Global Explanations with Decision Rules: a Co-learning Approach

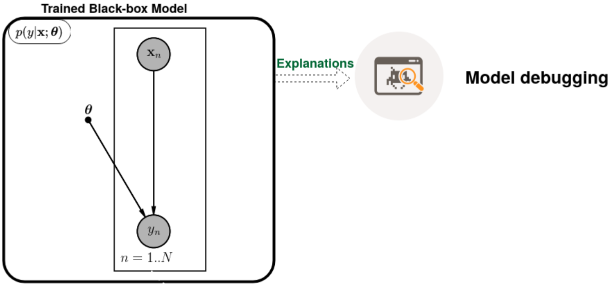
Géraldin Nanfack, Paul Temple and Benoît Frénay

June 10, 2021

UNIVERSITÉ DE NAMUR

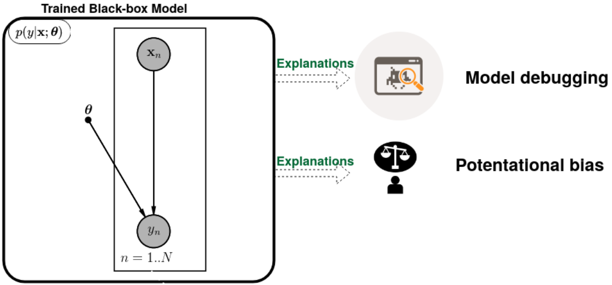
NADI Namur Digital Institute PReCISE

Explainable Artificial Intelligence



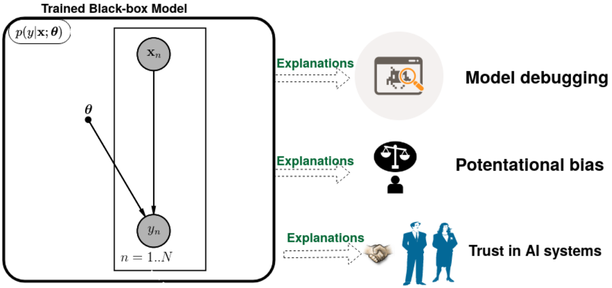
- Models should not only provide accurate predictions but also explanations in human-understandable terms

Explainable Artificial Intelligence



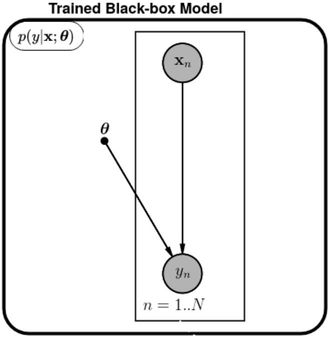
- Models should not only provide accurate predictions but also explanations in human-understandable terms

Explainable Artificial Intelligence



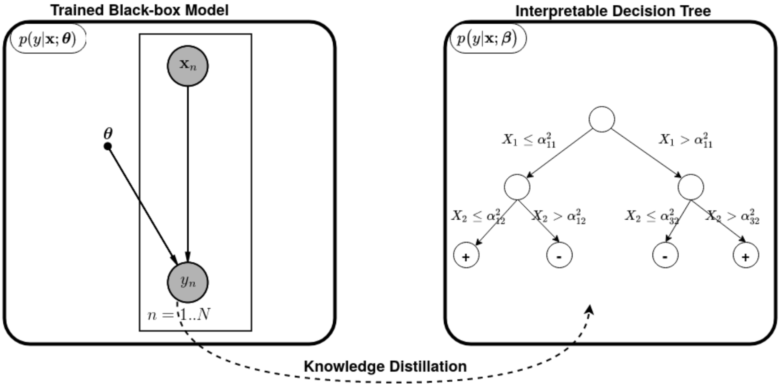
- Models should not only provide accurate predictions but also explanations in human-understandable terms

Posthoc Explainability Methods with Decision Rules



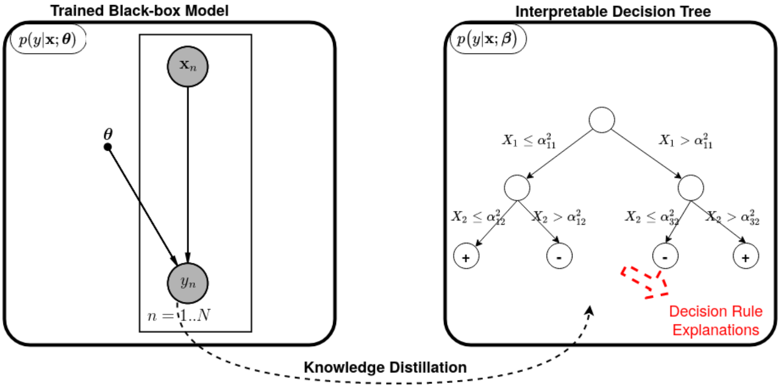
- Traditionally, train the black-box model and then use a surrogate rule learner to extract rule explanations

Posthoc Explainability Methods with Decision Rules



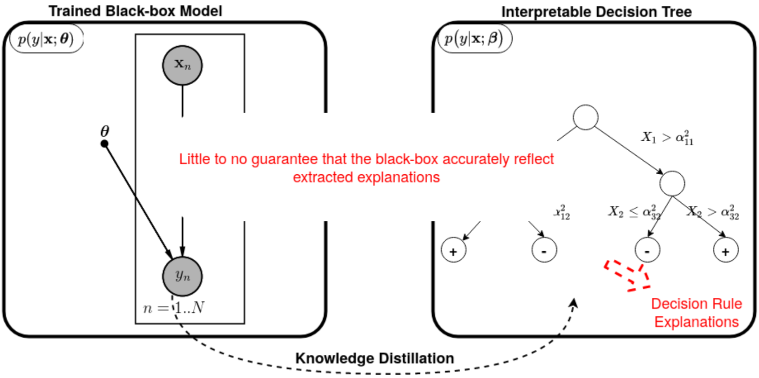
- Traditionally, train the black-box model and then use a surrogate rule learner to extract rule explanations

Posthoc Explainability Methods with Decision Rules



- Traditionally, train the black-box model and then use a surrogate rule learner to extract rule explanations

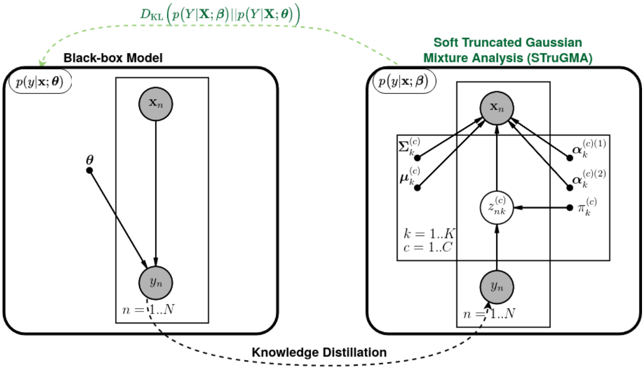
Posthoc Explainability Methods with Decision Rules



- Traditionally, train the black-box model and then use a surrogate rule learner to extract rule explanations

Co-learning for Global Explanations with Decision Rules

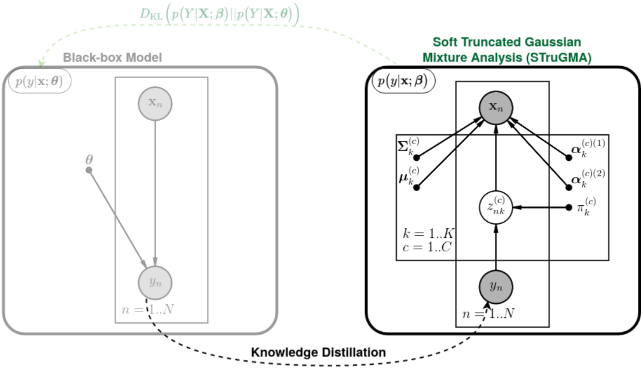
The co-learning framework



- The black-box model is regularised by its rule explanations
- To alleviate the non-differentiability of rule models, we propose STruGMA, a probabilistic model embedding a set of rules

Co-learning for Global Explanations with Decision Rules

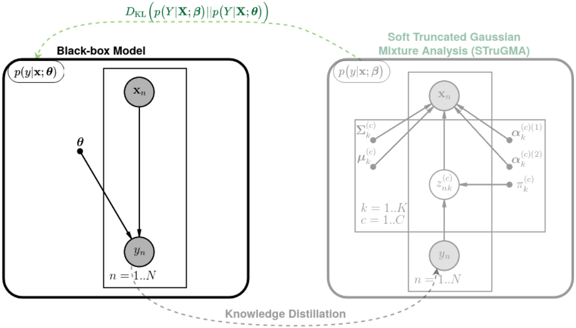
Learning STruGMA



- The soft truncated Gaussian mixture analysis (STruGMA) to embed decision rules with learnable splits on $\boldsymbol{\alpha}_k^{(1)}$ and $\boldsymbol{\alpha}_k^{(2)}$
- Training instances $\boldsymbol{X}$ are relabelled with the outputs $\boldsymbol{Y}_{\boldsymbol{\theta}}$ of the black-box model and STruGMA is learned from that

Co-learning for Global Explanations with Decision Rules

Learning the black-box model



- Loss: $\lambda^* \times \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) + (1 - \lambda^*) \times D_{\mathrm{KL}}\left(p(\mathrm{Y}|\mathrm{X};\boldsymbol{\beta})||p(\mathrm{Y}|\mathrm{X};\boldsymbol{\theta})\right)$,
- Divergence term: $\approx \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{c=1}^{C} p(y = c|\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) \log \frac{p(y=c|\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})}{p(y=c|\hat{\boldsymbol{x}}_i;\boldsymbol{\theta})}$,
- $\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})$ is the cross-entropy loss
- λ^* is set using the multiple gradient descent algorithm (MGDA)[1]

[1] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. Neurips 2018.

Co-learning for Global Explanations with Decision Rules

Results with deep neural networks

- Co-learning improves fidelity of decision rule explanations

Dataset	TreeExplainer	TreeCoExplainerHR	TreeCoExplainerBB
Bank	95.97 (0.74)	96.18 (0.63)	**96.49 (0.89)**
Credit	77.3 (3.47)	81.25 (3.47)	**81.5 (3.43)**
Ionosphere	87.32 (3.25)	**90.28 (3.42)**	88.87 (5.69)
Gamma	93.31 (2.08)	93.15 (0.85)	**95.6 (0.36)**
Pima	88.44 (2.41)	88.9 (1.35)	**92.01 (3.24)**
Waveform	80.26 (1.53)	80.52 (1.87)	**80.86 (1.28)**
Wine	89.17 (4.62)	**92.78 (4.93)**	89.72 (2.64)

Co-learning for Global Explanations with Decision Rules

Results with deep neural networks

- Co-learning improves fidelity of decision rule explanations

Dataset	TreeEx-plainer	TreeCoEx-plainerHR	TreeCoEx-plainerBB
Bank	95.97 (0.74)	96.18 (0.63)	**96.49 (0.89)**
Credit	77.3 (3.47)	81.25 (3.47)	**81.5 (3.43)**
Ionosphere	87.32 (3.25)	**90.28 (3.42)**	88.87 (5.69)
Gamma	93.31 (2.08)	93.15 (0.85)	**95.6 (0.36)**
Pima	88.44 (2.41)	88.9 (1.35)	**92.01 (3.24)**
Waveform	80.26 (1.53)	80.52 (1.87)	**80.86 (1.28)**
Wine	89.17 (4.62)	**92.78 (4.93)**	89.72 (2.64)

- Co-learning has a limited impact on the performance of the black-box model:

Dataset	coBB	BB
Bank	90.68 (0.77)	**90.99 (0.84)**
Credit	**75.65 (3.88)**	74.75 (3.5)
Ionosphere	**90.98 (3.88)**	90.56 (3.45)
Gamma	80.57 (0.49)	**82.79 (2.53)**
Pima	73.12 (2.31)	**75.39 (1.77)**
Waveform	85.97 (0.87)	**86.15 (0.7)**
Wine	96.94 (2.43)	**97.5 (2.05)**

Co-learning for Global Explanations with Decision Rules

Results with deep neural networks

- The co-learned black-box model continues to be competitive against the decision tree

Dataset	**coBB**	**DT**
Bank	90.68 (0.77)	**90.81 (0.96)**
Credit	**75.65 (3.88)**	71.05 (3.3)
Ionosphere	**90.98 (3.88)**	90.28 (4.43)
Gamma	80.57 (0.49)	**82.72 (0.43)**
Pima	**73.12 (2.31)**	72.02 (2.59)
Waveform	**85.97 (0.87)**	75.24 (1.23)
Wine	**96.94 (2.43)**	87.78 (4.93)

Co-learning for Global Explanations with Decision Rules

Results with deep neural networks

- The co-learned black-box model continues to be competitive against the decision tree

Dataset	coBB	DT
Bank	90.68 (0.77)	**90.81 (0.96)**
Credit	**75.65 (3.88)**	71.05 (3.3)
Ionosphere	**90.98 (3.88)**	90.28 (4.43)
Gamma	80.57 (0.49)	**82.72 (0.43)**
Pima	**73.12 (2.31)**	72.02 (2.59)
Waveform	**85.97 (0.87)**	75.24 (1.23)
Wine	**96.94 (2.43)**	87.78 (4.93)

- We also perform a qualitative evaluation with a medical doctor on two medical datasets
  - Explanations were mostly clinically correct

Thank you for your attention!

- Further details, see paper #231
- Questions?