Demystifying Black-box Models with Symbolic Metamodels

Ahmed M. Alaa
Mihaela van der Schaar

NeurIPS 2019

Demystifying ANY Black-box Model

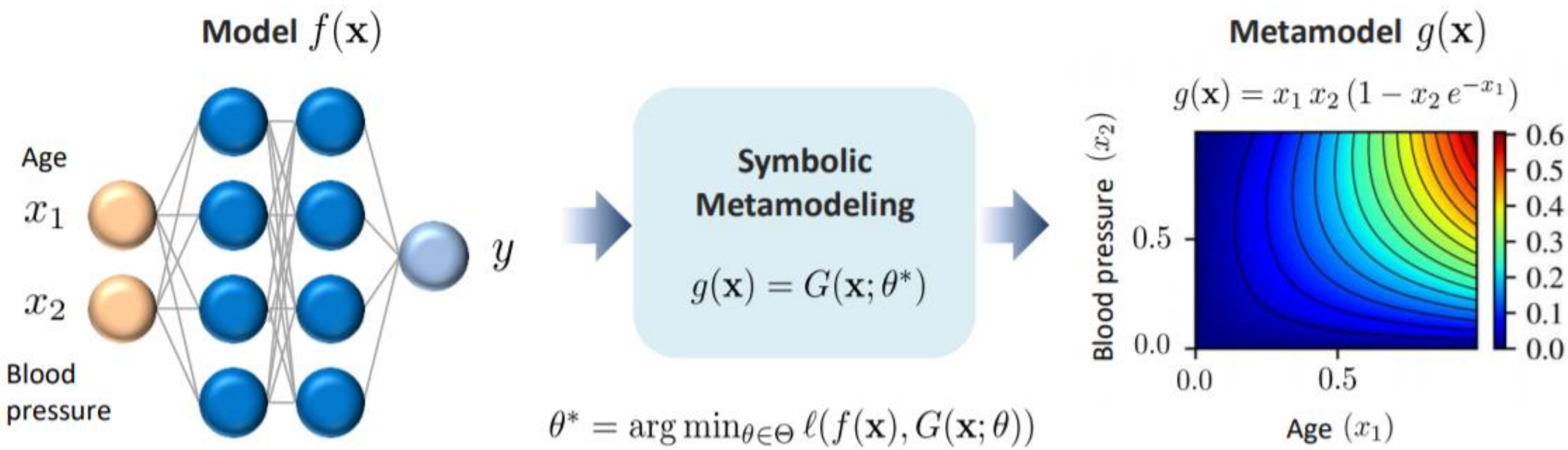
- **Our focus:** Black-box machine learning models with small to moderate number of features used in applications where the physical interpretation of features is important.

- **Key Example:** ML models for medical risk prediction…

  - Need a transparent risk equation describing the model for approval in practice guidelines.

  - Need to understand what the model discovered: feature importance, instance-wise feature importance, feature interactions, model non-linearity, etc.

Symbolic Metamodeling



- **A symbolic metamodel takes as an input a trained machine learning model and outputs a transparent mathematical equation describing the model's prediction surface.**

- **Metamodeling needs only query access to the black-box model.**

Symbolic Metamodeling

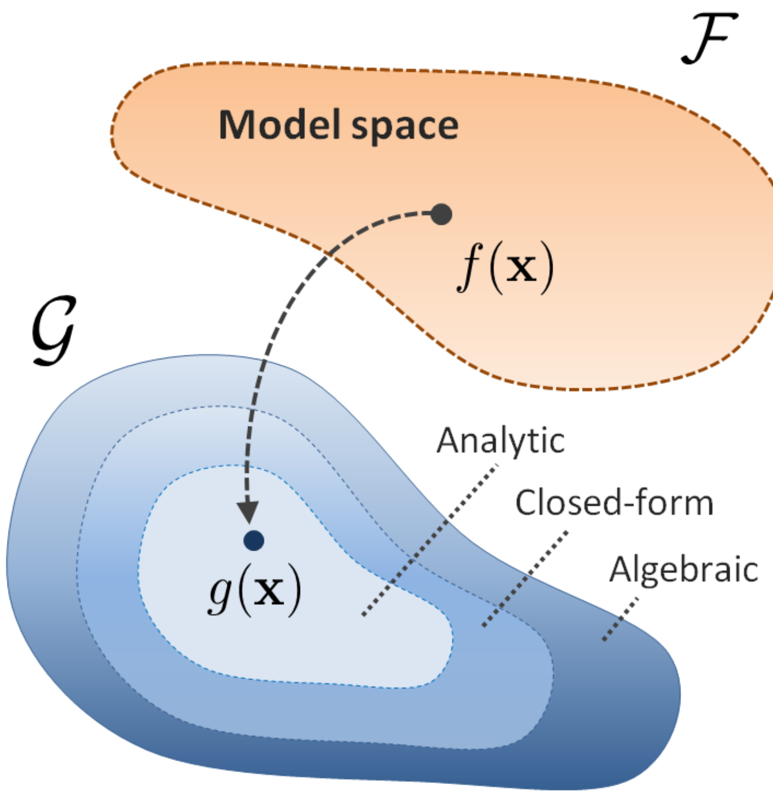
- **The symbolic metamodel problem formulation**

Model space

e.g., for neural networks: space of all uninterpretable composite functions with given activation

Metamodel space

Combination of simple functions (polynomials, closed-form or analytic expressions)



Symbolic Metamodeling

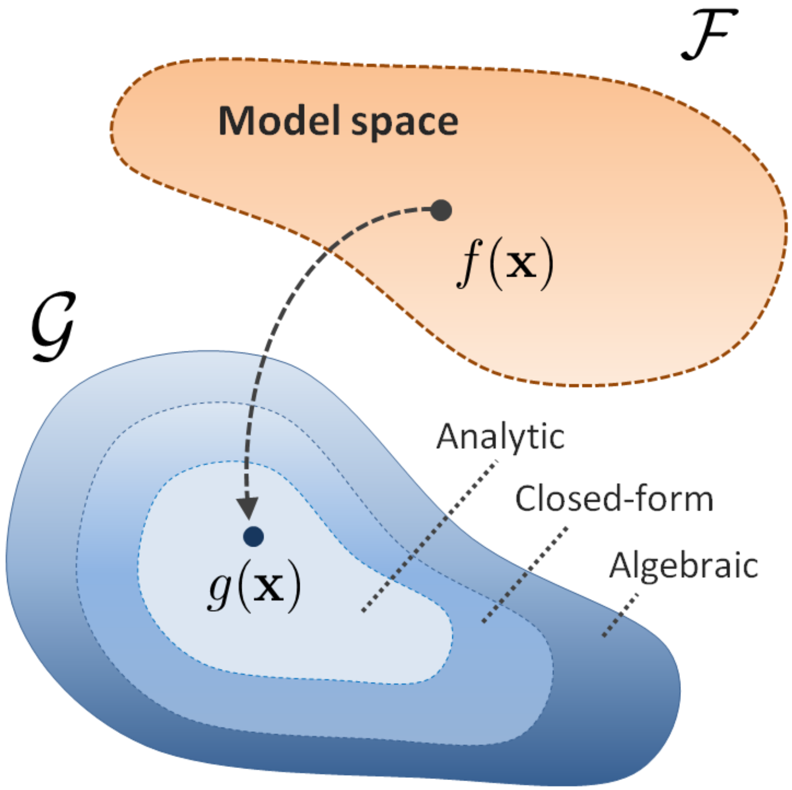
- **The symbolic metamodel problem formulation**

Metamodel optimization problem

$$g^* = \arg\min_{g \in \mathcal{G}} \ell(g, f)$$

Metamodeling loss

$$\ell(g, f) = \| f - g \|_2^2 = \int_{\mathcal{X}} (g(\mathbf{x}) - f(\mathbf{x}))^2 \, d\mathbf{x}$$



Metamodeling via Meijer-G functions

- **Parameterize the metamodeling space using two steps**

1 - Decompose the metamodel into univariate functions

$$g(\mathbf{x}) = g(x_1, \ldots, x_n) = \sum_{i=0}^{r} g_i^{out}\left(\sum_{j=1}^{d} g_{ij}^{in}(x_j)\right)$$

2 - Model basis functions via Meijer-G functions

$$G_{p,q}^{m,n}\left(\begin{smallmatrix} a_1,\ldots,a_p \\ b_1,\ldots,b_q \end{smallmatrix}\middle| x\right) = \frac{1}{2\pi i}\int_{\mathcal{L}} \frac{\prod_{j=1}^{m}\Gamma(b_j - s)\prod_{j=1}^{n}\Gamma(1 - a_j + s)}{\prod_{j=m+1}^{q}\Gamma(1 - b_j + s)\prod_{j=n+1}^{p}\Gamma(a_j + s)}\, x^s\, ds$$

What are Meijer-G functions?

- **A univariate special function given by the following line integral in the complex plane.**

$$G_{p,q}^{m,n}\left(\begin{smallmatrix} a_1,\ldots,a_p \\ b_1,\ldots,b_q \end{smallmatrix} \middle| x\right) = \frac{1}{2\pi i}\int_{\mathcal{L}} \frac{\prod_{j=1}^{m}\Gamma(b_j - s)\prod_{j=1}^{n}\Gamma(1-a_j+s)}{\prod_{j=m+1}^{q}\Gamma(1-b_j+s)\prod_{j=n+1}^{p}\Gamma(a_j+s)}\, x^s\, ds$$

- **Reduces to almost all known basic functions for different selections of the poles and zeros.**

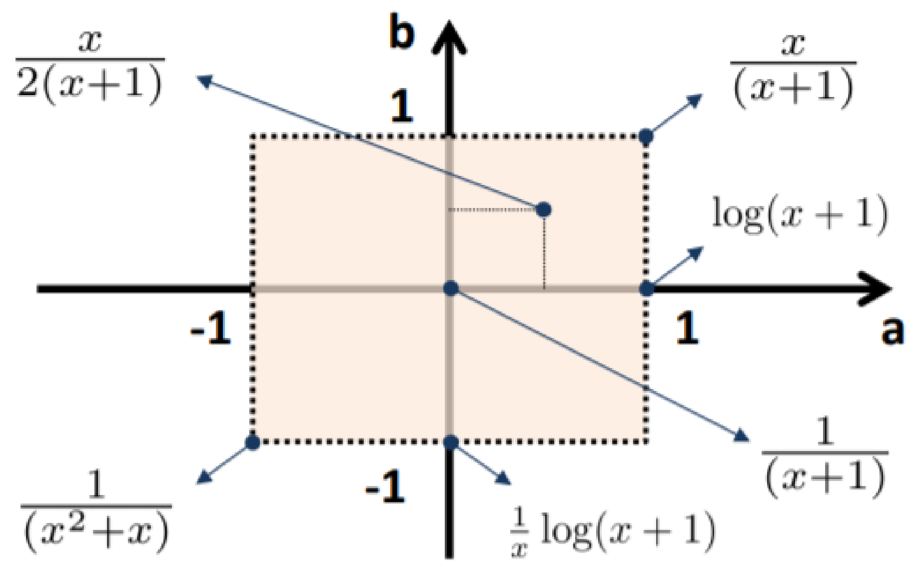
G-function	Equivalent function	G-function	Equivalent function
$G_{0,1}^{1,0}\left(\begin{smallmatrix} - \\ 0 \end{smallmatrix} \middle| -x\right)$	e^x	$G_{2,2}^{1,2}\left(\begin{smallmatrix} \frac{1}{2},1 \\ \frac{1}{2},0 \end{smallmatrix} \middle| x^2\right)$	$2\arctan(x)$
$G_{2,2}^{1,2}\left(\begin{smallmatrix} 1,1 \\ 1,0 \end{smallmatrix} \middle| x\right)$	$\log(1+x)$	$G_{1,2}^{2,0}\left(\begin{smallmatrix} 1 \\ \alpha,0 \end{smallmatrix} \middle| x\right)$	$\Gamma(\alpha, x)$
$G_{0,2}^{1,0}\left(\begin{smallmatrix} - \\ 0,\frac{1}{2} \end{smallmatrix} \middle| \frac{x^2}{4}\right)$	$\frac{1}{\sqrt{\pi}}\cos(x)$	$G_{1,2}^{2,0}\left(\begin{smallmatrix} 1 \\ 0,\frac{1}{2} \end{smallmatrix} \middle| x^2\right)$	$\sqrt{\pi}\,\mathrm{erfc}(x)$
$G_{0,2}^{1,0}\left(\begin{smallmatrix} - \\ \frac{1}{2},0 \end{smallmatrix} \middle| \frac{x^2}{4}\right)$	$\frac{1}{\sqrt{\pi}}\sin(x)$	$G_{0,2}^{1,0}\left(\begin{smallmatrix} - \\ \frac{a}{2},\frac{-a}{2} \end{smallmatrix} \middle| \frac{x^2}{4}\right)$	$J_a(x)$

Advantages

- This means that we can learn symbolic equations by tuning real-valued parameters using gradient descent!

- Example: Tuning symbolic expressions using 2 parameters a & b

$$\widehat{f}(x; a, b) = G^{1,2}_{2,2}\left({a,a \atop a,b} \middle| x \right)$$



Connection to related works

- **Can reduce to different forms of model explanation** by analytic derivations of the symbolic expression.

Derivative: instance-wise feature importance (INVASE, L2X, SHAP, DeepLIFT)

$$
\begin{aligned}
g(\mathbf{x}) \approx \; g(\mathbf{x}_0) \; & + \; (x_1 - x_{0,1}) \cdot g_{x_1}(\mathbf{x}_0) - x_{0,2} \cdot x_1 \cdot g_{x_1 x_2}(\mathbf{x}_0) + \tfrac{1}{2}\,(x_1 - x_{0,1})^2\, g_{x_1 x_1}(\mathbf{x}_0) \\
& + \; (x_2 - x_{0,2}) \cdot g_{x_2}(\mathbf{x}_0) - x_{0,1} \cdot x_2 \cdot g_{x_1 x_2}(\mathbf{x}_0) + \tfrac{1}{2}\,(x_2 - x_{0,2})^2\, g_{x_2 x_2}(\mathbf{x}_0) \\
& + \; x_1 \cdot x_2 \cdot g_{x_1 x_2}(\mathbf{x}_0),
\end{aligned}
$$

Interaction terms: feature interactions (GAM2)

- **Metamodeling provides symbolic equations for instance-wise feature importance!**

Experiments

- **Synthetic experiments: Can recover richer symbolic expressions compared to existing symbolic regression methods based on genetic programming**

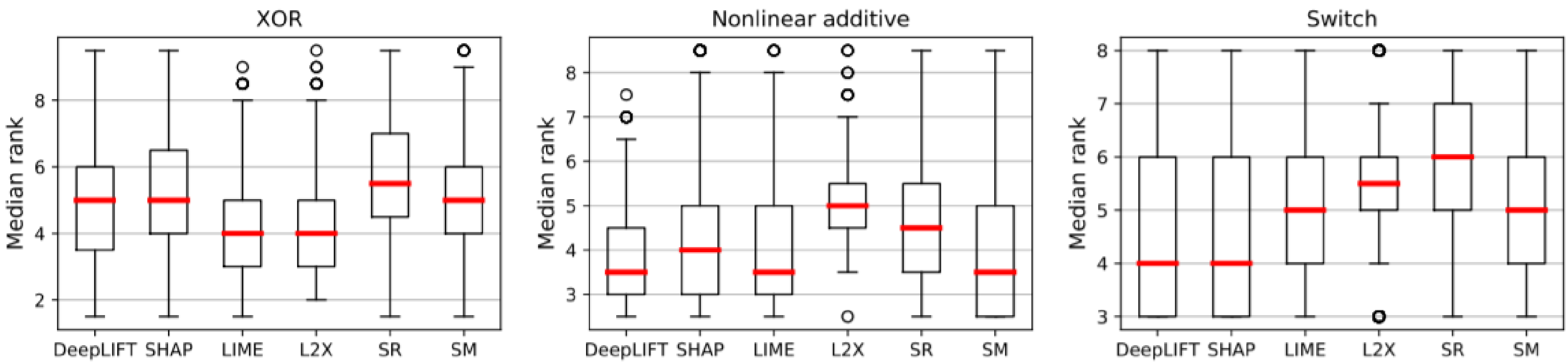
	$f_1(x) = e^{-3x}$	$f_2(x) = \frac{x}{(x+1)^2}$	$f_3(x) = \sin(x)$	$f_4(x) = J_0(10\sqrt{x})$
SM^p	$-x^3 + \frac{5}{2}(x^2 - x) + 1$ R^2: 0.995	$\frac{x^3}{3} - \frac{4x^2}{5} + \frac{2x}{3}$ R^2: 0.985	$\frac{-1}{4}\, x^2 + x$ R^2 : 0.999	$-7\,(x^2 - x) - 1.4$ R^2 : -4.75
SM^c	$x^{4\times10^{-6}}\, e^{-2.99x}$ R^2: 0.999	$x\,(x+1)^{-2}$ R^2: 0.999	$1.4\, x^{1.12}$ R^2 : 0.999	$I_{0.0003}\left(10\, e^{\frac{j\pi}{2}}\, \sqrt{x}\right)$ R^2 : 0.999
SR	$x^2 - 1.9x + 0.9$ R^2: 0.970	$\frac{0.7x}{x^2+0.9x+0.75}$ R^2 0.981	$-0.17\, x^2 + x + 0.016$ R^2 : 0.998	$-x\,(x - 0.773)$ R^2 : 0.116

- **SMp =** meta-modeling restricted to polynomials
- **SMc =** meta-modeling restricted to closed-form expressions
- **SR=** symbolic regression

Experiments

- **Instance-wise feature importance:** 3 standard synthetic datasets with different levels of complexity for which true feature importance is known.

- **Symbolic metamodeling** performs competitively **compared to methods tailored** for feature importance



Experiments

- **Medical applications:** debugs the discrepancies in assigning feature importance in machine learning models and existing medical scores.
- Medical score ignores interactions and hence does not correctly quantify the importance of individual features.