New methods for global interpretability of differentiable machine learning models

Χρήστος Δίου / Christos Diou

Department of Informatics and Telematics
Harokopio University of Athens

Ημερίδα: «Η Επιστήμη των Δεδομένων στη Διαδικασία Λήψης Αποφάσεων σε Αβέβαιο Περιβάλλον: Σύγχρονες Τάσεις και Εφαρμογές»
Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
11/12/2023

Contents

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-learning/

[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-learning/

[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-learning/

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

Questions

- Why did a model make a specific decision?
- What could we change so that the model will make a different decision?
- Can we summarize and predict the model's behavior?

Today we focus on the last question

Taxonomy of interpretability methods

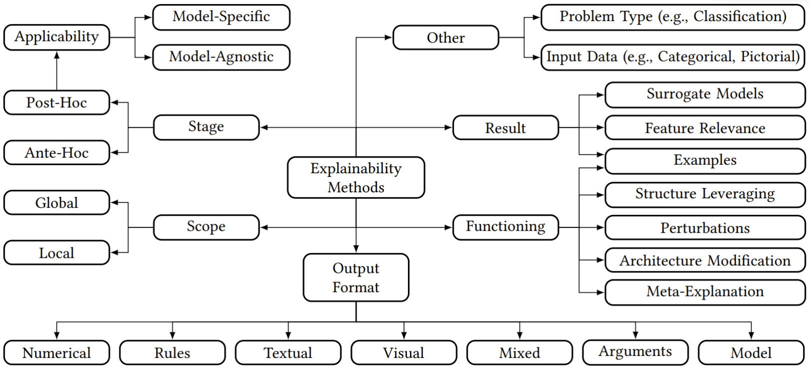


Figure: Timo Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods". In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 2022 [8]

Contents

Interpretable models (ante-hoc)

- Some models afford explanations
  - interpretable-by-design
- Examples, (generalized) linear models, decision trees, k-NN
- Example: Linear regression

$$\hat{y} = w_1 x_1 + \ldots + w_p x_p + b$$

Interpretable models (ante-hoc)

- Result in the bike sharing dataset (model weights)

$$\hat{y} = w_1 x_1 + \ldots + w_p x_p + b$$

	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSPRING	899.3	122.3	7.4
seasonSUMMER	138.2	161.7	0.9
seasonFALL	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

Figure: C. Molnar, IML book, 2022 [7]

Interpretable models (ante-hoc)

- Feature effects (visualization)

$$effect_j^{(i)} = w_j x_j^{(i)}$$

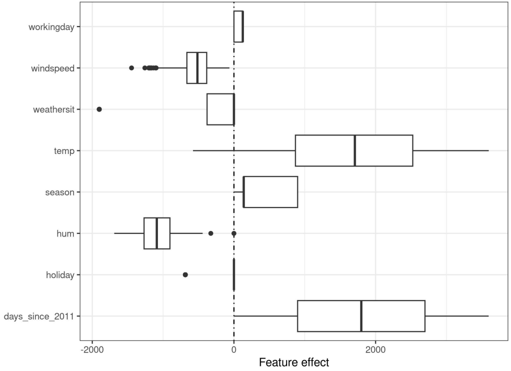


Figure: C. Molnar, IML book, 2022 [7]

Feature effect methods (1)

- Black-box model $f(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$, trained on $\mathcal{D}$
- Goal:
  - For single variable: Plot illustrating the effect of a feature x_s on f for all values of x_s
  - For pairs of variables: Plot illustrating the effect of pair (x_s, x_l) on f for all values of x_s and x_l

Feature Effect: global, model-agnostic, outputs plot

Feature Effect methods (2)

$y = f(x_s) \rightarrow$ plot showing the effect of x_s on the output y

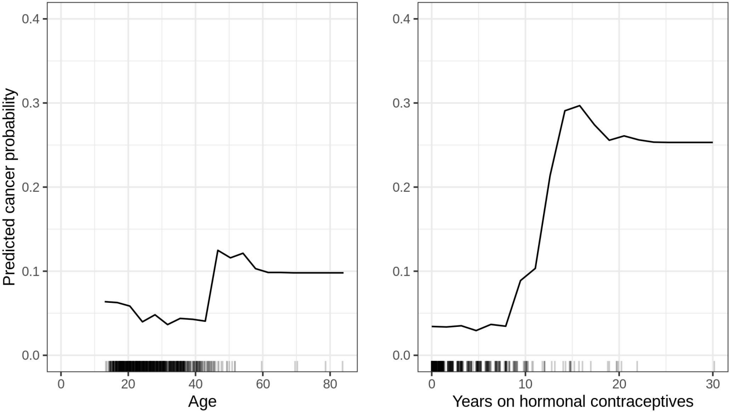


Figure: C. Molnar, IML book, 2022 [7]

Feature Effect is simple and intuitive

Feature Effect Methods (3)

- $x_s \rightarrow$ feature of interest, $\boldsymbol{x_c} \rightarrow$ other features
- How can we isolate x_s?
- Difficult task:
  - features are correlated
  - f has learned complex interactions

PDP, MPlot and ALE

- PDP (Friedman, 2001) [3]
  - $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}}[f(x_s, \boldsymbol{x_c})]$
  - **Unrealistic instances**
  - e.g. $f(x_{\texttt{age}} = 20, x_{\texttt{years\_contraceptives}} = 20) = ??$

PDP, MPlot and ALE

- PDP (Friedman, 2001) [3]
  - $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}}[f(x_s, \boldsymbol{x_c})]$
  - **Unrealistic instances**
  - e.g. $f(x_{\texttt{age}} = 20, x_{\texttt{years\_contraceptives}} = 20) = ??$

- MPlot Apley & Zhu, 2020 [1]
  - $\boldsymbol{x_c}|x_s$: $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}|x_s}[f(x_s, \boldsymbol{x_c})]$
  - **Aggregated effects**
  - Real effect: $x_{\texttt{age}} = 20 \rightarrow 10$, $x_{\texttt{years\_contraceptives}} = 20 \rightarrow 10$
  - MPlot may assign 17 to both

PDP, MPlot and ALE

- PDP (Friedman, 2001) [3]
  - $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}}[f(x_s, \boldsymbol{x_c})]$
  - **Unrealistic instances**
  - e.g. $f(x_{\texttt{age}} = 20, x_{\texttt{years\_contraceptives}} = 20) = ??$

- MPlot Apley & Zhu, 2020 [1]
  - $\boldsymbol{x_c}|x_s$: $f(x_s) = \mathbb{E}_{\boldsymbol{x_c}|x_s}[f(x_s, \boldsymbol{x_c})]$
  - **Aggregated effects**
  - Real effect: $x_{\texttt{age}} = 20 \rightarrow 10$, $x_{\texttt{years\_contraceptives}} = 20 \rightarrow 10$
  - MPlot may assign 17 to both

- ALE Apley & Zhu, 2020 [1]
  - $f(x_s) = \int_{x_{min}}^{x_s} \mathbb{E}_{\boldsymbol{x_c}|z}[\frac{\partial f}{\partial x_s}(z, \boldsymbol{x_c})]\partial z$
  - **Resolves both failure modes**

ALE approximation

ALE definition: $f(x_s) = \int_{x_{s,min}}^{x_s} \mathbb{E}_{\boldsymbol{x_c}|z}[\frac{\partial f}{\partial x_s}(z, \boldsymbol{x_c})]\partial z$

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

ALE approximation

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

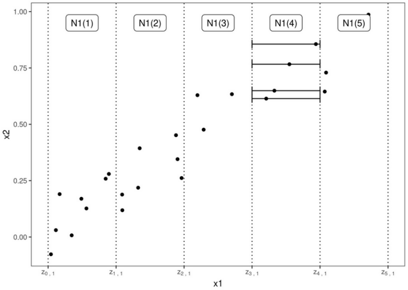


Figure: Image taken from Interpretable ML book [7]

Bin splitting (parameter K) is crucial!

ALE approximation - weaknesses

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

- Point Effect $\Rightarrow$ evaluation at bin limits
  - 2 evaluations of f per point $\rightarrow$ slow
  - change bin limits, pay again $2 * N$ evaluations of f $\rightarrow$ restrictive
  - broad bins may create out of distribution (OOD) samples $\rightarrow$ not-robust in wide bins

Contents

V. Gkolemis, T. Dalamagas and C. Diou, “DALE: Differential Accumulated Local Effects for efficient and accurate global explanations”, ACML 2022 [4]

Work in collaboration with Vasilis Gkolemis (PhD student @ HUA) and Theodoros Dalamagas (Researcher, ATHENA RC)

Our proposal: Differential ALE

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

- 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
  - Unbiased estimator of ALE (bias / variance proofs in the paper and supporting material)

For differentiable models, DALE resolves ALE weaknesses

DALE is faster and more versatile - theory

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

- Faster
    - gradients wrt all features $\nabla_{\boldsymbol{x}} f(\boldsymbol{x^i})$ in a single pass (via the Jacobian)
    - auto-differentiation must be available (deep learning)
- Versatile
    - Change bin limits, with near zero computational cost

DALE is faster and allows redefintion of the bin limits

DALE is faster and versatile - Experiments

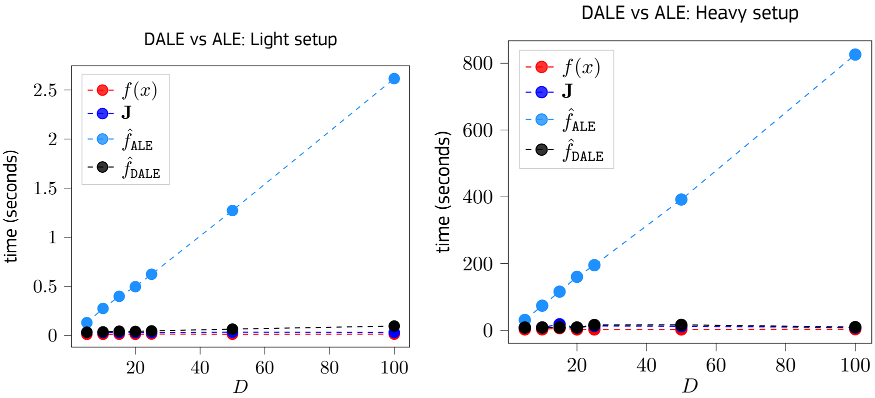


Figure: Light setup; small dataset ($N = 10^2$ instances), computationally light f. Heavy setup; big dataset ($N = 10^5$ instances), computationally heavy f. D is the number of dimensions.

DALE considerably accelerates the estimation

DALE uses on-distribution samples - Theory

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

- point effect independent of bin limits
  - $\frac{\partial f}{\partial x_s}(x_s^i, \boldsymbol{x_c^i})$ computed on real instances $\boldsymbol{x}^i = (x_s^i, \boldsymbol{x_c^i})$
- bin limits affect only the resolution of the plot
  - wide bins $\rightarrow$ low resolution plot, bin estimation from more points
  - narrow bins $\rightarrow$ high resolution plot, bin estimation from less points

DALE enables wide bins without creating out of distribution instances

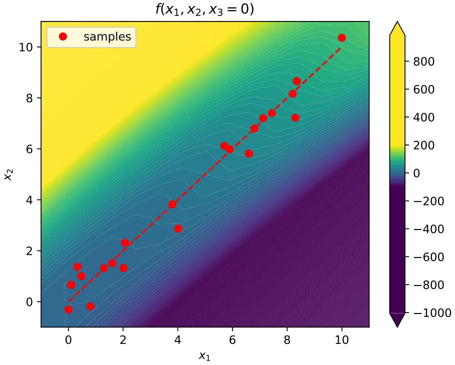
DALE uses on-distribution samples - Experiments

$$f(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 \pm g(x)$$

$$x_1 \in [0, 10], x_2 \sim x_1 + \epsilon, x_3 \sim \mathcal{N}(0, \sigma^2)$$

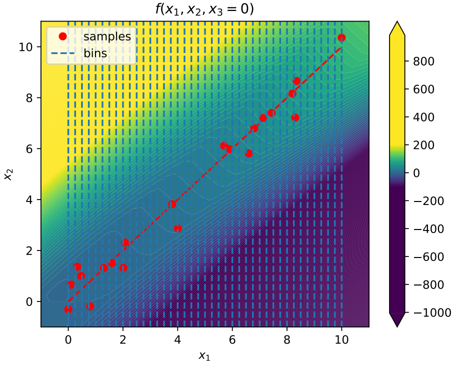
$$f_{\mathtt{ALE}}(x_1) = \frac{x_1^2}{2}$$

- point effects affected by $(x_1 x_3)$ (σ is large)
- bin estimation is noisy (samples are few)



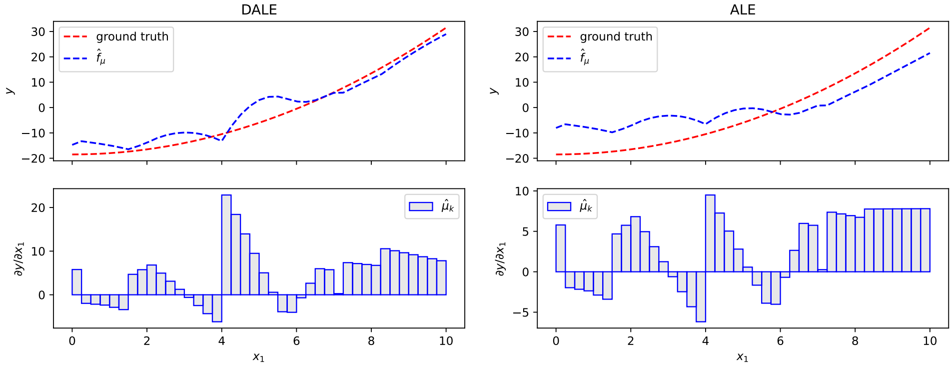
Intuition: we need wider bins (more samples per bin)

DALE vs ALE - 40 Bins



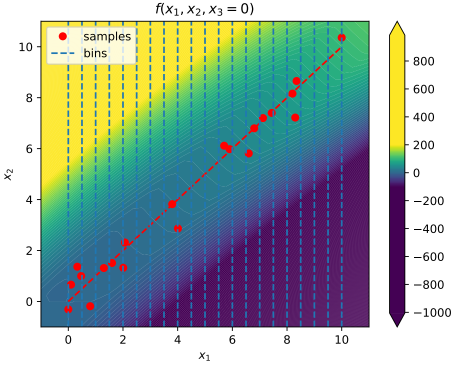
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation

DALE vs ALE - 40 Bins



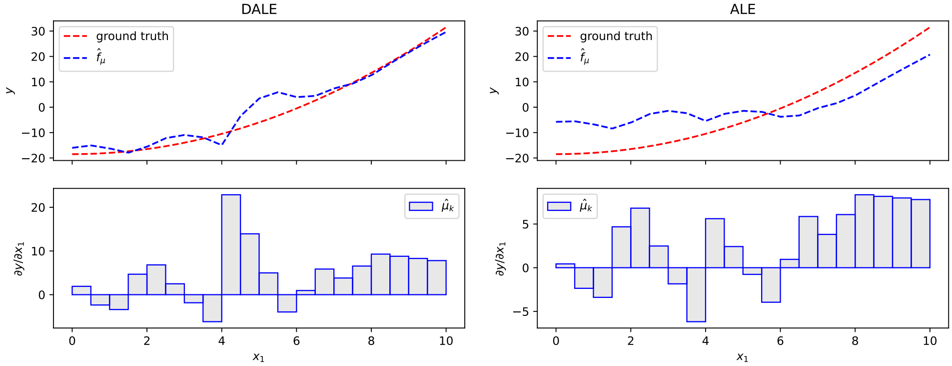
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation

DALE vs ALE - 20 Bins



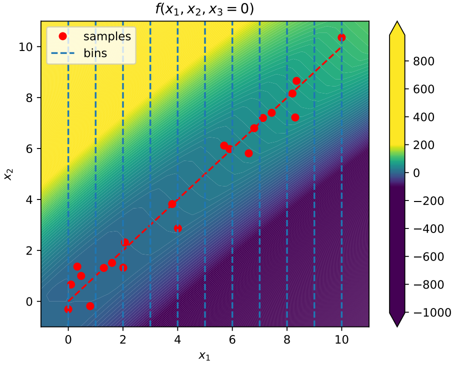
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation

DALE vs ALE - 20 Bins



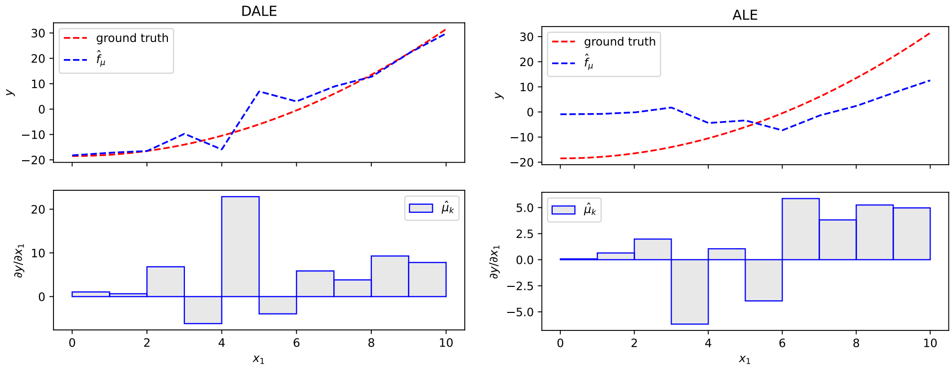
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation

DALE vs ALE - 10 Bins



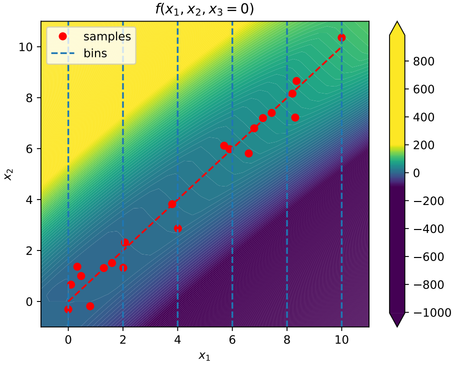
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: starts being OOD, noisy bin effect $\rightarrow$ poor estimation

DALE vs ALE - 10 Bins



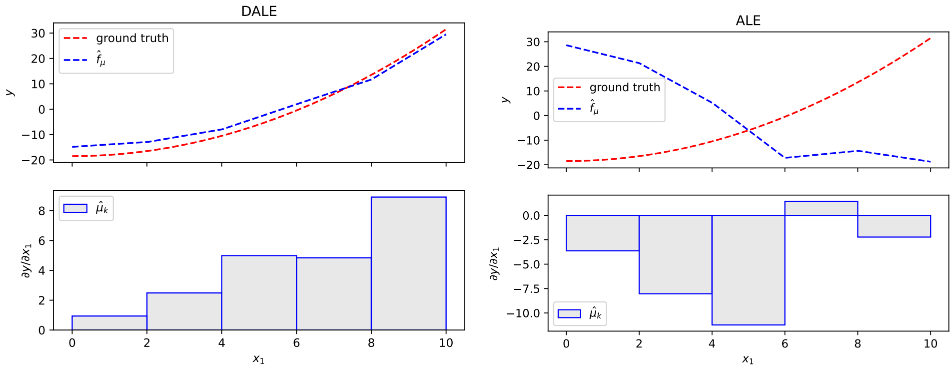
- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: starts being OOD, noisy bin effect $\rightarrow$ poor estimation

DALE vs ALE - 5 Bins



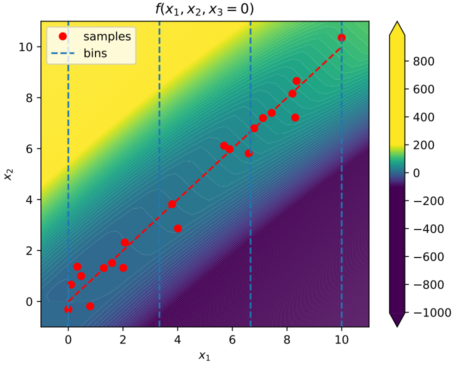
- DALE: on-distribution, robust bin effect $\rightarrow$ good estimation
- ALE: completely OOD, robust bin effect $\rightarrow$ poor estimation

DALE vs ALE - 5 Bins



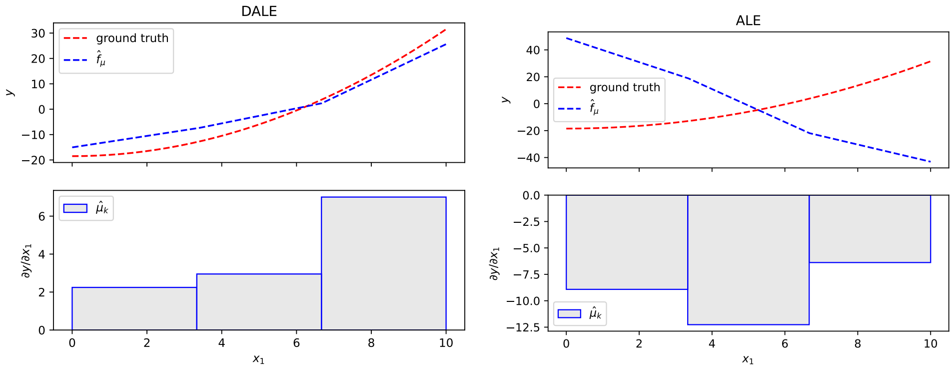
- DALE: on-distribution, robust bin effect $\rightarrow$ good estimation
- ALE: completely OOD, robust bin effect $\rightarrow$ poor estimation

DALE vs ALE - 3 Bins



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

DALE vs ALE - 3 Bins



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

Real Dataset Experiments - Efficiency

- Bike-sharing dataset [2]
- $y \rightarrow$ daily bike rentals
- x : 10 features, most of them characteristics of the weather

Efficiency on Bike-Sharing Dataset (Execution Times in seconds)

	Number of Features										
	1	2	3	4	5	6	7	8	9	10	11
DALE	1.17	**1.19**	**1.22**	**1.24**	**1.27**	**1.30**	**1.36**	**1.32**	**1.33**	**1.37**	**1.39**
ALE	**0.85**	1.78	2.69	3.66	4.64	5.64	6.85	7.73	8.86	9.9	10.9

DALE requires almost same time for all features

Real Dataset Experiments - Accuracy

- Difficult to compare in real world datasets
- We do not know the ground-truth effect
- In most features, DALE and ALE agree.
- Only $X_{\texttt{hour}}$ is an interesting feature

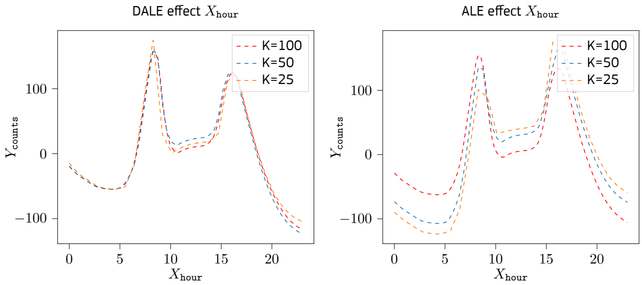


Figure: (Left) DALE (Left) and ALE (Right) plots for $K = \{25, 50, 100\}$

Contents

V. Gkolemis, T. Dalamagas, E. Ntoutsi and C. Diou, “RHALE: Robust and Heterogeneity-aware Accumulated Local Effects ”, ECAI 2023 [5]

Work in collaboration with Vasilis Gkolemis (PhD student @ HUA), Theodoros Dalamagas (Researcher, ATHENA RC) and Eirini Ntoutsi (Prof, Universität der Bundeswehr, München)



Next step: Heterogeneity and optimal bin selection

Using DALE, one has the computational margin to worry about additional issues:

- Computation of heterogeneity of local effects (i.e., standard error of the mean)
- Optimal selection of bins such that the effect does not have a high variation within the bin

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

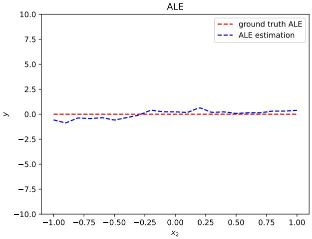
- Robust: Automatic bin splitting (result does not depend on arbitrary bin selection)
- Heterogeneity aware: $\pm$ from the average

Example (based on Goldstein et al [6])

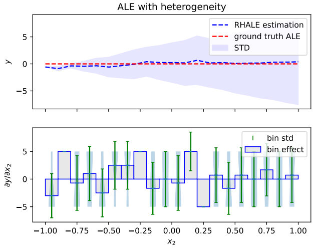
Aggregation bias

$$Y = 0.2X_1 - 5X_2 + 10X_2\mathbb{1}_{X_3>0} + \mathcal{E}$$

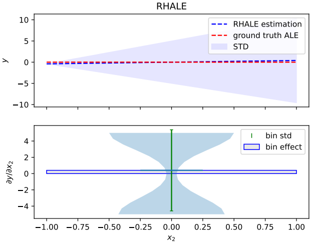
$$\mathcal{E} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad X_1, X_2, X_3 \overset{i.i.d.}{\sim} \mathcal{U}(-1,1)$$



x_2 ALE plot (20 bins)



x_2 ALE + heterogeneity (20 bins)



x_2 RHALE (auto-binning)

Definitions and Approximations - Main effect

ALE main effect definition

$$f^{\texttt{ALE}}(x_s) = \int_{x_{s,\min}}^{x_s} \underbrace{\mathbb{E}_{X_c|X_s=z}\left[f^s(z, X_c)\right]}_{\mu(z)} \partial z$$

ALE main effect approximation

$$\hat{f}^{\texttt{ALE}}(x_s) = \Delta x \sum_k^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\boldsymbol{x}^i \in \mathcal{S}_k} [\frac{\partial f}{\partial x_s}(x_s^i, \boldsymbol{x_c^i})]}_{\text{bin effect:}\hat{\mu}(z)}$$

Simple but wrong: ALE + Heterogeneity

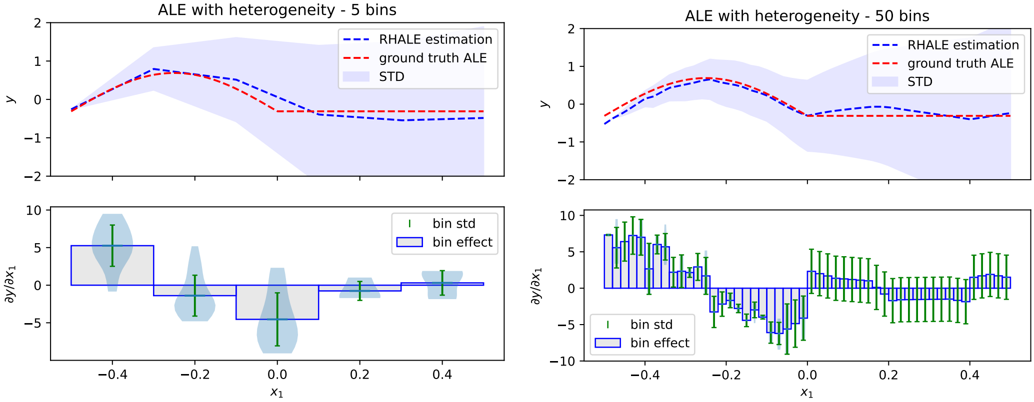


Figure: Left: approximation with narrow bin-splitting (5 bins) and (Right) with dense-bin splitting

- Fixed-size bin splitting can ruin the estimation of the heterogeneity

Definitions and Approximations - Heterogeneity

ALE heterogeneity definition

$$\sigma(x_s) = \sqrt{\int_{x_{s,\min}}^{x_s} \underbrace{\mathbb{E}_{X_c|X_s=z}\left[\left(f^s(z, X_c) - \mu(z)\right)^2\right]\partial z}_{\sigma^2(z)}}$$

ALE heterogeneity approximation

$$\mathtt{STD}(x_s) = \sqrt{\sum_{k=1}^{k_x}(z_k - z_{k-1})^2 \underbrace{\frac{1}{|\mathcal{S}_k| - 1} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left(f^s(\mathbf{x}^i) - \hat{\mu}(z_{k-1}, z_k)\right)^2}_{\hat{\sigma^2}(z)}}$$

Derivations

In the paper we formally prove

1. the conditions under which the above definition is an unbiase estimator of the heterogeneity
2. the conditions under which a bin splitting minimizes the estimator variance

Based on the above, we formulate bin-splitting as an optimization problem and propose an efficient solution using dynamic programming.

RHALE: Robust and Heterogeneity-aware ALE

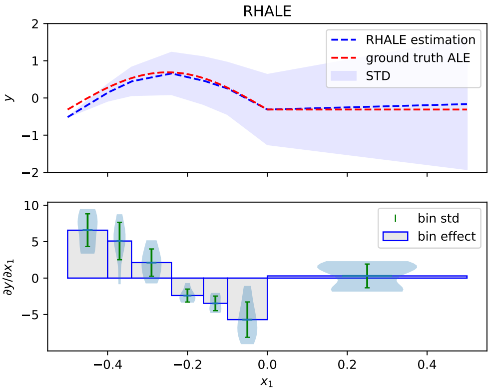


Figure: Variable bin size leads to improved estimation

Simple but correct:

- Automatically finds the **optimal** bin-splitting
- Optimal $\Rightarrow$ best approximation of the average (ALE) effect
- Optimal $\Rightarrow$ best approximation of the heterogeneity

Impact

In case you work with a differentiable model, as in Deep Learning, use the combination of DALE and RHALE to:

- compute ALE fast, for multiple bin sizes in one pass
- quantify the heterogeneity of the ALE plot, i.e., the deviation of the instance-level effects from the average effect
- get a robust approximation of (a) the main ALE effect and (b) the heterogeneity, using automatic bin-splitting

A python package will soon be released to provide these functionalities

Contents

V. Gkolemis, A. Tzerefos, T. Dalamagas, E. Ntoutsi and C. Diou, “Regionally Additive Models: Explainable-by-design models minimizing feature interactions”, Uncertainty meets Explainability workshop, ECML 2023
https://xai-uncertainty.github.io

Work in collaboration with Vasilis Gkolemis (PhD student @ HUA), Anargiros Tzerefos, Theodoros Dalamagas (Researcher, ATHENA RC) and Eirini Ntoutsi (Prof, Universität der Bundeswehr, München)




After DALE & RHALE: Regional effects

- Similar to the way one can select optimal bin splits to minimize heterogeneity, one can also identify optimal subregions of the features $\boldsymbol{x}_c$ where the effect is homogeneous
- Work in progress (also work by others)
- Also part of the soon-to-be-released python package
- In this final, brief part we will discuss something a little different, based on the same idea

Generalized Additive Models (GAMs)

Wikipedia says:

> *In statistics, a generalized additive model (GAM) is a generalized linear model in which the response variable depends linearly on unknown smooth functions of some predictor variables.*

Generalized Additive Models (GAMs)

Wikipedia says:

> *In statistics, a generalized additive model (GAM) is a generalized linear model in which the response variable depends linearly on unknown smooth functions of some predictor variables.*

$$y$$

Generalized Additive Models (GAMs)

Wikipedia says:

> *In statistics, a generalized additive model (GAM) is a generalized linear model in which the response variable depends linearly on unknown smooth functions of some predictor variables.*

$$y = \cdot + \ldots + \cdot$$

Generalized Additive Models (GAMs)

Wikipedia says:

> *In statistics, a generalized additive model (GAM) is a generalized linear model in which the response variable depends linearly on unknown smooth functions of some predictor variables.*

$$y = f_1(x_1) + \ldots + f_D(x_D)$$

Introductory Example

Output/target variable:

- $y_{\texttt{bike-rentals}}$: the expected number of bike rentals per hour

Input/covariates:

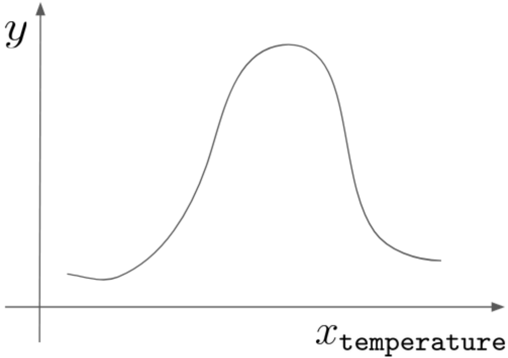
- $x_{\texttt{temperature}}$: temperature per hour
- $x_{\texttt{humidity}}$: humidity per hour
- $x_{\texttt{is\_weekday}}$: if it is weekday or weekend

Let's fit a GAM:

$$y = f_1(x_{\texttt{temperature}}) + f_2(x_{\texttt{humidity}}) + f_3(x_{\texttt{is\_weekday}})$$

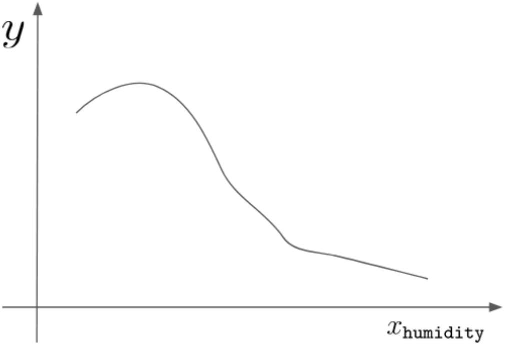
GAMs - Interpretability (1)

$f_1(x_{\texttt{temperature}})$



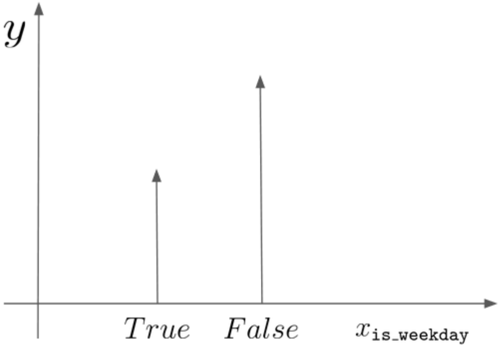
GAMs - Interpretability (2)

$f(x_{\texttt{humidity}})$

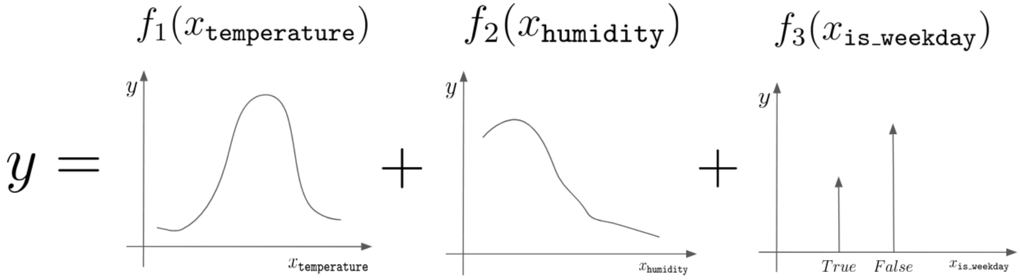


GAMs - Interpretability (3)

$f(x_{\texttt{is\_weekday}})$



GAMs is explainable!



GAMs - Limitations/Extensions

Limitations:

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\texttt{temperature}}, x_{\texttt{is\_weekday}})$

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\texttt{temperature}}, x_{\texttt{is\_weekday}})$
- Solution 2: Model two conditional terms
  - $f(x_{\texttt{temperature}} | weekday)$
  - $f(x_{\texttt{temperature}} | weekend)$

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\texttt{temperature}}, x_{\texttt{is\_weekday}})$
- Solution 2: Model two conditional terms
  - $f(x_{\texttt{temperature}}|weekday)$
  - $f(x_{\texttt{temperature}}|weekend)$

Extensions:

- Solution 1: $GA^2M =$ GAM + pairwise interactions (Yin Lou et. al)

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\texttt{temperature}}, x_{\texttt{is\_weekday}})$
- Solution 2: Model two conditional terms
  - $f(x_{\texttt{temperature}}|weekday)$
  - $f(x_{\texttt{temperature}}|weekend)$

Extensions:

- Solution 1: $GA^2M =$ GAM + pairwise interactions (Yin Lou et. al)
- Solution 2: $RAM =$ GAM at subregions

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\texttt{temperature}}, x_{\texttt{is\_weekday}})$ Explainable
- Solution 2: Model two conditional terms
  - $f(x_{\texttt{temperature}}|weekday)$ Explainable
  - $f(x_{\texttt{temperature}}|weekend)$ Explainable

Extensions:

- Solution 1: $GA^2M =$ GAM + pairwise interactions (Yin Lou et. al)
- Solution 2: $RAM =$ GAM at subregions

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

$RA^{(2)}Ms$ solve that:

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?

$RA^{(2)}Ms$ solve that:

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!

$RA^{(2)}Ms$ solve that:

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it is a workday? and bike is the only transport?

$RA^{(2)}Ms$ solve that:

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it is a workday? and bike is the only transport?
- model $f(x_{\texttt{temperature}}, x_{\texttt{humidity}}, x_{\texttt{is\_weekday}})$?

$RA^{(2)}Ms$ solve that:

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it is a workday? and bike is the only transport?
- model $f(x_{\texttt{temperature}}, x_{\texttt{humidity}}, x_{\texttt{is\_weekday}})$? Not explainable

$RA^{(2)}Ms$ solve that:

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it is a workday? and bike is the only transport?
- model $f(x_{\texttt{temperature}}, x_{\texttt{humidity}}, x_{\texttt{is\_weekday}})$? Not explainable

$RA^{(2)}Ms$ solve that:

- $f(x_{\texttt{temperature}}, x_{\texttt{humidity}} | x_{\texttt{is\_weekday}}) \rightarrow RA^2M$

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it is a workday? and bike is the only transport?
- model $f(x_{\texttt{temperature}}, x_{\texttt{humidity}}, x_{\texttt{is\_weekday}})$? Not explainable

$RA^{(2)}Ms$ solve that:

- $f(x_{\texttt{temperature}}, x_{\texttt{humidity}} | x_{\texttt{is\_weekday}}) \rightarrow RA^2M$
- $f(x_{\texttt{temperature}} | x_{\texttt{humidity}} = \{high, low\}, x_{\texttt{is\_weekday}}) \rightarrow$ RAM with two conditions

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it is a workday? and bike is the only transport?
- model $f(x_{\texttt{temperature}}, x_{\texttt{humidity}}, x_{\texttt{is\_weekday}})$? Not explainable

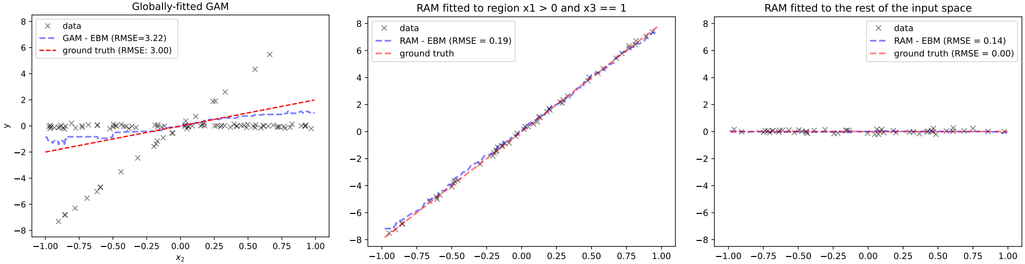
$RA^{(2)}Ms$ solve that:

- $f(x_{\texttt{temperature}}, x_{\texttt{humidity}} | x_{\texttt{is\_weekday}}) \rightarrow RA^2M$ Explainable
- $f(x_{\texttt{temperature}} | x_{\texttt{humidity}} = \{high, low\}, x_{\texttt{is\_weekday}}) \rightarrow$ RAM with two conditions Explainable

RAM on toy example

$$f(\boldsymbol{x}) = 8x_2 \mathbb{1}_{x_1>0} \mathbb{1}_{x_3=0}$$

$$x_1, x_2 \sim \mathcal{U}(-1, 1), x_3 \sim Bernoulli(0, 1)$$



(a) $f_2(x_2)$ (b) $f_2(x_2)\mathbb{1}_{x_1>0 \text{ and } x_3=1}$ (c) $f_2(x_2)\mathbb{1}_{x_1\leq 0 \text{ or } x_3\neq 1}$

Figure: (Left) GAM, (Middle and Right) RAM

How RAM works

3-step approach:

How RAM works

3-step approach:

- Fit a black-box model to learn complex feature interactions
  - it should be differentiable
  - neural network is a good option

How RAM works

3-step approach:

- Fit a black-box model to learn complex feature interactions
    - it should be differentiable
    - neural network is a good option
- Use a Regional Effect method to isolate the important interactions
    - RHALE
    - Feature Interactions - Herbinger et. al
    - finds which features $f(x_i)$ should be split into subregions $f(x_i | x_j \lessgtr \tau)$

How RAM works

3-step approach:

- Fit a black-box model to learn complex feature interactions
  - it should be differentiable
  - neural network is a good option
- Use a Regional Effect method to isolate the important interactions
  - RHALE
  - Feature Interactions - Herbinger et. al
  - finds which features $f(x_i)$ should be split into subregions $f(x_i|x_j \lessgtr \tau)$
- Fit a univariate function on each detected subregion
  - learn all $f(x_i|x_j \lessgtr \tau)$

Step 1

- Fit a black-box model to capture all complex structures
  - it should be differentiable
  - A neural network is a good option

Step 2

- Regional Effect method to find important interactions
  - RHALE
  - Feature Interactions - Herbinger et. al
- Idea:
  - Feature effect is the average effect of each feature x_s on the output y
  - It is computed by averaging the instance-level effects
  - Heterogeneity $\mathcal{H}$ (or uncertainty) measures the deviation of the instance-level effects from the average effect
  - we want to split the dataset in subgroups in order to minimize the heterogeneity
- In mathematical terms:

$$\underbrace{\mathcal{H}(f_i(x_i))}_{\mathcal{H}\text{ before split}} >> \underbrace{\mathcal{H}(f_i(x_i|x_j > \tau)) + \mathcal{H}(f_i(x_i|x_j \leq \tau))}_{\text{sum of }\mathcal{H}\text{ after split}}$$

Step 3

- Step 2 defines a new feature space $\mathcal{X}^{\texttt{RAM}}$
- Every feature is split to T_s subregions which are defined by $\mathcal{R}_{st}$:

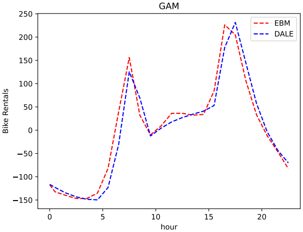
$$
\begin{aligned}
\mathcal{X}^{\texttt{RAM}} &= \{x_{st} | s \in \{1, \ldots, D\}, t \in \{1, \ldots, T_s\}\} \\
x_{st} &= \begin{cases} x_s, & \text{if } \mathbf{x}_{/\mathbf{s}} \in \mathcal{R}_{st} \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{1}
$$

- Fit a univariate function on each subregion:

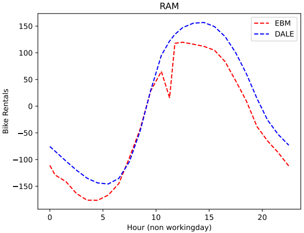
$$
f^{\texttt{RAM}}(\mathbf{x}) = c + \sum_{s,t} f_{st}(x_{st}) \quad \mathbf{x} \in \mathcal{X}^{\texttt{RAM}} \tag{2}
$$

Bike Sharing dataset

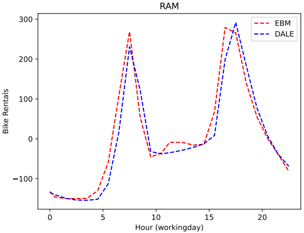
Predict bike-rentals per hour



(a) $f(X_{\texttt{hour}})$



(b) $f(X_{\texttt{hour}})\mathbb{1}_{X_{\texttt{workingday}} \neq 1}$



(c) $f(X_{\texttt{hour}})\mathbb{1}_{X_{\texttt{workingday}} = 1}$

Experimental Results

Tested on Bike Sharing and California Housing Datasets.

	Black-box	**x-by-design**			
	all orders	1st order		2nd order	
	DNN	**GAM**	**RAM**	**GA^{2}M**	**RA^{2}M**
Bike (MAE)	0.254	0.549	**0.430**	0.298	**0.278**
Bike (RMSE)	0.389	0.734	**0.563**	0.438	**0.412**
Housing (MAE)	0.373	0.600	**0.553**	0.554	**0.533**
Housing (RMSE)	0.533	0.819	**0.754**	0.774	**0.739**

What is next?

- Results are preliminary
  - Compare RAM vs GAM and RA^2M vs GA^2M in more datasets
  - Check robustness on edge cases:
    - highly correlated features
    - limited training examples
- Can we model uncertainty?
  - Uncertain because we do not model higher-order interactions
  - Uncertain about the conditionals, i.e., detected subregions
  - Uncertain about the univariate functions we learn
- Could we make it a 1-step process?
  - a network that automatically learns both the univariate functions and the conditions

Recap

- DALE can help with the computation of fast and accurate feature effect explanations for differentiable models
  - One can change the resolution of the explanation (i.e., number of bins K) for free
- RHALE can improve explanations by selecting variable bin splits, in an optimal way
  - Unbiased estimation of heterogeneity
  - Select optimal bin splits to minimize heterogeneity and improve the robustness of the explanation
- Regionally Additive Models have the potential to improve the model accuracy, while maintaining explainability
  - Selection of optimal feature space subregions and fit a GAM
  - Preliminary work, a lot to be done

Thank you!

References I

[1] Daniel W. Apley and Jingyu Zhu. “Visualizing the effects of predictor variables in black box supervised learning models”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.4 (2020), pp. 1059–1086. ISSN: 14679868. DOI: 10.1111/rssb.12377. arXiv: 1612.08468.

[2] Hadi Fanaee-T and Joao Gama. “Event labeling combining ensemble detectors and background knowledge”. In: *Progress in Artificial Intelligence* (2013), pp. 1–15. ISSN: 2192-6352. DOI: 10.1007/s13748-013-0040-3. URL: [WebLink].

[3] Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine”. In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232. ISSN: 00905364. DOI: 10.1214/aos/1013203451.

[4] Vasilis Gkolemis, Theodore Dalamagas, and Christos Diou. “DALE: Differential Accumulated Local Effects for efficient and accurate global explanations”. In: *Asian Conference on Machine Learning*. PMLR. 2023, pp. 375–390.

References II

[5] Vasilis Gkolemis et al. "RHALE: Robust and heterogeneity-aware accumulated local effects". In: (2023).

[6] Alex Goldstein et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". In: *journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65.

[7] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[8] Timo Speith. "A review of taxonomies of explainable artificial intelligence (XAI) methods". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 2239–2250.