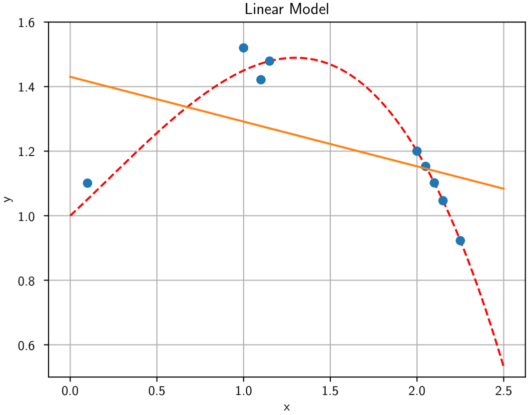
Example

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

Example

Linear model → Underfiting!

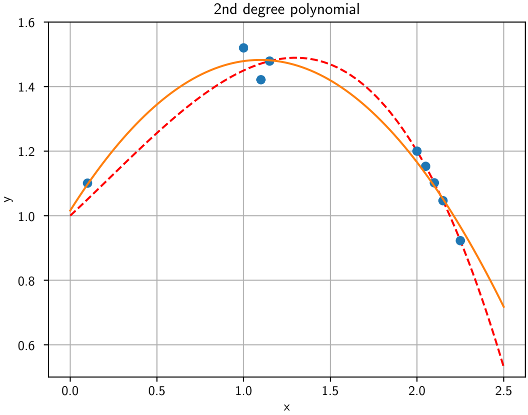
$$y = w_1 \cdot x + w_0$$



Example

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

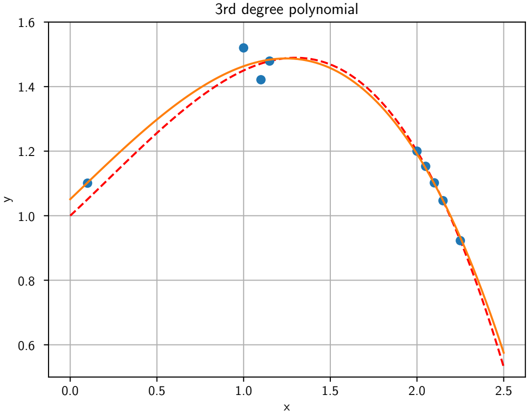
$$y = w_2 \cdot x^2 + w_1 \cdot x + w_0$$



Example

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

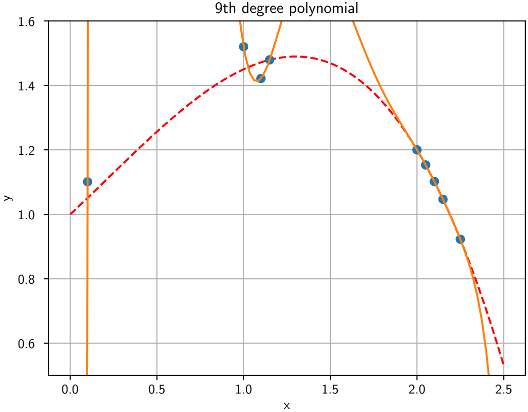
$$y = w_3 \cdot x^3 + w_2 \cdot x^2 + w_1 \cdot x + w_0$$



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 $y = \hat{f}(\mathbf{x})$
- 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?

Partial Dependence Plots (PDP)

- Proposed by J. Friedman on 2001[4] and is the marginal **effect** of a feature to the model output:

$$f_s(x_s) = E_{X_c}\left[\hat{f}(x_s, X_c)\right]$$

- Computation:

$$\hat{f}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_s, \mathbf{x}_c^{(i)})$$

[4] J. Friedman. "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232

Partial Dependence Plots (PDP)

Bike sharing Dataset:

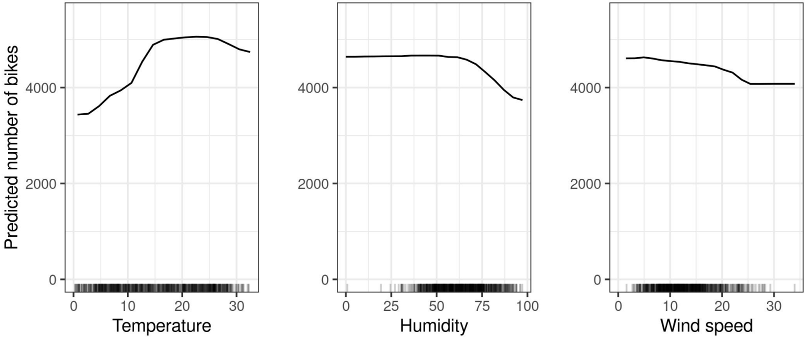


Figure: C. Molnar, IML book

[4] J. Friedman. "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232

Issues with PDPs

- The marginal distribution ignores correlated features!
- To compute the effect of temperature = 33 degrees it will (also) use an instance with month = January

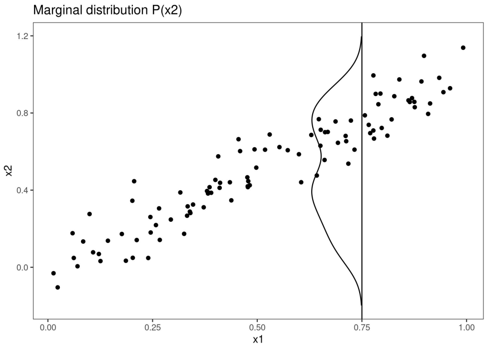


Figure: C. Molnar, IML book

Accumulated Local Effects (ALE)[5]

- Resolves problems that result from the feature correlation by computing differences over a (small) window
- Definition: $f(x_s) = \int_{x_{min}}^{x_s} \mathbb{E}_{\mathbf{x}_c|z}[\frac{\partial f}{\partial x_s}(z, \mathbf{x}_c)]\partial z$

[5] D. Apley and J. Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4 (2020): 1059-1086.

ALE approximation

$$\text{Approximation: } f(x_s) = \sum_{k=1}^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \underbrace{[f(z_k, \mathbf{x}_c^i) - f(z_{k-1}, \mathbf{x}_c^i)]}_{\text{point effect}}}_{\text{bin effect}}$$

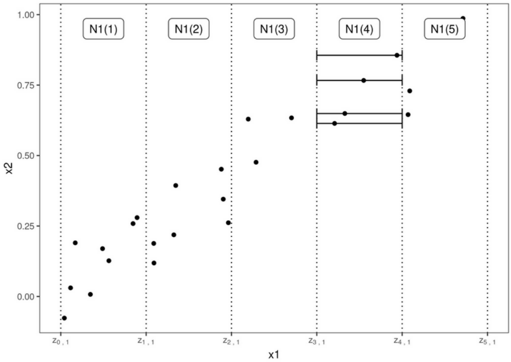


Figure: C. Molnar, IML book

ALE plots - examples

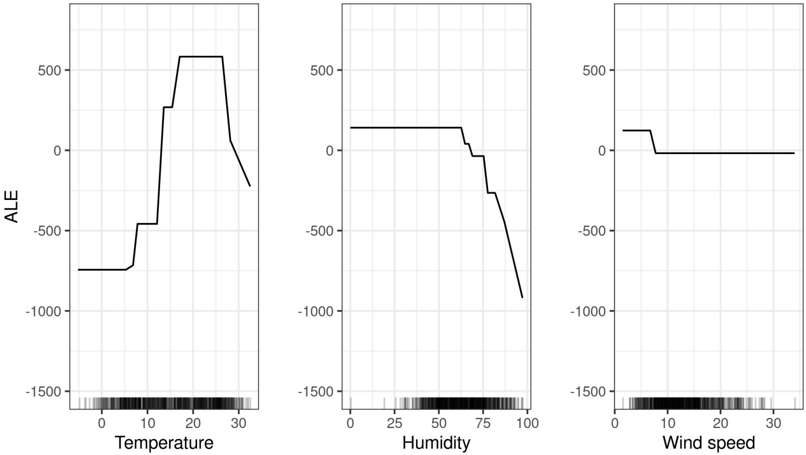


Figure: C. Molnar, IML book

Our work

- Differential Accumumulated Local Effects (DALE)
  - Asian Conference in Machine Learning (ACML 2022)
  - work done with my supervisors: Christos Diou, Theodore Dalamagas
- More efficient and accurate extension of ALE
- Works only with differential models (like Neural Networks)
- https://arxiv.org/abs/2210.04542