Institute for Enterprise Systems
Empowering Digital Transformation

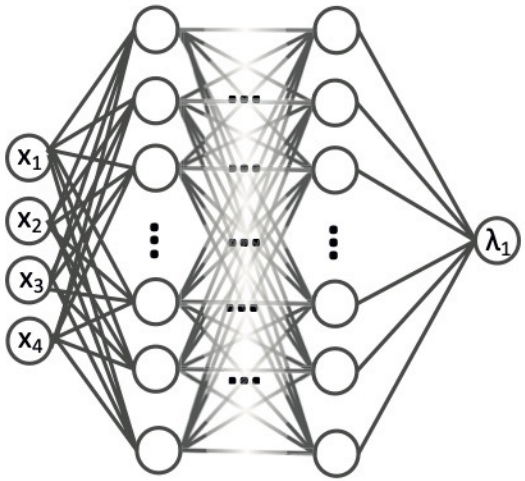
UNIVERSITY OF MANNHEIM

Machine Learning for Converting Black-Box Models to Interpretable Functions

Sascha Marton, Christian Bartelt, Heiner Stuckenschmidt

ECML-PKDD 2020, Virtual Conference, September 14, 2020

Motivation and Goal



Neural Network Representation of a Function λ

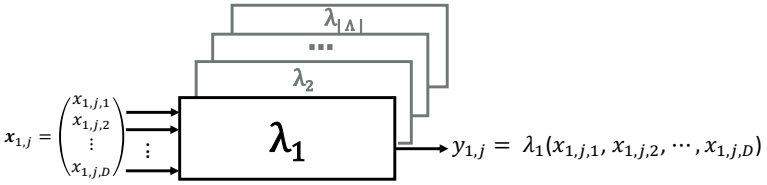
$$\lambda = P(x) = 4x_1^2x_3 - 3x_2 + x_4^3 - 5$$

Polynomial Representation of a Function λ

- **Goal**: Find a mapping from the model internals (e.g. the weights and biases of a neural network) to semantically well defined and human-understandable domain

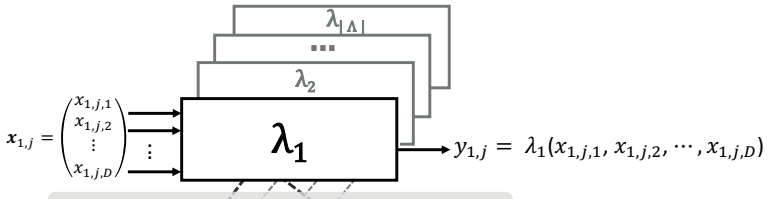
General Framework

1. Generate functions $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$ and datasets $\left\{\mathcal{D}_{\lambda_i} = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1}^{N}\right\}_{i=1}^{|\Lambda|}$



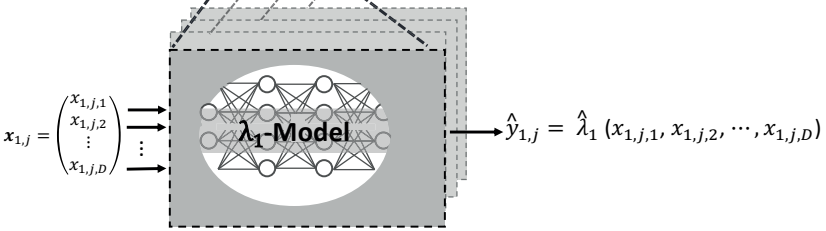
General Framework

1. Generate functions $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$ and datasets $\left\{\mathcal{D}_{\lambda_i} = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1}^{N}\right\}_{i=1}^{|\Lambda|}$



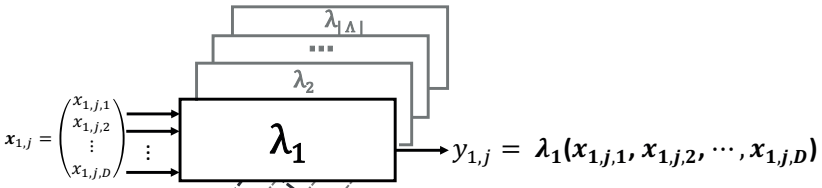
2. Train black-box model for $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$

3. Generate training data for $\mathcal{I}$-Model



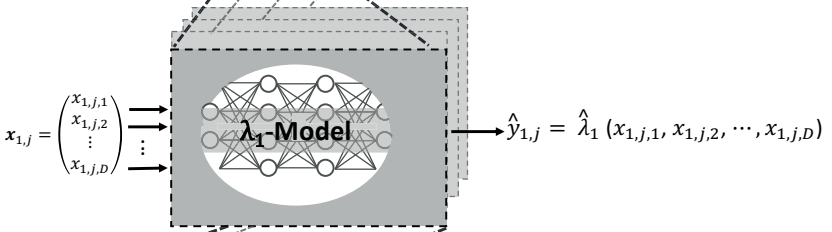
General Framework

1. Generate functions $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$ and datasets
$\left\{\mathcal{D}_{\lambda_i} = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1}^{N}\right\}_{i=1}^{|\Lambda|}$



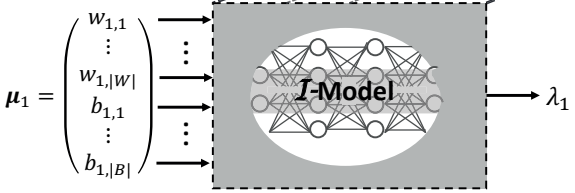
2. Train black-box model for $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$

3. Generate training data for $\mathcal{I}$-Model
   a) $\mathcal{D}_{I_a} = \{(\boldsymbol{\mu}_i, \lambda_i)\}_{i=1}^{|\Lambda|}$



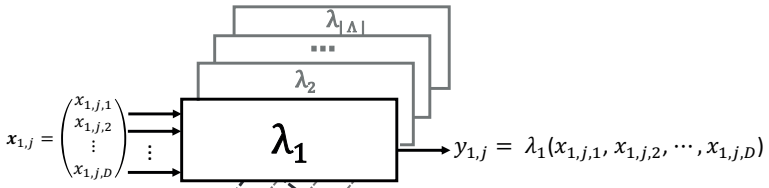
4. Train $\mathcal{I}$-Model

Version a



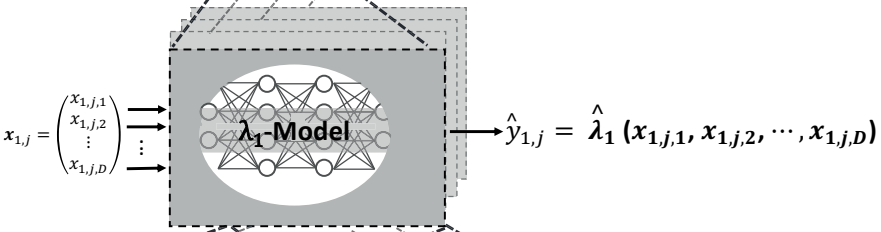
General Framework

1. Generate functions $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$ and datasets
$\left\{\mathcal{D}_{\lambda_i} = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1}^{N}\right\}_{i=1}^{|\Lambda|}$



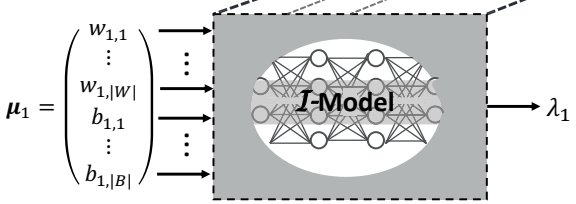
2. Train black-box model for $\lambda_{1,2,\dots,|\Lambda|} \in \Lambda$

3. Generate training data for $\mathcal{I}$-Model
   a) $\mathcal{D}_{I_a} = \{(\boldsymbol{\mu}_i, \lambda_i)\}_{i=1}^{|\Lambda|}$
   b) $\mathcal{D}_{I_b} = \{(\boldsymbol{\mu}_i, \tilde{\hat{\lambda}}_i)\}_{i=1}^{|\Lambda|}$

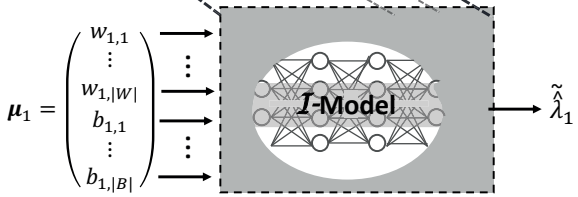


4. Train $\mathcal{I}$-Model

Version a



Version b



Conclusion

- Using a trained Interpretation-Model we are able to interpret arbitrary black-box models as terms of the previously estimated algebra
- To achieve good results, we need to make two assumptions:
  1. It is possible to approximate the model we want to interpret accurately using the estimated algebra
  2. We have found an appropriate representation for functions from this algebra