Regionally Additive Models: Explainable-by-design models minimizing feature interactions

Vasilis Gkolemis[1,2] Anargiros Tzerefos[1] Theodore Dalamagas[1]
Eirini Ntoutsi[3] Christos Diou[2]

[1]ATHENA Research and Innovation Center

[2]Harokopio University of Athens

[3]Universitat der Bundeswehr Munchen

November 2023 @ IML-LMU

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:

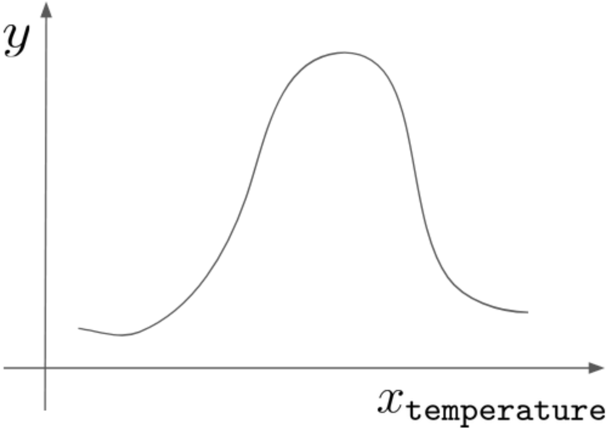
- $x_{\texttt{temperature}}$: temperature per hour
- $x_{\texttt{humidity}}$: humidity per hour
- $x_{\texttt{is\_weekday}}$: if it is weekday or weekend

Let's fit a GAM:

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

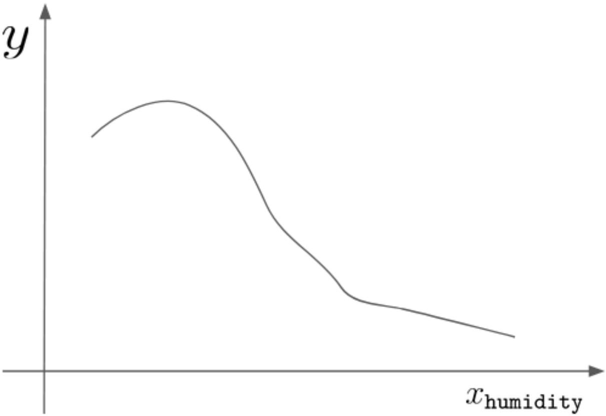
GAMs - Interpretability (1)

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

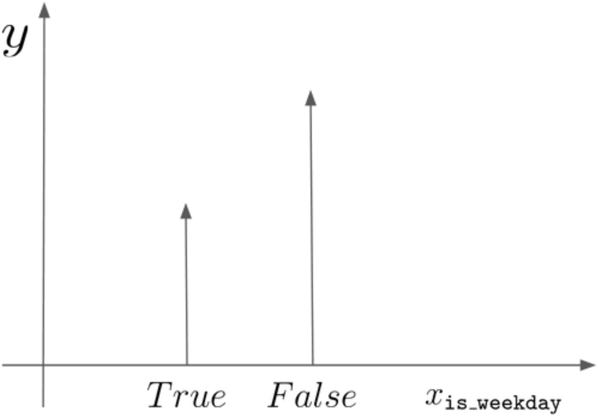


GAMs - Interpretability (2)

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

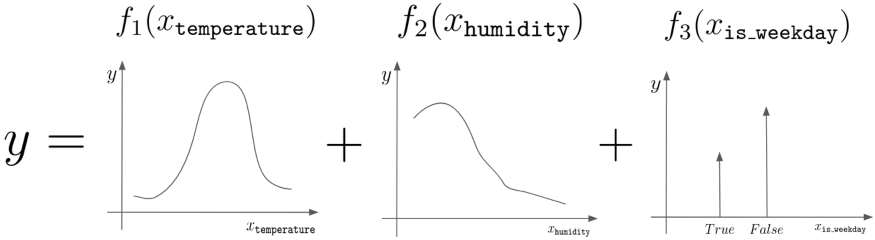


$f(x_{\texttt{is\_weekday}})$



GAMs - Interpretability (4)

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 $f(x_{\mathtt{temperature}}, x_{\mathtt{is\_weekday}})$
- Solution 2: Model two conditional terms
  - $f(x_{\mathtt{temperature}} | weekday)$
  - $f(x_{\mathtt{temperature}} | weekend)$

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:

- cold day with humidity $>>$ cold $+$ humidity!

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

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

GA^2Ms Limitations:

- cold day with humidity $>>$ cold $+$ humidity!
- If it is weekend, you can stay at home!

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

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

GA^2Ms Limitations:

- cold day with humidity $>>$ cold $+$ humidity!
- If it is weekend, you can stay at home!
- But if it workday? and bike is the only transport?

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

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

GA^2Ms Limitations:

- cold day with humidity $>>$ cold $+$ humidity!
- If it is weekend, you can stay at home!
- But if it 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:

- cold day with humidity $>>$ cold $+$ humidity!
- If it is weekend, you can stay at home!
- But if it 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:

- cold day with humidity $>>$ cold $+$ humidity!
- If it is weekend, you can stay at home!
- But if it 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:

- $f(x_{\texttt{temperature}}, x_{\texttt{humidity}} | x_{\texttt{is\_weekday}} = \{T, F\}) \rightarrow RA^2M : 2 \times 2D$-plots

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

GA^2Ms Limitations:

- cold day with humidity $>>$ cold $+$ humidity!
- If it is weekend, you can stay at home!
- But if it 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:

- $f(x_{\texttt{temperature}}, x_{\texttt{humidity}} | x_{\texttt{is\_weekday}} = \{T, F\}) \rightarrow RA^2M : 2 \times 2D$-plots
- $f(x_{\texttt{temperature}} | x_{\texttt{humidity}} = \{high, low\}, x_{\texttt{is\_weekday}} = \{T, F\}) \rightarrow \text{RAM}$ with two conditions: $4 \times 2D$ plots

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

GA^2Ms Limitations:

- cold day with humidity $>>$ cold + humidity!
- If it is weekend, you can stay at home!
- But if it workday? and bike is the only transport?
- model $f(x_{\texttt{temperature}}, x_{\texttt{humidity}}, x_{\texttt{is\_weekday}})$? Not explainable

---

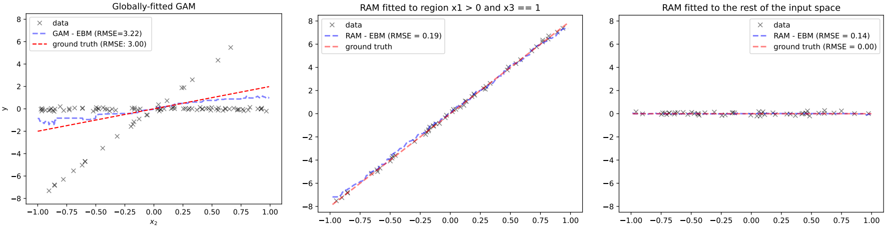
$RA^{(2)}Ms$ solve that:

- $f(x_{\texttt{temperature}}, x_{\texttt{humidity}} | x_{\texttt{is\_weekday}} = \{T, F\}) \rightarrow RA^2M : 2 \times 2D$-plots Explainable
- $f(x_{\texttt{temperature}} | x_{\texttt{humidity}} = \{high, low\}, x_{\texttt{is\_weekday}} = \{T, F\}) \rightarrow \text{RAM}$ with two conditions: $4 \times 2D$ plots Explainable

RAM on toy example

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

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



(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 - Gkolemis et. al
  - Feature Interactions - Herbinger et. al
  - finds which features $f(x_i)$ should be split into subregions $f(x_i | x_j \lesseqgtr \tau)$

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 - Gkolemis et. al
  - Feature Interactions - Herbinger et. al
  - finds which features $f(x_i)$ should be split into subregions $f(x_i | x_j \lesseqgtr \tau)$
- Fit a univariate function on each detected subregion
  - learn all $f(x_i | x_j \lesseqgtr \tau)$

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 - Gkolemis et. al
  - 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 $\mathcal{H}$ measures the deviation of the instance-level effects from the average effect due to feature interactions
  - 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 $\mathcal{R}_{st}$:

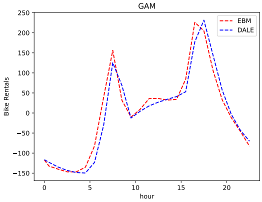
$$\mathcal{X}^{\texttt{RAM}} = \{x_{st} | s \in \{1, \ldots, D\}, t \in \{1, \ldots, T_s\}\}$$
$$x_{st} = \begin{cases} x_s, & \text{if } \mathbf{x}_{/\mathbf{s}} \in \mathcal{R}_{st} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

- Fit a univariate function on each subregion:

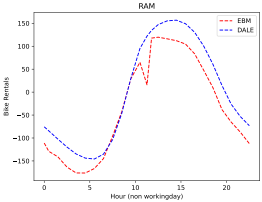
$$f^{\texttt{RAM}}(\mathbf{x}) = c + \sum_{s,t} f_{st}(x_{st}) \quad \mathbf{x} \in \mathcal{X}^{\texttt{RAM}} \tag{2}$$

Bike Sharing dataset

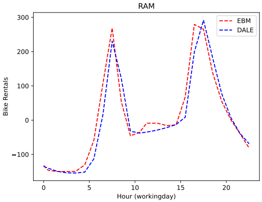
Predict bike-rentals per hour



(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	1^{st} order		2^{nd} 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**

Conclusions

- RAM $\Rightarrow$ Use XAI to understand something and **take some action** given **some goal**
  - What is the goal?
  - What to understand?
  - What is the action?

Conclusions

- RAM $\Rightarrow$ Use XAI to understand something and **take some action** given **some goal**
  - What is the goal?
    - Big picture: Fit an accurate x-by-design model
    - Small picture: Fit multiple GAMs in subregions of the input space
  - What to understand?
  - What is the action?

Conclusions

- RAM ⇒ Use XAI to understand something and **take some action** given **some goal**
  - What is the goal?
    - Big picture: Fit an accurate x-by-design model
    - Small picture: Fit multiple GAMs in subregions of the input space
  - What to understand?
    - Which subregions minimize feature interactions?
  - What is the action?

Conclusions

- RAM $\Rightarrow$ Use XAI to understand something and **take some action** given **some goal**
  - What is the goal?
    - Big picture: Fit an accurate x-by-design model
    - Small picture: Fit multiple GAMs in subregions of the input space
  - What to understand?
    - Which subregions minimize feature interactions?
  - What is the action?
    - Split the input space in these subregions and fit a GAM in each subregion

What is next?

- Results are preliminary
  - Compare *RAM* vs *GAM* and RA^2M vs GA^2M in more datasets
  - Compare with novel works, papers are coming each year
  - Check robustness on edge cases:
    - highly correlated features
    - limited training examples
    - which regional effect method works better?
- Theoretical work - Can we prove it?
  - minimizing the heterogeneity maximizes the accuracy of a RAM model?
  - under which conditions?
- Can we model the uncertainty?
  - Uncertainty due to not modeling higher-order interactions
  - Uncertain about the conditionals, i.e., the 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

Thank you for your attention

- For more discussion or future ideas on RAM, please, contact me:
  - vgkolemis@athenarc.gr
  - gkolemis@hua.gr
- More info about the paper: https://arxiv.org/abs/2309.12215
- Questions?