Global Explanations with Decision Rules: a Co-learning Approach

Géraldin Nanfack, Paul Temple and Benoît Frénay

PReCISE Research Center, Namur Digital Institute (NADI), University of Namur, Belgium

Introduction

- Medical or justice models may require models to provide explanations for their predictions
- Interpretable decision lists and trees exist but in practice, more powerful models like deep networks perform well even for tabular data [1]
- Post-hoc explainability methods exist but there are little guarantees that explanations accurately reflect knowledge learned by the black-box model
- Recent works have proposed to regularise black-box models for explainability but they require prior knowledge

Our Approach

The Soft Truncated Gaussian Mixture Analysis (STruGMA) is a differentiable probabilistic model designed to embed a set of hyper-rectangle rules.

- Given a black-box model, we aim to learn global decision rule explanations that reflect knowledge embedded in the black-box model.
- We propose to co-learn the black-box model and STruGMA and show that thanks to co-learning (i) the black-box model becomes easier to explain and (ii) remains competitive in terms of accuracy.

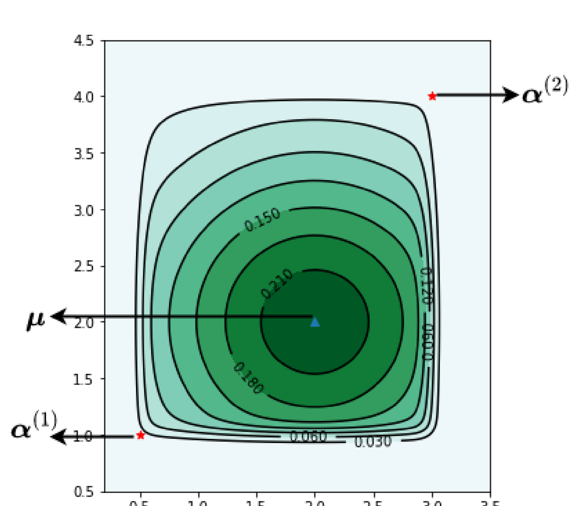

Figure: A soft Truncated Gaussian Distribution

$$p\left(\boldsymbol{x} \middle| z=k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}\right) \approx \frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\int_{\boldsymbol{\alpha}_k^{(1)}}^{\boldsymbol{\alpha}_k^{(2)}} \mathcal{N}(\boldsymbol{t}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{t}} \prod_{d=1}^{D} \sigma_\eta\left(x_d - \alpha_{kd}^{(1)}\right)\left(1 - \sigma_\eta\left(x_d - \alpha_{kd}^{(2)}\right)\right),$$

The Co-learning Framework

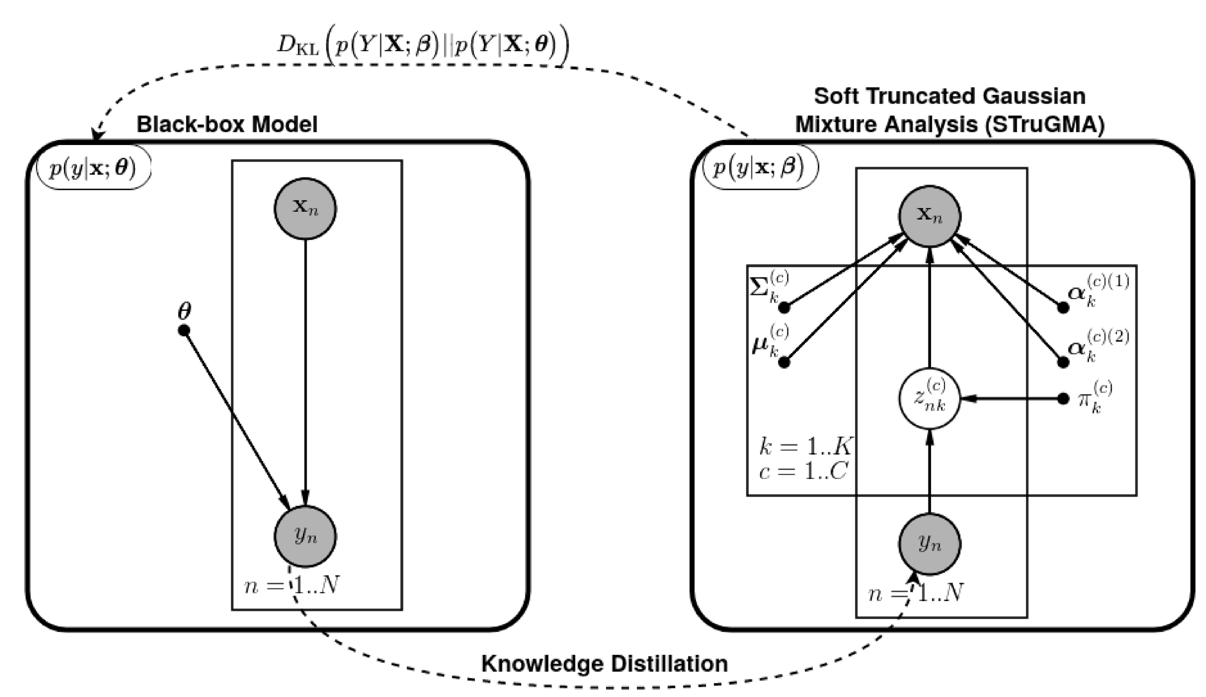

Figure: Left: the black-box model, right: STruGMA

Learning the black-box model

The loss of the black-box model is

$$\lambda^* \times \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) + (1-\lambda^*) \times D_{\mathrm{KL}}\big(p(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})||p(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta})\big),$$

where

$$D_{\mathrm{KL}}\big(p(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})||p(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta})\big) = \mathbb{E}_{\boldsymbol{x}\sim p(\mathbf{x};\boldsymbol{\beta})}[D_{\mathrm{KL}}(p(\mathbf{Y}|\boldsymbol{x};\boldsymbol{\beta})||p(\mathbf{Y}|\boldsymbol{x};\boldsymbol{\theta}))] \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{c=1}^{C} p(y=c|\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) \log \frac{p(y=c|\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})}{p(y=c|\hat{\boldsymbol{x}}_i;\boldsymbol{\theta})},$$

$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})$ is the cross-entropy loss and using the multiple gradient descent algorithm (MGDA)[2, 3]

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}} \Big\| \lambda \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) + (1-\lambda)\nabla_{\boldsymbol{\theta}} D_{\mathrm{KL}}\big(p(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})||p(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta})\big) \Big\|$$

Learning STruGMA

STruGMA is learned using the EM algorithm with challenges that we solved, namely:

- No closed-form solution: gradient descent in M-Step
- Hard constraints $\boldsymbol{\alpha}^{(1)} < \boldsymbol{\alpha}^{(2)}$: projected gradient
- Overlapping rules: simple yet effective heuristic to enforce the constraint:

$$\max_d \left(\left| \frac{1}{2}\left(\alpha_{id}^{(1)} + \alpha_{id}^{(2)}\right) - \frac{1}{2}\left(\alpha_{jd}^{(1)} + \alpha_{jd}^{(2)}\right) \right| - \frac{1}{2}\left(\alpha_{id}^{(2)} - \alpha_{id}^{(1)}\right) - \frac{1}{2}\left(\alpha_{jd}^{(2)} - \alpha_{jd}^{(1)}\right) \right) \geq 0.$$

Finally, we use **knowledge distillation** so that STruGMA globally explains the black-box model.

Results

- **Fidelity is improved** thanks to co-learning

Table: TreeExplainer is the baseline and TreeCoExplainerHR and TreeCoExplainerBB are ours.

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)

- **Model's accuracy** is **little** impacted ($\pm 2\%$)

Table: Predictive accuracy of co-learned black-box models (coBB) and black-box models without co-learning (BB).

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)**

- And the **co-learned black-box model** continues to be **competitive**

Table: Predictive accuracy of co-learned black-box models (coBB) and decision trees (DT).

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)

References

[1] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *ICLR*, 2017.

[2] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Neurips*, 2018.

[3] Jean-Antoine Désidéri. *Multiple-Gradient Descent Algorithm (MGDA)*. PhD thesis, INRIA, 2009.