Fast Generic Interaction Detection for Model Interpretability and Compression

ICLR 2022 (Poster)

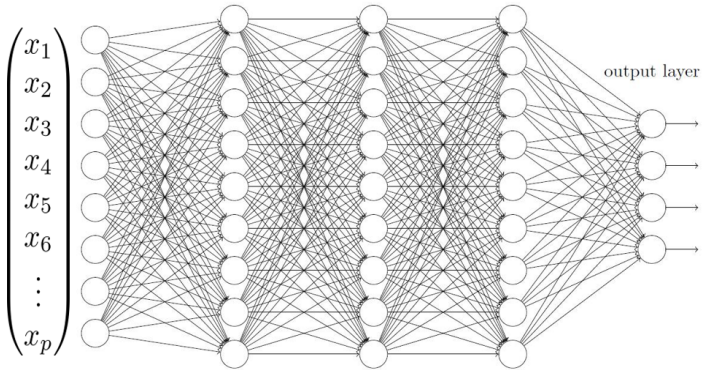
Tianjian Zhang, Feng Yin, Zhi-Quan Luo

The Chinese University of Hong Kong, Shenzhen
Shenzhen Research Institute of Big Data

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

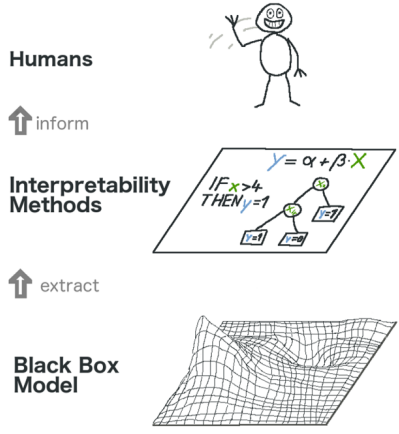
深圳市大数据研究院
Shenzhen Research Institute of Big Data

Black-box model's limitation



- Current black box ML models are often heavily parameterized and hard to interpret.

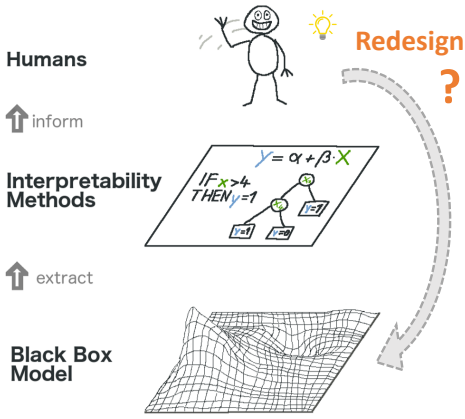
Interpretation matters



Credit to Christoph Molnar

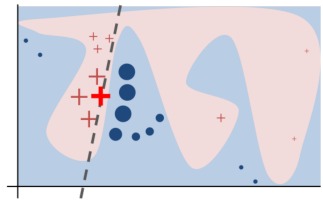
Can we improve the model?

(Model performance, interpretability, model size, etc.)

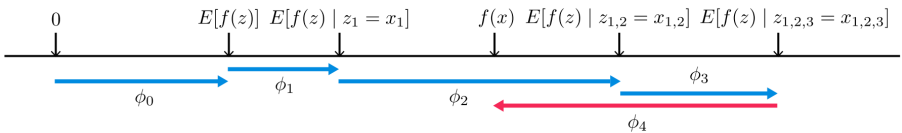


Additive attribution: LIME and SHAP

- Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al. (2016)]



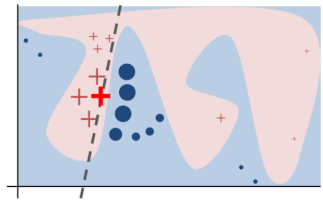
- SHapley Additive exPlanations (SHAP) [Lundberg and Lee (2017)]



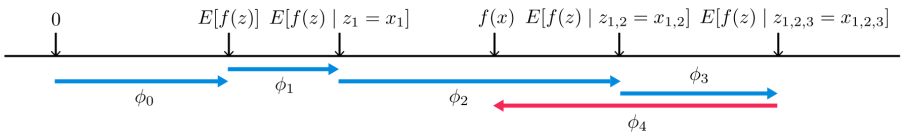
- Nonlinear interpretations are missing.

Additive attribution: LIME and SHAP

- Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al. (2016)]



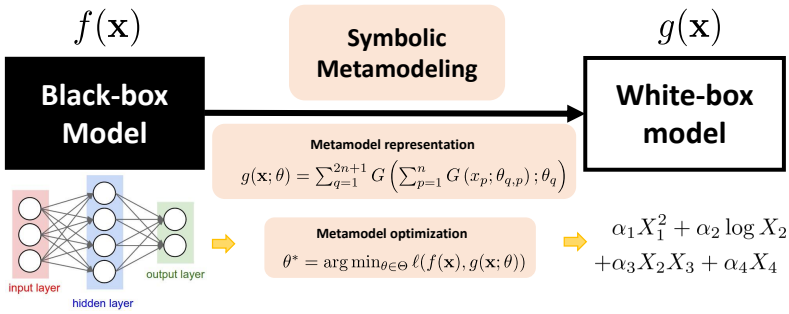
- SHapley Additive exPlanations (SHAP) [Lundberg and Lee (2017)]



- Nonlinear interpretations are missing.

Metamodel

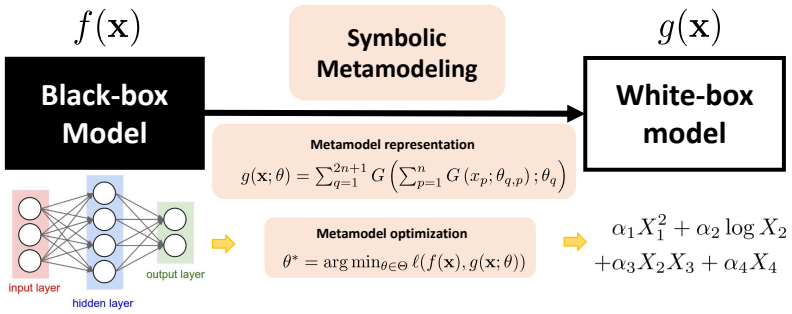
[Alaa and van der Schaar (2019)]



- $g(\mathbf{x})$ can approximate $f(\mathbf{x})$ well, and capture nonlinear interaction, e.g. $X_2 X_3$.
- But the performance of g cannot surpass f.

Metamodel

[Alaa and van der Schaar (2019)]



- $g(\mathbf{x})$ can approximate $f(\mathbf{x})$ well, and capture nonlinear interaction, e.g. $X_2 X_3$.
- But the performance of g cannot surpass f.

Interaction detection

- SOTA methods
  - Neural Interaction Detection (NID) [Tsang et al. (2018b)]
  - Persistence Interaction Detection (PID) [Liu et al. (2020)]

    Heuristic and not general: Restricted to special NN structures, e.g. a sparse ReLU network.
- Other methods
  - Integrated Hessian [Janizek et al. (2020)]
  - Shapley Interactions [Sundararajan et al. (2020)]
  - and more...

    Computationally heavy.
- **Our contribution:**
  - A fast and generic interaction detection method.
  - High performance, lightweight, interpretable Parametric ACE model (ParaACE)

Interaction detection

- SOTA methods
  - Neural Interaction Detection (NID) [Tsang et al. (2018b)]
  - Persistence Interaction Detection (PID) [Liu et al. (2020)]
    Heuristic and not general: Restricted to special NN structures, e.g. a sparse ReLU network.
- Other methods
  - Integrated Hessian [Janizek et al. (2020)]
  - Shapley Interactions [Sundararajan et al. (2020)]
  - and more...
    Computationally heavy.
- **Our contribution:**
  - A fast and generic interaction detection method.
  - High performance, lightweight, interpretable Parametric ACE model (ParaACE)

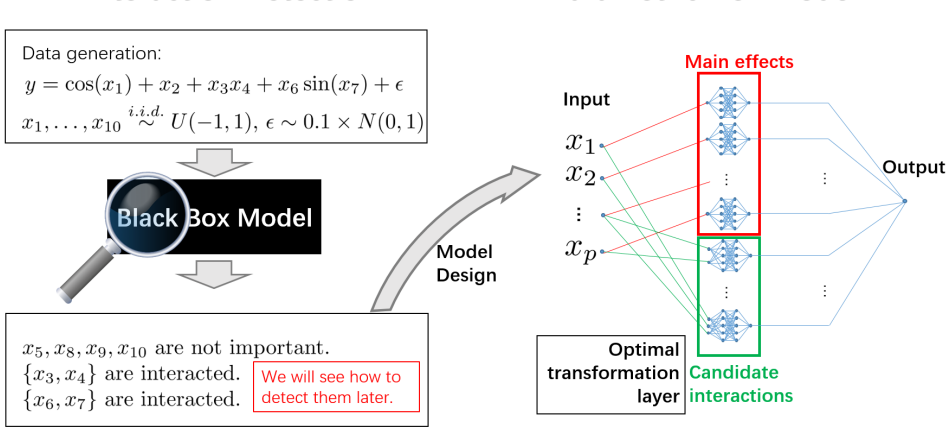
Interaction detection

- SOTA methods
  - Neural Interaction Detection (NID) [Tsang et al. (2018b)]
  - Persistence Interaction Detection (PID) [Liu et al. (2020)]
    Heuristic and not general: Restricted to special NN structures, e.g. a sparse ReLU network.
- Other methods
  - Integrated Hessian [Janizek et al. (2020)]
  - Shapley Interactions [Sundararajan et al. (2020)]
  - and more...
    Computationally heavy.
- **Our contribution:**
  - A fast and generic interaction detection method.
  - High performance, lightweight, interpretable Parametric ACE model (ParaACE)

Interaction detection

- SOTA methods
  - Neural Interaction Detection (NID) [Tsang et al. (2018b)]
  - Persistence Interaction Detection (PID) [Liu et al. (2020)]
    Heuristic and not general: Restricted to special NN structures, e.g. a sparse ReLU network.
- Other methods
  - Integrated Hessian [Janizek et al. (2020)]
  - Shapley Interactions [Sundararajan et al. (2020)]
  - and more...
    Computationally heavy.
- **Our contribution:**
  - A fast and generic interaction detection method.
  - High performance, lightweight, interpretable Parametric ACE model (ParaACE)

Overview of the Framework



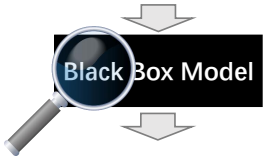
Overview of the Framework

Interaction Detection

Data generation:

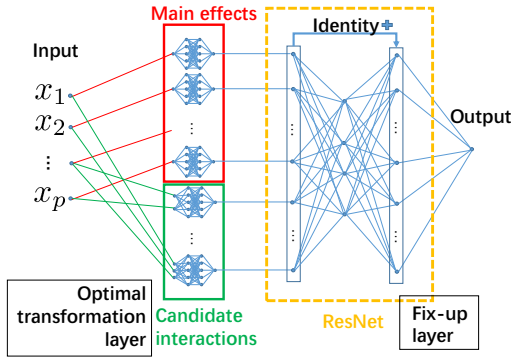
$$y = \cos(x_1) + x_2 + x_3 x_4 + x_6 \sin(x_7) + \epsilon$$

$$x_1, \ldots, x_{10} \overset{i.i.d.}{\sim} U(-1, 1),\ \epsilon \sim 0.1 \times N(0, 1)$$

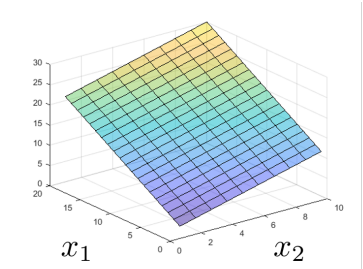


x_5, x_8, x_9, x_{10} are not important.
$\{x_3, x_4\}$ are interacted.
$\{x_6, x_7\}$ are interacted.

Parametric ACE model



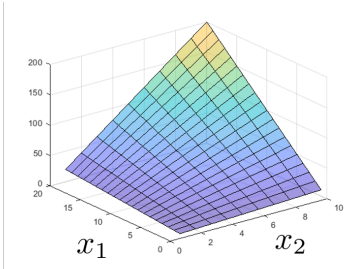
What is interaction?



$$F(x_1, x_2) = x_1 + x_2$$

$$F(x_1, x_2) = h(x_1) + g(x_2)$$

No Interaction



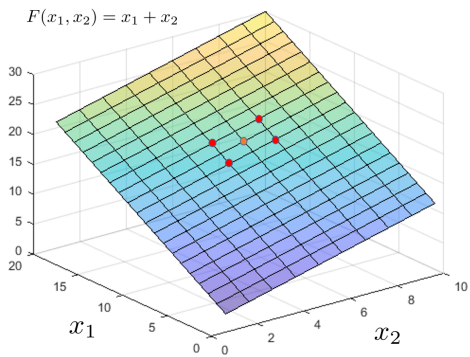
$$F(x_1, x_2) = x_1 \times x_2$$

$$F(x_1, x_2) \neq h(x_1) + g(x_2)$$

Interacted

$$\frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} = 0 \quad \text{or} \quad \neq 0?$$

How to detect interactions?



- Pick a local point (x_1, x_2)

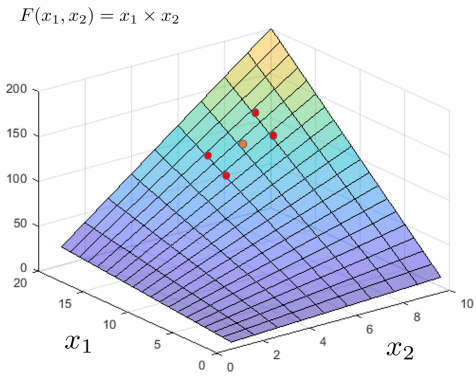
$$
\begin{aligned}
&F(x_1 + h, x_2 + h) \\
-&F(x_1 + h, x_2 - h) \\
-&F(x_1 - h, x_2 + h) \\
+&F(x_1 - h, x_2 - h)
\end{aligned}
$$

$$
\shortparallel
$$

$$
0
$$

- No interaction locally at (x_1, x_2)

How to detect interactions?



- Pick a local point (x_1, x_2)

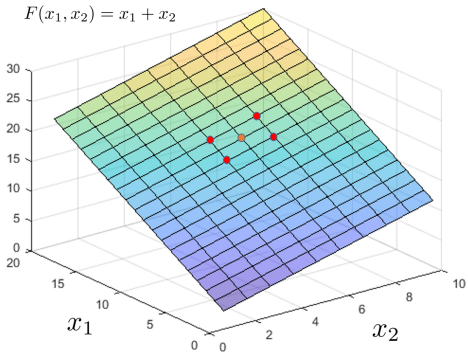
$$
\begin{aligned}
&F(x_1 + h, x_2 + h) \\
-&F(x_1 + h, x_2 - h) \\
-&F(x_1 - h, x_2 + h) \\
+&F(x_1 - h, x_2 - h)
\end{aligned}
$$

$$\not{\mathbf{H}}$$

$$\mathbf{0}$$

- Interacted locally at (x_1, x_2)

How to detect interactions?



- Average over all N data points

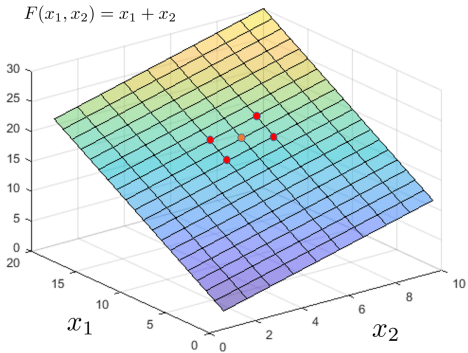
$$\mathbb{E}_{X_1,X_2} \begin{bmatrix} F(X_1+h, X_2+h) \\ -F(X_1+h, X_2-h) \\ -F(X_1-h, X_2+h) \\ +F(X_1-h, X_2-h) \end{bmatrix}^2$$

$$\Downarrow$$

$$\mathbf{0}$$

- No interaction globally
- How to choose h ?
  (a small h?)

How to detect interactions?



- Average over all N data points

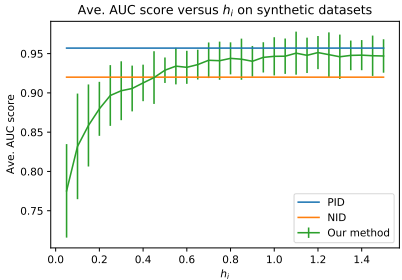
$$\mathbb{E}_{X_1,X_2} \begin{bmatrix} F(X_1+h, X_2+h) \\ -F(X_1+h, X_2-h) \\ -F(X_1-h, X_2+h) \\ +F(X_1-h, X_2-h) \end{bmatrix}^2$$

$$\Downarrow$$

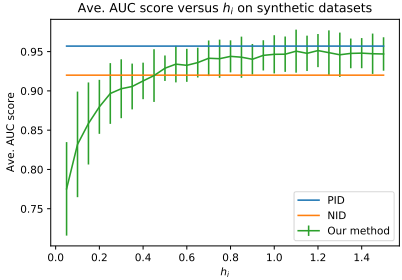
$$0$$

- No interaction globally
- How to choose h ?
  (a small h?)

Surprisingly, large h works!



Surprisingly, large h works!



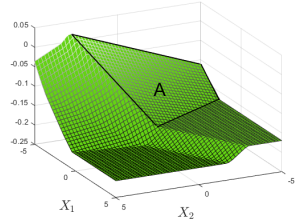
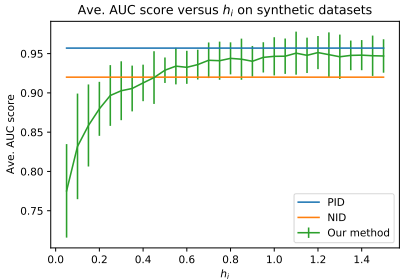


Figure: Landscape of ReLU networks

Surprisingly, large h works!



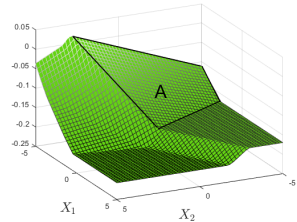
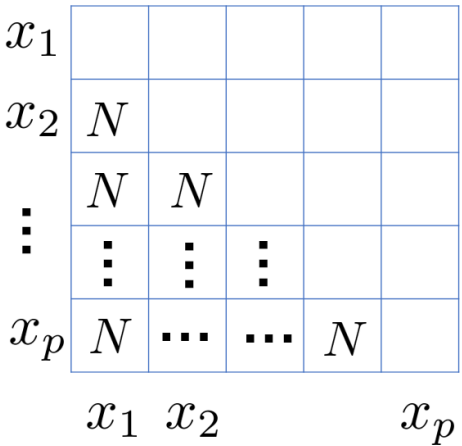


Figure: Landscape of ReLU networks

Theorem 1.
For any x and y, function F shows no interaction between x and y, i.e. $F(x,y) = h(x) + g(y)$ **iff**, *for any $h, k > 0$, $F(x+h, y+k) - F(x+h, y-k) - F(x-h, y+k) + F(x-h, y-k) = 0$.*

Accelerating via UCB algorithm

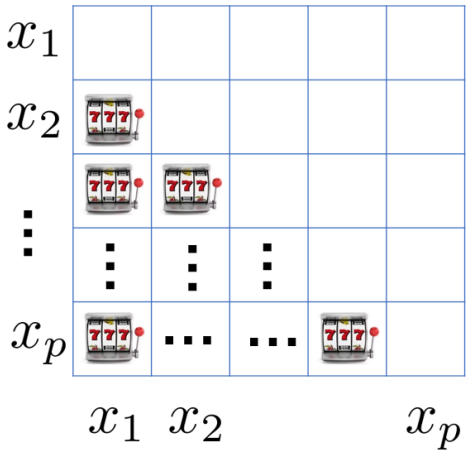


- Consider $f(x_1, \ldots, x_p)$.
- The # data samples is N.
- # **total evaluations:**

$$4Np(p-1)/2,$$

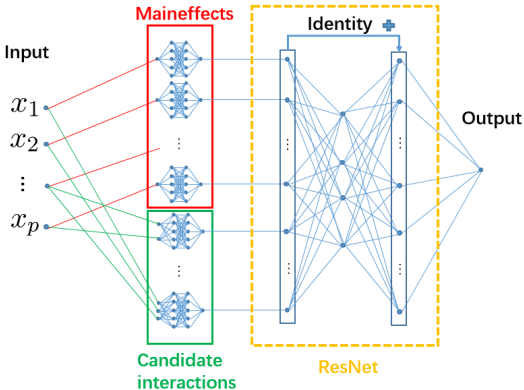
which can be largely reduced.

Accelerating via UCB algorithm

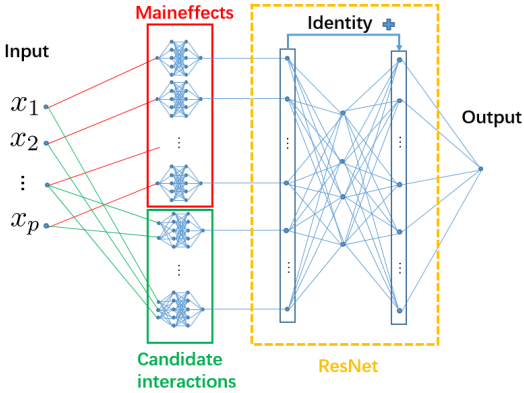


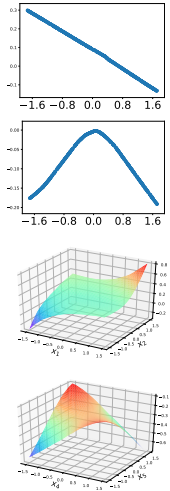
- To find top k-strongest interactions, # **evaluations** is $\mathcal{O}(p^2 \log(p^2 N) + kN)$.
- previous: $\mathcal{O}(p^2 N)$

Parametric ACE



Parametric ACE



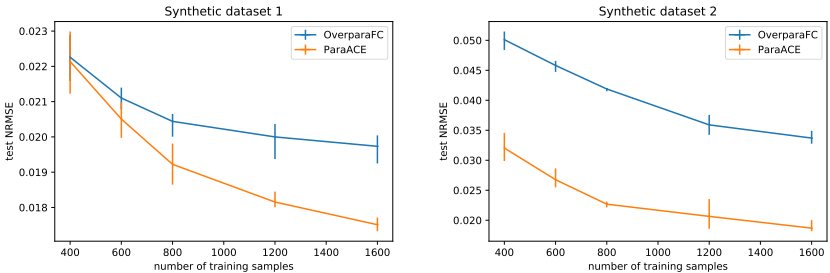


Performance gain and model compression

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	Average	CR
OverparaFC	0.026	0.054	0.059	0.062	0.042	0.041	0.018	0.024	0.032	0.023	0.038	1
KD (student)	0.029	0.055	0.061	0.064	0.041	0.042	0.019	0.024	0.032	0.024	0.039	278
LTH	0.027	0.044	0.032	0.032	0.039	0.033	0.018	0.021	0.029	0.025	0.030	18
SynFlow	0.025	0.048	0.034	0.035	0.039	0.031	**0.015**	**0.018**	0.026	0.022	0.029	280
ParaACE	**0.025**	**0.035**	**0.031**	**0.030**	**0.038**	**0.025**	0.016	0.019	**0.023**	**0.021**	**0.026**	**283**

			OverparaFC		LTH		SynFlow		**ParaACE**		
Datasets	N	p	NRMSE	Parameters	NRMSE	CR	NRMSE	CR	NRMSE	Parameters	CR
Elevators	16599	18	0.0483	4999461	0.0523	87	0.0479	120	**0.0475**	39848	**125**
Parkinsons	5875	20	0.0251	5009461	**0.0180**	87	0.0229	120	0.0204	40946	**122**
Skillcraft	3338	19	0.0937	5004461	0.1228	87	0.0968	120	**0.0929**	40397	**124**
Bike sharing	17379	15	**0.0403**	4984461	0.0404	87	0.0405	120	0.0420	38201	**130**
Cal housing	20640	8	0.1059	4949461	0.1038	87	0.1026	120	**0.1022**	16388	**302**

ParaACE is sample efficient



- To achieve the same test accuracy, ParaACE requires fewer samples.

Conclusion

Conclusion

- We propose a fast generic and model-agnostic interaction detection method.
- ParaACE model achieves improvements in predictive performance, interpretability, sample efficiency, and model size.
- There are lots of potential applications in finance, smart medicine, biology, wireless communication, etc.

Reach out to me at

- Email: tianjianzhang@link.cuhk.edu.cn
- Code: https://github.com/zhangtj1996/ParaACE

Conclusion

Conclusion

- We propose a fast generic and model-agnostic interaction detection method.
- ParaACE model achieves improvements in predictive performance, interpretability, sample efficiency, and model size.
- There are lots of potential applications in finance, smart medicine, biology, wireless communication, etc.

Reach out to me at

- Email: tianjianzhang@link.cuhk.edu.cn
- Code: https://github.com/zhangtj1996/ParaACE

Reference I

Alaa, A. M. and van der Schaar, M. (2019). Demystifying black-box models with symbolic metamodels. In Advances in Neural Information Processing Systems, pages 11304–11314.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. Journal of the American statistical Association, 80(391):580–598.

Janizek, J. D., Sturmfels, P., and Lee, S.-I. (2020). Explaining explanations: Axiomatic feature interactions for deep networks.

Liu, Z., Song, Q., Zhou, K., Wang, T. H., Shan, Y., and Hu, X. (2020). Towards interaction detection using topological analysis on neural networks. arXiv preprint arXiv:2010.13015.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems 30, pages 4765–4774. Curran Associates, Inc.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The shapley taylor interaction index. In International Conference on Machine Learning, pages 9259–9268. PMLR.

Tsang, M., Cheng, D., and Liu, Y. (2018a). Detecting statistical interactions from neural network weights. In International Conference on Learning Representations.

Tsang, M., Liu, H., Purushotham, S., Murali, P., and Liu, Y. (2018b). Neural interaction transparency (NIT): Disentangling learned interactions for improved interpretability. In NeurIPS, pages 5809–5818.