Global Explanations with Decision Rules: a Co-learning Approach (supplementary material)

Géraldin Nanfack[1] **Paul Temple**[1] **Benoît Frénay**[1]

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

A GRADIENT OF THE EXPECTED LOG-LIKELIHOOD

This section reports the gradient of the expected log-likelihood of STruGMA w.r.t. its parameters $\boldsymbol{\beta}$.

We remind that when omitting class conditioning c, the expected log-likelihood is:

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^t) = \sum_n \sum_k r_{nk} \log \pi_k + \sum_n \sum_k r_{ik} \left[\log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_d \log \sigma_\eta \left(x_{nd} - \alpha_{kd}^{(1)} \right) + \log \left(1 - \sigma_\eta \left(x_{nd} - \alpha_{kd}^{(2)} \right) \right) \right] - \sum_n \sum_k r_{nk} \log \int_{\boldsymbol{\alpha}_k^{(1)}}^{\boldsymbol{\alpha}_k^{(2)}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\mathbf{x}.$$

We consider the case where $\boldsymbol{\Sigma}_k$ is diagonal i.e. $\boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{k1}^2 & & \\ & \ddots & \\ & & \sigma_{kd}^2 \end{bmatrix}$. To avoid any confusion, the σ_{kd} are different here from $\sigma_\eta(x) = 1/\left(1 + \exp(-\eta x)\right)$.

Derivatives with respect to each parameter of $\boldsymbol{\beta}$ are:

$$\frac{\partial Q(\boldsymbol{\beta}, \boldsymbol{\beta}^t)}{\partial \mu_{kd}} = \sum_n r_{nk} \left(\frac{x_{nd} - \mu_{kd}}{\sigma_{kd}^2} - \frac{\mathcal{N}(\alpha_{kd}^{(1)}; \mu_{kd}, \sigma_{kd}^2) - \mathcal{N}(\alpha_{kd}^{(2)}; \mu_{kd}, \sigma_{kd}^2)}{F(\alpha_{kd}^{(2)}; \mu_{kd}, \sigma_{kd}^2) - F(\alpha_{kd}^{(1)}; \mu_{kd}, \sigma_{kd}^2)} \right),$$

$$\frac{\partial Q(\boldsymbol{\beta}, \boldsymbol{\beta}^t)}{\partial \sigma_{kd}} = \sum_n r_{nk} \left(\frac{(x_{nd} - \mu_{kd})^2}{\sigma_{kd}^2} - \frac{1}{\sigma_{kd}} - \frac{1}{\sigma_{kd}} \times \frac{\mathcal{N}(\alpha_{kd}^{(1)}; \mu_{kd}, \sigma_{kd}^2) - \mathcal{N}(\alpha_{kd}^{(2)}; \mu_{kd}, \sigma_{kd}^2)}{F(\alpha_{kd}^{(2)}; \mu_{kd}, \sigma_{kd}^2) - F(\alpha_{kd}^{(1)}; \mu_{kd}, \sigma_{kd}^2)} \right),$$

$$\frac{Q(\boldsymbol{\beta}, \boldsymbol{\beta}^t)}{\partial \alpha_{kd}^{(1)}} = \sum_n \eta r_{nk} \left(1 - \sigma_\eta (x_{nd} - \alpha_{kd}^{(1)}) \right),$$

$$\frac{Q(\boldsymbol{\beta}, \boldsymbol{\beta}^t)}{\partial \alpha_{kd}^{(2)}} = \sum_n -\eta r_{nk} \sigma_\eta (x_{nd} - \alpha_{kd}^{(2)}),$$

where d is a feature number, n is a data-instance number, k is a component number of STru GMA and F is the cumulative distribution function of the univariate normal distribution.

B BREAKING THE OVERLAPPING

This section provides details for the heuristic we used to break the overlapping.

Considering two hyper-rectangles i and j, the non-overlapping constraint is

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

By looking at the form of the constraint, it can be seen that breaking overlapping can be done on only one particular dimension d. In cases where the gradient-based updates of parameters violate the constraint in Eq. 1, given a dimension d, our heuristic considers 4 adaptations:

$$\begin{cases} \text{(i) } \alpha_{id}^{(2)} = \alpha_{jd}^{(1)} & \text{if } \alpha_{jd}^{(1)} > \alpha_{id}^{(1)} \\ \text{(ii) } \alpha_{jd}^{(2)} = \alpha_{id}^{(1)} & \textit{otherwise} \end{cases} \tag{2}$$

and

$$\begin{cases} \text{(iii) } \alpha_{jd}^{(1)} = \alpha_{id}^{(2)} & \text{if } \alpha_{jd}^{(2)} > \alpha_{id}^{(2)} \\ \text{(iv) } \alpha_{id}^{(1)} = \alpha_{jd}^{(2)} & \textit{otherwise.} \end{cases} \tag{3}$$

Note that (ii) and (iv) are simply alternatives of (i) and (iii) respectively when permuting i and j. Furthermore, by applying any of these adaptations, it can be checked that the constraints $\boldsymbol{\alpha}_i^{(2)} > \boldsymbol{\alpha}_i^{(1)}$ and $\boldsymbol{\alpha}_j^{(2)} > \boldsymbol{\alpha}_i^{(1)}$ are satisfied.

Table 1: Details of neural network architectures.

Network	Architecture	Datasets
Network1	Dense layer - 128, ELU Dense layer - 128, ELU Dense layer - C, Softmax	Marketing, Credit, Pima, Waveform, Wine
Network3	Dense layer - 40, ELU Dropout - 0.4 Dense layer - 25, ELU Dense layer - 10, ELU Dropout - 0.4 Dense layer - C, Softmax	Magic gamma
Network2	Dense layer - 128, ELU Dense layer - 128, ELU Dropout - 0.4 Dense layer - 256, ELU Dense layer - 256, ELU Dropout - 0.4 Dense layer - C, Softmax	Ionosphere

Figure 1 illustrates adaptations along a particular dimension d.

As we have 4 adaptations per dimension, therefore, there are $4D$ choices, where D is the size of the input space. The best choice is taken as the choice that maximises the expected log-likelihood score. Another interpretation is that it is the choice that minimises the loss in *coverage* of data-instances by STruGMA.

In summary, with this heuristic, breaking overlapping has the algorithmic complexity $\mathcal{O}(K^2 \times D)$, where K is the number of hyper-rectangles and D is the number of features.

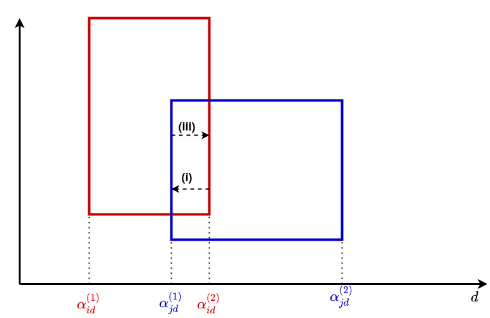

Figure 1: Illustration of adaptations along a particular dimension d. (i) and (iii) correspond to adaptations described in Eq. 2 and Eq. 3.

C DETAILS OF NEURAL NETWORK ARCHITECTURES

The three architectures in the table 1 where used in the experiments. Neural networks with different architectures that gave best results were chosen for each dataset as the black-box.