Global Explainability (XAI) Techniques

State of the art, challenges and the role of uncertainty

Vasilis Gkolemis[!*]

[!]ATHENA Research and Innovation Center

[*]Harokopio University of Athens

March 2023

Program

1. Intro to XAI (5')
2. Feature Effect (15')
   - PDP
   - ALE
   - DALE
3. Feature Interaction (5')
4. Heterogeneous effects and uncertainty (5')
5. Feature Importance (5')
6. Non-Tabular case (5')
7. Summary/Discussion (2')

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
- If models are knowledge extractors, what have they learned? 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 (2017)
[5] Kim et. al (2016)
[6] Murdoch et. al (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 (2017)
[5] Kim et. al (2016)
[6] Murdoch et. al (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 (2017)
[5] Kim et. al (2016)
[6] Murdoch et. al (2019)

Global vs Local

- Global
  - Provide a general interpretation of the model's behavior
  - Extract interpretable quantity that holds for $x \in \mathcal{X}$
  - Example: Feature Effect $x_s \rightarrow y$
- Local
  - Interpret the model's output for a particular input
  - Extract interpretable quantity that holds for x close to $x^{(i)}$
  - Example: Linear model that replaces f around $x^{(i)}$ (LIME)

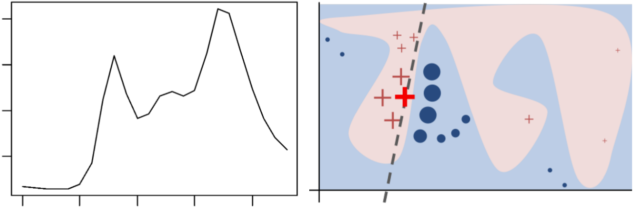


Figure: (Left) Global vs (Right) Local

Advantages and challenges of global methods

- Advantages:
  - Interpretable quantity holds for $x \in \mathcal{X}$
  - Global summary of the model's behavior
- Challenges:
  - Interpretable quantity holds for $x \in \mathcal{X}$
  - Level of fidelity
  - Level of interpretability
  - Can we have both?
    - if yes, replace the original model
    - if no, deal with the trade-off

---

if no, can uncertainty quantify the level of fidelity?

Methods we will discuss

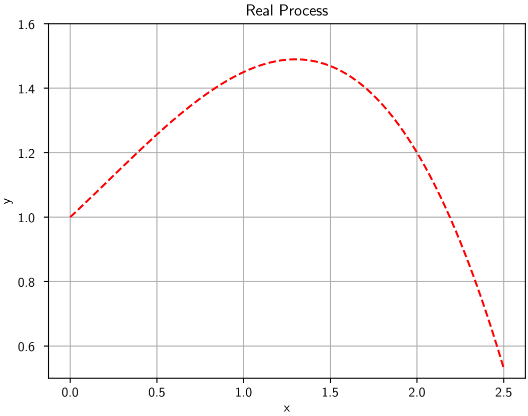
1. Feature Effect
   - $1D$ plot: $f_s(x_s) : x_s \rightarrow y$
   - Effect (mapping) of a single feature x_s on the output y
   - Apley et. al (2019)
2. Feature Interaction
   - Number
   - Level of interaction between features x_i and x_j
   - Greenwell et. al (2018)
3. (1) + (2) $\rightarrow$ Heterogeneous Effects / Uncertainty
4. Feature Importance
   - Number
   - To what extent the model accuracy would drop, if x_s was absent
   - Fisher et. al (2018)

Program

1. Intro to XAI (5')
2. Feature Effect (15')
   - PDP
   - ALE
   - DALE
3. Feature Interaction (5')
4. Heterogeneous effects and uncertainty (5')
5. Feature Importance (5')
6. Non-Tabular case (5')
7. Summary/Discussion (2')

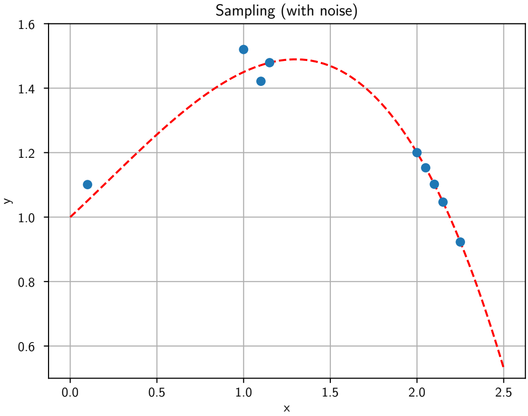
Example

Consider the following mapping $x \to y$



Example

Process unknown → we only have samples



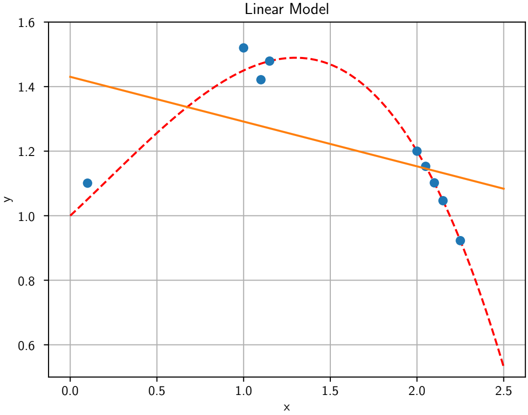
Example

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

Example

Linear model → Underfitting!

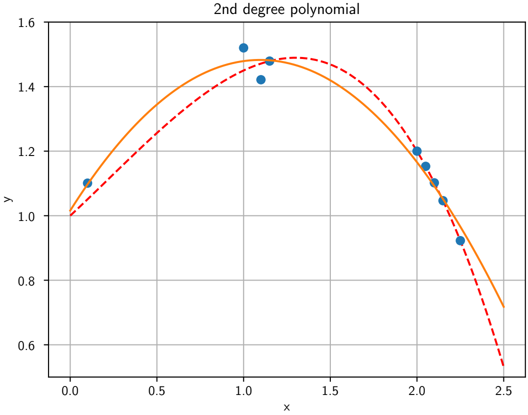
$$y = w_1 \cdot x + w_0$$



Example

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

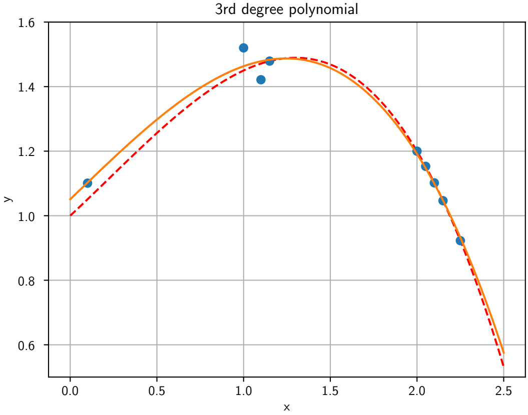
$$y = w_2 \cdot x^2 + w_1 \cdot x + w_0$$



Example

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

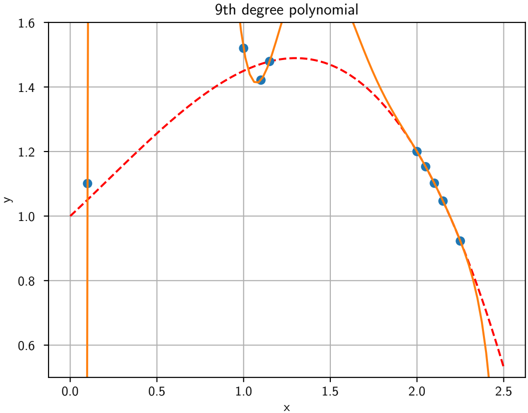
$$y = w_3 \cdot x^3 + w_2 \cdot x^2 + w_1 \cdot x + w_0$$



Example

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

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



Feature Effect Methods

- Model behavior is *explained* by the shape of the function
- Overfitting, Underfitting are easily diagnosed
- If high-dimensional input $\boldsymbol{x} \in \mathbb{R}^D$?
  - Tabular data; tens or hundreds of features
  - Images and signals; several thousands of input dimensions
- $x_s \rightarrow$ feature of interest
- $\mathbf{x}_c \rightarrow$ other features
- How do we isolate the effect of x_s?

Running Example: Bike Sharing Problem

- Predict Bike rentals per hour in California
- We have 11 features
  - e.g., month, hour, temperature, humidity, windspeed
- We fit a ML model $y = \hat{f}(\boldsymbol{x})$

How each feature affects the output?

Partial Dependence Plots (PDP)

- Proposed by J. Friedman on 2001[7] 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:

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

[7] Friedman et. al (2001)

Partial Dependence Plots (PDP)

Bike sharing Dataset:

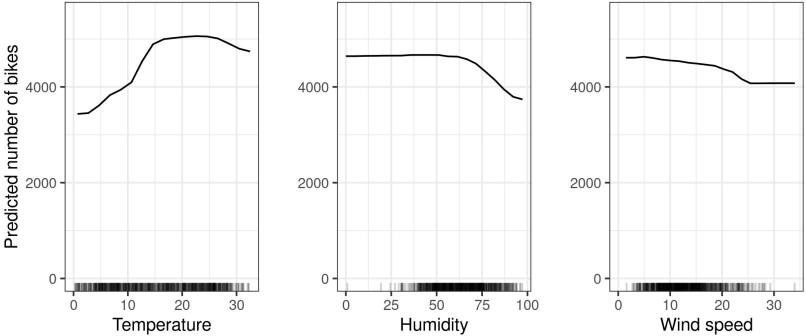


Figure: C. Molnar, IML book

[7] Friedman et. al (2001)

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

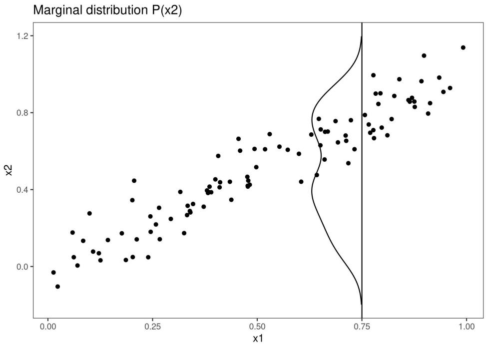


Figure: C. Molnar, IML book

Accumulated Local Effects (ALE)[8]

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

[8] 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

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

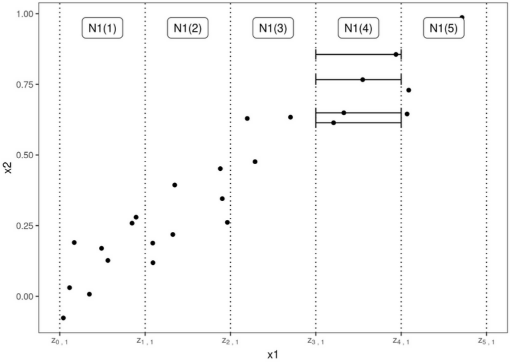


Figure: C. Molnar, IML book

ALE plots - examples

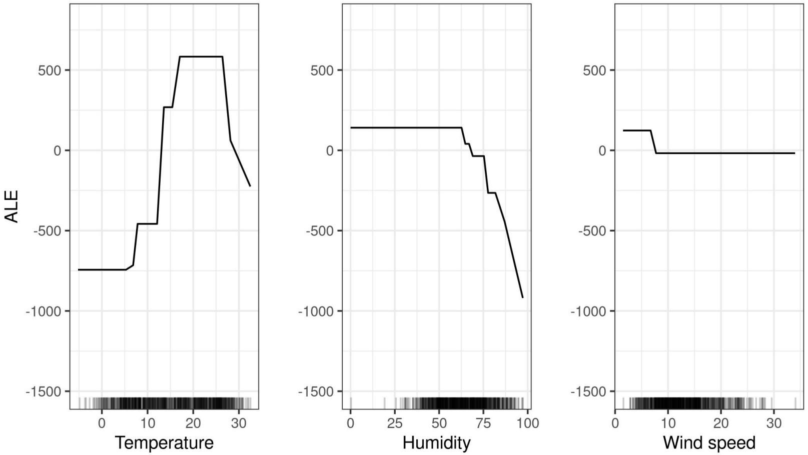


Figure: C. Molnar, IML book

Our work

- Differential Accumulated Local Effects (DALE)
  - Asian Conference in Machine Learning (ACML 2022)
  - Work done with: 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

Our proposal: Differential ALE

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

- Point Effect $\Rightarrow$ evaluation on instances
  - Fast $\rightarrow$ use of auto-differentiation, all derivatives in a single pass
  - Versatile $\rightarrow$ point effects computed once, change bins without cost
  - Secure $\rightarrow$ does not create artificial instances

For differentiable models, DALE resolves ALE weaknesses

DALE is faster

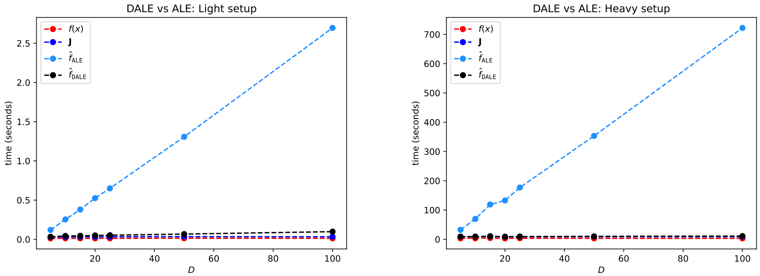
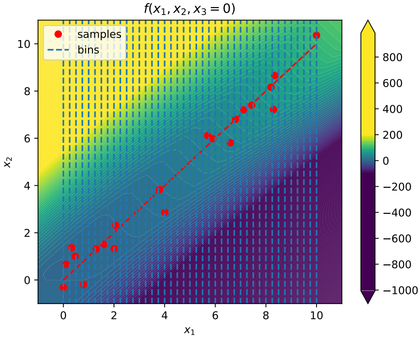


Figure: Light setup; small dataset ($N = 10^2$ instances), light f. Heavy setup; big dataset ($N = 10^5$ instances), heavy f

---

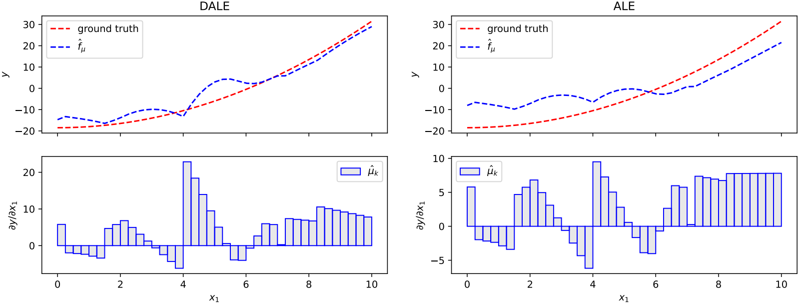
DALE accelerates the estimation

DALE may be more accurate - 40 Bins



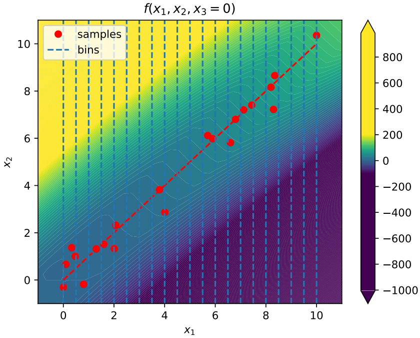
- DALE: on-distribution, noisy bin effect → poor estimation
- ALE: on-distribution, noisy bin effect → poor estimation

DALE may be more accurate - 40 Bins



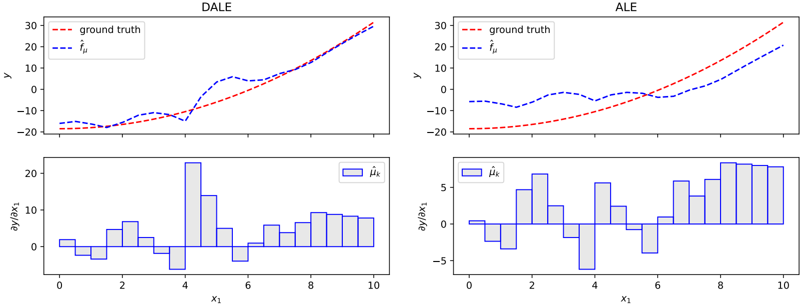
- DALE: on-distribution, noisy bin effect → poor estimation
- ALE: on-distribution, noisy bin effect → poor estimation

DALE may be more accurate - 20 Bins



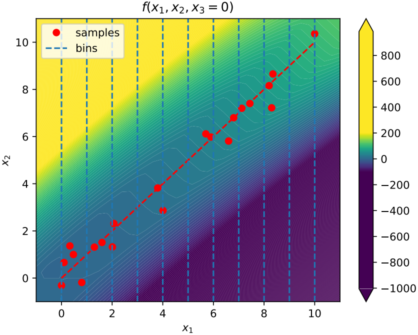
- DALE: on-distribution, noisy bin effect → poor estimation
- ALE: on-distribution, noisy bin effect → poor estimation

DALE may be more accurate - 20 Bins



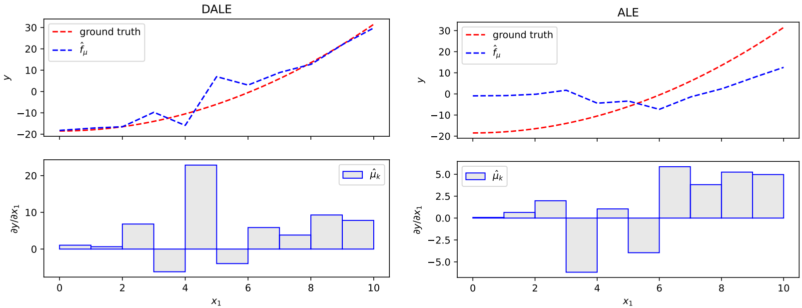
- DALE: on-distribution, noisy bin effect → poor estimation
- ALE: on-distribution, noisy bin effect → poor estimation

DALE may be more accurate - 10 Bins



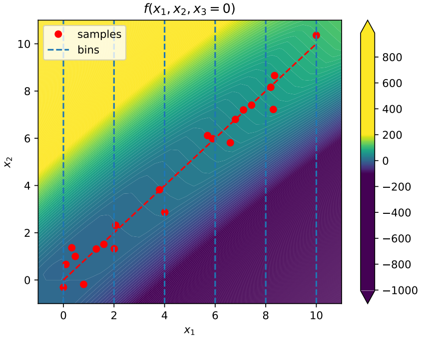
- DALE: on-distribution, noisy bin effect → poor estimation
- ALE: starts being OOD, noisy bin effect → poor estimation

DALE may be more accurate - 10 Bins



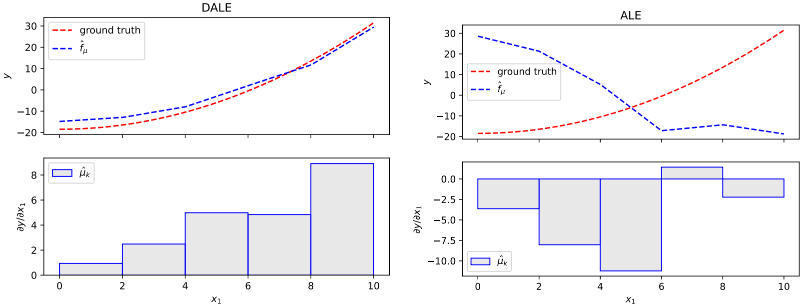
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: starts being OOD, noisy bin effect $\rightarrow$ poor estimation

DALE may be more accurate - 5 Bins



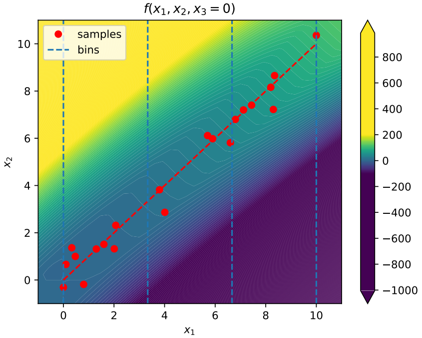
- DALE: on-distribution, robust bin effect → good estimation
- ALE: completely OOD, robust bin effect → poor estimation

DALE may be more accurate - 5 Bins



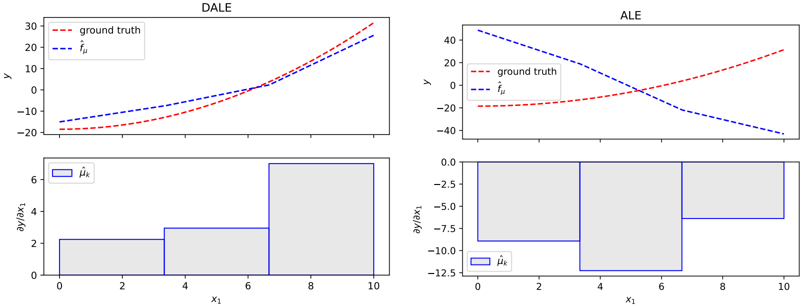
- DALE: on-distribution, robust bin effect $\rightarrow$ good estimation
- ALE: completely OOD, robust bin effect $\rightarrow$ poor estimation

DALE may be more accurate - 3 Bins



- DALE: on-distribution, robust bin effect → good estimation
- ALE: completely OOD, robust bin effect → poor estimation

DALE may be more accurate - 3 Bins



- DALE: on-distribution, robust bin effect → good estimation
- ALE: completely OOD, robust bin effect → poor estimation

Program

1. Intro to XAI (5')
2. Feature Effect (15')
   - PDP
   - ALE
   - DALE
3. Feature Interaction (5')
4. Heterogeneous effects and uncertainty (5')
5. Feature Importance (5')
6. Non-Tabular case (5')
7. Summary/Discussion (2')

Feature Interaction - Motivation

- Is Feature Effect a good approach?
  - Interpretability $\rightarrow$ very good, easy intuition
  - Fidelity $\rightarrow$ it depends..
- Additive case: $f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2)$
  - Generalized Additive Models
  - X-by-design
- Non-additive case: $f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2) + \underbrace{f_{12}(x_1, x_2)}_{\textit{interaction}}$
  - how to distribute $f_{12}(x_1, x_2)$ to x_1 and x_2?
  - Research question; uncertainty could help
- f is unknown,
- what is the magnitude of the interaction terms?
- Feature Interaction methods!

Problem Statement

When features interact with each other in a prediction model, the prediction cannot be expressed as the sum of the feature effects, because the effect of one feature depends on the value of the other feature. Aristotle's predicate "The whole is greater than the sum of its parts" applies in the presence of interactions.[9]

[9] Interpretable Machine Learning book

H-statistic

- Level of interaction between feature i and feature j

$$\mathcal{H}_{jk}^2 = \frac{\sum_{i=1}^n \left(PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right)^2}{\sum_{i=1}^n PD_{jk}^2(x_j^{(i)}, x_k^{(i)})}$$

- Level of interaction between feature i and all the other features

$$\mathcal{H}_{j}^2 = \frac{\sum_{i=1}^n \left(f(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)}) \right)^2}{\sum_{i=1}^n f^2(x^{(i)})}$$

H-statistic

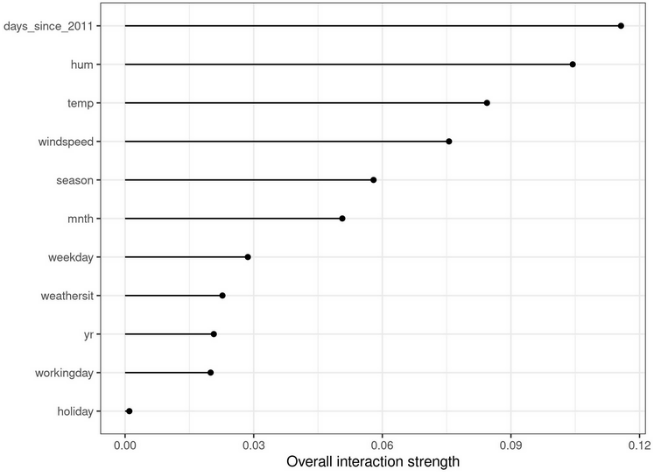


Figure: C. Molnar, IML book

Other approaches

- Greenwell's interaction index
  - PDP-based method
  - A Simple and Effective Model-Based Variable Importance Measure
- SHAP interaction index
  - SHAP-based method
  - Consistent Individualized Feature Attribution for Tree Ensembles

Program

Interaction implies heterogeneity

High interaction index $\rightarrow$ heterogeneous effects $\rightarrow$ low fidelity of Feature Effect plot

Example

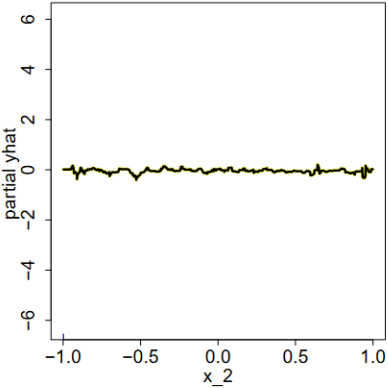


Figure: PDP plot, taken from Goldstein et. al

Interpretation? Maybe $y \perp\!\!\!\perp x_2$

Example

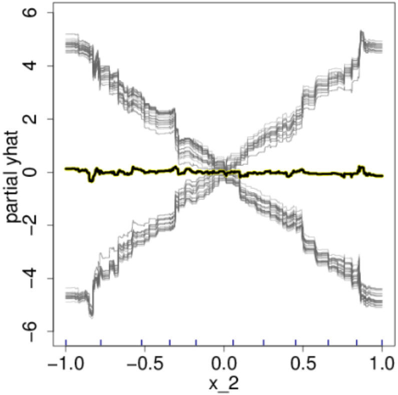


Figure: PDP-ICE plot, taken from Goldstein et. al

Interpretation now? Maybe $y \approx \pm 6x_2$ depending on a condition

Heterogeneity on PDP is called ICE

- Local effects, often, deviate from the global effect
- Aggregation bias → Mehrabi et. al. (2019)
- $\texttt{ICE}^{(i)}(x_s) = f(x_s, \mathbf{x}_c^{(i)})$
- Another approach

$$\rightarrow \mathbb{V}(x_s) = \frac{1}{N-1} \sum_i \left(f(x_s, x_c^{(i)}) - \underbrace{\frac{1}{N} \sum_i f(x_s, x_c^{(i)})}_{\mu(x_s)} \right)^2$$

- ICE show the *type* of heterogeneity, variance shows only the *magnitude*
- They both model the uncertainty of the feature effect!

ICE have the same limitations as PDPs under correlations!

Heterogeneity/Uncertainty on ALE

- The variance idea?
- Estimate the variance inside each bin?
- And then aggregate the variances?
- Bin splitting is important, otherwise biased estimation of the variance

---

Spoiler; we are working on it!

Regional Effect plots

- Heterogeneity → subspaces with homogeneous effects

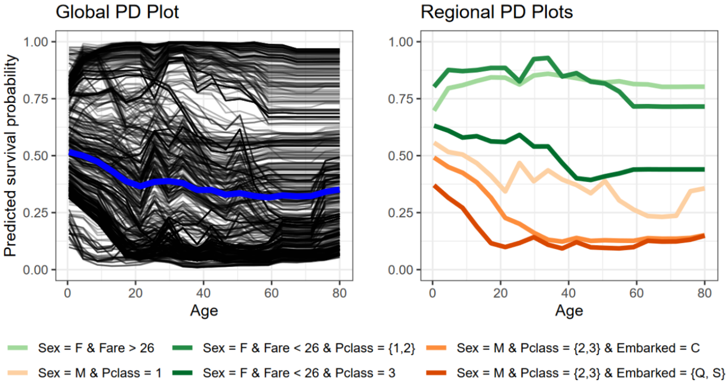


Figure: REPID: Regional Effect plots, taken from Herbinger et. al

Program

1. Intro to XAI (5')
2. Feature Effect (15')
   - PDP
   - ALE
   - DALE
3. Feature Interaction (5')
4. Heterogeneous effects and uncertainty (5')
5. Feature Importance (5')
6. Non-Tabular case (5')
7. Summary/Discussion (2')

Feature importance

- Many ways to define *importance*
- Permutation Feature Importance (PFI) measures the accuracy drop if we permute a feature

---

Algorithm:

- Estimate the original model error $e_{orig} = L(y, f(X))$
- For each feature $d \in \{1, \ldots D\}$
  - Generate feature matrix X_{perm} by permuting feature d in the data X
  - Estimate error $e_{perm} = L(y, f(X_{perm}))$
  - Calculate permutation feature importance as quotient $FI_d = \frac{e_{perm}}{e_o rig}$ or $FI_d = e_{orig} - e_{perm}$

Permutation Feature Importance (PFI)

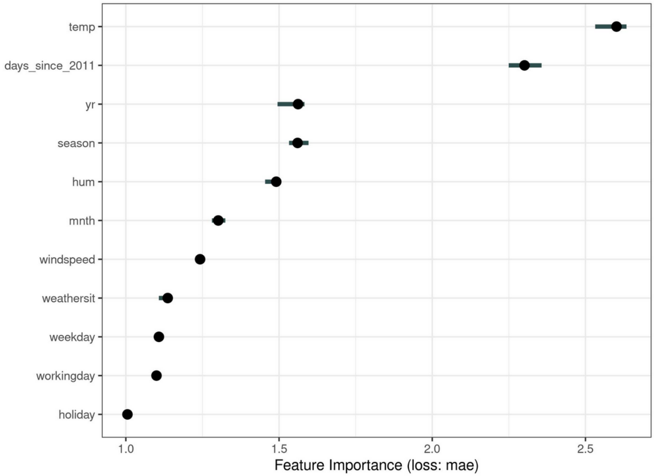


Figure: Image taken from Interpretable Machine Learning

Challenges/ideas for feature importance

- Other ideas:
  - Connection with feature effect
  - $\texttt{FI}_s = \int_{x_s} f_s(x_s)$, i.e. energy of the signal
- Challenges:
  - Just permute the feature and measure the accuracy or retrain on the permuted dataset?
  - If just permute, two highly correlated features may divide their importance (seem less important)
  - If just permute, we suffer from unrealistic instances
  - If retrain, two highly correlated features may cover each other (seem unimportant)

Program

Can global methods be applied in Images?

- Raw pixels do not have semantics

Figure: Grad-Cam, image taken from Adebayo et. al (2017)

Can global methods be applied in Images?

- Focus on the reasoning process of the CNN
- What makes images (in general) be classified as cats?
- Find prototypes!
- Unfortunately, not yet available, as a post-hoc explainability technique
- Only local prototypes can be found post-hoc

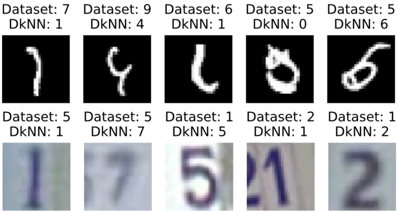


Figure: Deep KNN, image taken from Papernot et. al (2018)

Can global methods be applied in Images?

But prototype learning can be enforced in the model architecture

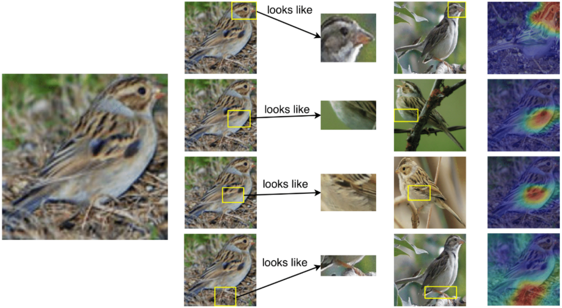


Figure: Deep KNN, image taken from Chen et. al (2018)

Program

1. Intro to XAI (5')
2. Feature Effect (15')
   - PDP
   - ALE
   - DALE
3. Feature Interaction (5')
4. Heterogeneous effects and uncertainty (5')
5. Feature Importance (5')
6. Non-Tabular case (5')
7. Summary/Discussion (2')

Summary

Global explainability techniques:

- provide important model summaries
- they suffer from fidelity issues
- uncertainty can help, i.e., global surrogate models that quantify the uncertainty
- straightforward application only on tabular data where features are meaningful

Summary

Papers:

- Uncertainty as a form of transparency, Bhatt et al. (2021)
- Reliable Post hoc Explanations: Modeling Uncertainty in Explainability, Slack et al. (2021)
- Explaining Hyperparameter Optimization via Partial Dependence Plots, Moosbauer (2021)

Questions?

Thank you for your attention!