Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 14

Exercise 1: Class Prototypes (25 P)

Consider the linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ mapping some input $\boldsymbol{x}$ to an output $f(\boldsymbol{x})$. We would like to interpret the function f by building a prototype $\boldsymbol{x}^\star$ in the input domain which produces a large value f. Activation maximization produces such interpretation by optimizing

$$\max_{\boldsymbol{x}} \big[f(\boldsymbol{x}) + \Omega(\boldsymbol{x})\big].$$

Find the prototype $\boldsymbol{x}^\star$ obtained by activation maximization subject to $\Omega(\boldsymbol{x}) = \log p(\boldsymbol{x})$ with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ and Σ are the mean and covariance.

Exercise 2: Shapley Values (25 P)

Consider the function $f(\boldsymbol{x}) = \min(x_1, \max(x_2, x_3))$. Compute the Shapley values ϕ_1, ϕ_2, ϕ_3 for the prediction $f(\boldsymbol{x})$ with $\boldsymbol{x} = (1, 1, 1)$. (We assume a reference point $\widetilde{\boldsymbol{x}} = \boldsymbol{0}$, i.e. we set features to zero when removing them from the coalition).

Exercise 3: Taylor Expansions (25 P)

Consider the simple radial basis function

$$f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\mu}\| - \theta$$

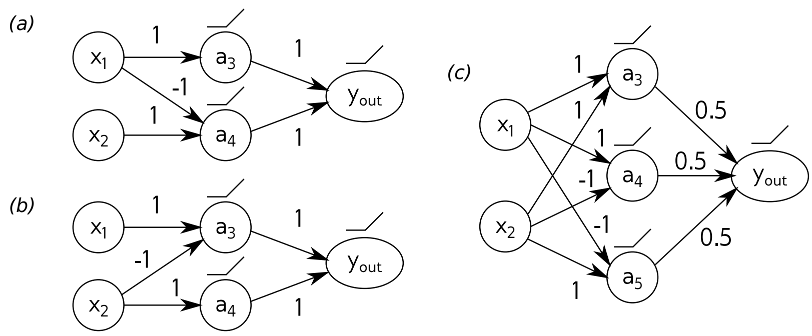
with $\theta > 0$. For the purpose of extracting an explanation, we would like to build a first-order Taylor expansion of the function at some root point $\widetilde{\boldsymbol{x}}$. We choose this root point to be taken on the segment connecting $\boldsymbol{\mu}$ and $\boldsymbol{x}$ (we assume that $f(\boldsymbol{x}) > 0$ so that there is always a root point on this segment).

Show that the first-order terms of the Taylor expansion are given by

$$\phi_i = \frac{(x_i - \mu_i)^2}{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2} \cdot (\|\boldsymbol{x} - \boldsymbol{\mu}\| - \theta)$$

Exercise 4: Layer-Wise Relevance Propagation (25 P)

We would like to test the dependence of layer-wise relevance propagation (LRP) on the structure of the neural network. For this, we consider the function $y = \max(x_1, x_2)$, where $x_1, x_2 \in \mathbb{R}^+$ are the input activations. This function can be implemented as a ReLU network in multiple ways. Three examples are given below.



We consider the propagation rule:

$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

where j and k are indices for two consecutive layers and where $()^+$ denotes the positive part. This propagation rule is applied to both layers.

Give for each network the computational steps that lead to the scores R_1 and R_2, and the obtained relevance values. More specifically, express R_1 and R_2 as a function of R_3 and R_4 (and R_5), and express the latter relevances as a function of $R_{\text{out}} = y$.

Exercise 1: Class Prototypes (25 P)

Consider the linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ mapping some input $\boldsymbol{x}$ to an output $f(\boldsymbol{x})$. We would like to interpret the function f by building a prototype $\boldsymbol{x}^\star$ in the input domain which produces a large value f. Activation maximization produces such interpretation by optimizing

$$\max_{\boldsymbol{x}} \left[f(\boldsymbol{x}) + \Omega(\boldsymbol{x}) \right].$$

Find the prototype $\boldsymbol{x}^\star$ obtained by activation maximization subject to $\Omega(\boldsymbol{x}) = \log p(\boldsymbol{x})$ with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ and Σ are the mean and covariance.

$$f(x) + \Omega(x) = w^T x + b + \log p(x)$$

$$= w^T x + b + \log \frac{1}{\sqrt{(2\pi)^R \det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$= w^T x + b - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \underbrace{\log \frac{1}{\sqrt{(2\pi)^R \det(\Sigma)}}}_{= \text{const}}$$

$$\Rightarrow \max_x \left\{ f(x) + \Omega(x) \right\} = \max_x \left\{ w^T x + b - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\}$$

$$\nabla_x \left(w^T x + b - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right) = w - \Sigma^{-1}(x-\mu) = \vec{0}$$

$$\Sigma \cdot \Sigma^{-1}(x-\mu) = \Sigma w$$ (To simplify the Calculation, we assume Σ is PD)

$$x - \mu = \Sigma w$$

$$x^* = \mu + \Sigma w$$

Exercise 2: Shapley Values (25 P)

Consider the function $f(\boldsymbol{x}) = \min(x_1, \max(x_2, x_3))$. Compute the Shapley values ϕ_1, ϕ_2, ϕ_3 for the prediction $f(\boldsymbol{x})$ with $\boldsymbol{x} = (1, 1, 1)$. (We assume a reference point $\widetilde{\boldsymbol{x}} = \boldsymbol{0}$, i.e. we set features to zero when removing them from the coalition).

$$\phi_i = \sum_{S: i \notin S} \underbrace{\frac{|S|!\,(d-|S|-1)!}{d!}}_{\alpha_S} \underbrace{\left[f(x_{S \cup \{i\}}) - f(x_S) \right]}_{\Delta_S}$$

ϕ_1:

S	α_S	Δ_S
$\{\}$	$\frac{0!(3-0-1)!}{3!} = \frac{1}{3}$	$f(1,0,0) - f(0,0,0) = 0 - 0 = 0$
$\{2\}$	$\frac{1!(3-2)!}{3!} = \frac{1}{6}$	$f(1,1,0) - f(0,1,0) = 1 - 0 = 1$
$\{3\}$	$\frac{1}{6}$	$f(1,0,1) - f(0,0,1) = 1 - 0 = 1$
$\{2,3\}$	$\frac{2!(3-2-1)!}{3!} = \frac{1}{3}$	$f(1,1,1) - f(0,1,1) = 1 - 0 = 1$

$$\phi_1 = \frac{1}{3} \cdot 0 + \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 1 + \frac{1}{3} \cdot 1 = \frac{2}{3}$$

ϕ_2:

S	α_S	Δ_S
$\{\}$	$\frac{1}{3}$	$f(0,1,0) - f(0,0,0) = 0 - 0 = 0$
$\{1\}$	$\frac{1}{6}$	$f(1,1,0) - f(1,0,0) = 1 - 0 = 1$
$\{3\}$	$\frac{1}{6}$	$f(0,1,1) - f(0,0,1) = 0 - 0 = 0$
$\{1,3\}$	$\frac{1}{3}$	$f(1,1,1) - f(1,0,1) = 1 - 1 = 0$

$$\phi_2 = \frac{1}{3} \cdot 0 + \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 0 + \frac{1}{3} \cdot 0 = \frac{1}{6}$$

$$\phi_1 = \frac{2}{3} \rightarrow \phi_2 + \phi_3 = \frac{1}{3} \xrightarrow{\text{Symmetric}} \phi_2 = \phi_3 = \frac{1}{6}$$

ϕ_3:

S	α_S	Δ_S
$\{\}$	$\frac{1}{3}$	$f(0,0,1)-f(0,0,0) = 0-0=0$
$\{1\}$	$\frac{1}{6}$	$f(1,0,1)-f(1,0,0) = 1-0=1$
$\{2\}$	$\frac{1}{6}$	$f(0,1,1)-f(0,1,0) = 0$
$\{1,2\}$	$\frac{1}{3}$	$f(1,1,1)-f(1,1,0) = 1-1=0$

$\phi_3 = \frac{1}{6}$

Exercise 3: Taylor Expansions (25 P)

Consider the simple radial basis function

$$f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\mu}\| - \theta$$

with $\theta > 0$. For the purpose of extracting an explanation, we would like to build a first-order Taylor expansion of the function at some root point $\widetilde{\boldsymbol{x}}$. We choose this root point to be taken on the segment connecting $\boldsymbol{\mu}$ and $\boldsymbol{x}$ (we assume that $f(\boldsymbol{x}) > 0$ so that there is always a root point on this segment).

Show that the first-order terms of the Taylor expansion are given by

$$\phi_i = \frac{(x_i - \mu_i)^2}{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2} \cdot (\|\boldsymbol{x} - \boldsymbol{\mu}\| - \theta)$$

$$f(x) = \underbrace{f(\widetilde{x})}_{=0} + \sum_{i=1}^{d} \underbrace{[\nabla f(\widetilde{x})]_i \cdot (x_i - \widetilde{x}_i)}_{\phi_i} + \ldots$$

$$f(x) = \|x-\mu\| - \theta = \left((x-\mu)^T(x-\mu)\right)^{\frac{1}{2}} - \theta$$

$$\nabla f(x) = \frac{1}{2} \cdot \left((x-\mu)^T \cdot (x-\mu)\right)^{-\frac{1}{2}} \cdot 2(x-\mu)$$

$$= \frac{x-\mu}{\|x-\mu\|}$$

In Taylor expansion we know $f(\tilde{x}) = \|\tilde{x} - \mu\| - \theta = 0$

$$\Rightarrow \|\tilde{x} - \mu\| = \theta \quad \ldots (1)$$

we know that $\tilde{x}$ lies on the sphere centered at μ with radius θ.

$$\Rightarrow \tilde{x} - \mu = \theta \cdot \frac{x-\mu}{\|x-\mu\|} \quad \ldots (2)$$

$$\phi_i = [\nabla f(\tilde{x})]_i \cdot (x_i - \tilde{x}_i) = \frac{\tilde{x}_i - \mu_i}{\|\tilde{x} - \mu\|} \cdot (x_i - \tilde{x}_i)$$

$$= \frac{\tilde{x}_i - \mu_i}{\theta} (x_i - \tilde{x}_i) \quad \Big| (1)$$

$$= \frac{\theta \cdot \frac{x_i - \mu_i}{\|x-\mu\|}}{\theta} (x_i - \tilde{x}_i) \quad \Big| (2)$$

$$= \frac{x_i - \mu_i}{\|x-\mu\|} (x_i - \tilde{x}_i)$$

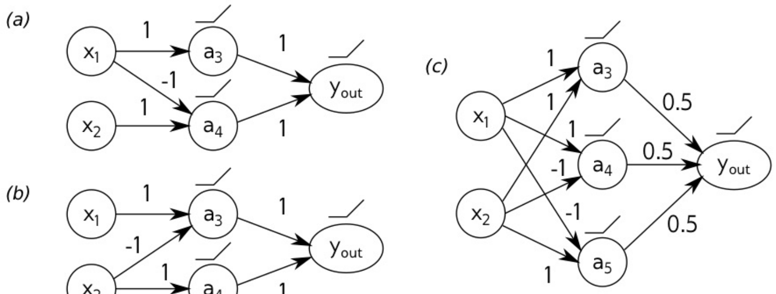
$$= \frac{x_i - \mu_i}{\|x-\mu\|} \cdot \left(x_i - \frac{x_i - \mu_i}{\|x-\mu\|} \cdot \theta - \mu_i\right) \quad \Big| (2)$$

$$= \frac{(x_i - \mu_i)^2}{\|x-\mu\|} \cdot \left(1 - \frac{\theta}{\|x-\mu\|}\right)$$

$$= \frac{(x_i - \mu_i)^2}{\|x-\mu\|^2} (\|x-\mu\| - \theta)$$

Exercise 4: Layer-Wise Relevance Propagation (25 P)

We would like to test the dependence of layer-wise relevance propagation (LRP) on the structure of the neural network. For this, we consider the function $y = \max(x_1, x_2)$, where $x_1, x_2 \in \mathbb{R}^+$ are the input activations. This function can be implemented as a ReLU network in multiple ways. Three examples are given below.



We consider the propagation rule:

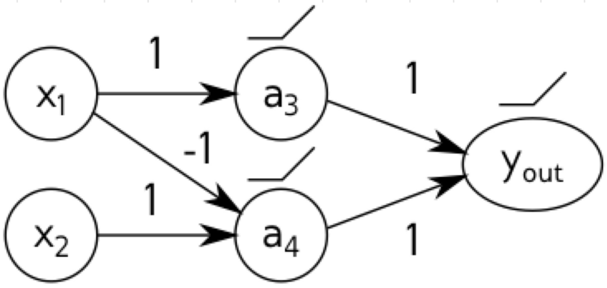
$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

j is all nodes in left sides

where j and k are indices for two consecutive layers and where $()^+$ denotes the positive part. This propagation rule is applied to both layers.

Give for each network the computational steps that lead to the scores R_1 and R_2, and the obtained relevance values. More specifically, express R_1 and R_2 as a function of R_3 and R_4 (and R_5), and express the latter relevances as a function of $R_{\text{out}} = y$.

(a)



$$R_5 = R_{out} = Y$$

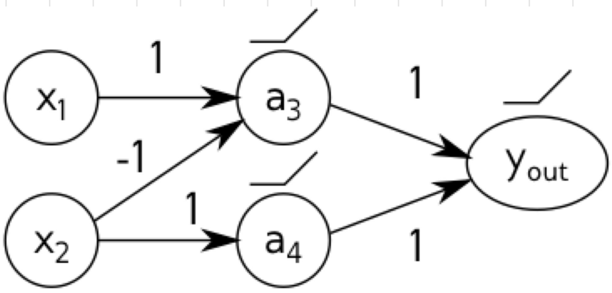
$$R_4 = \frac{a_4 w_{45}}{\sum_{k=3}^{4} a_k w_{k5}} \cdot R_5 = \frac{a_4}{a_3 + a_4} Y$$

$$R_3 = \frac{a_3 V_3}{\sum_{k=3}^{4} a_k V_k} \cdot R_5 = \frac{a_3}{a_3 + a_4} Y$$

$$R_2 = \sum_{k=3}^{4} \frac{a_2 w_{2k}^+}{\sum_{j=1}^{2} a_j w_{jk}^+} R_k = R_3 \cdot \frac{a_2 \overset{0}{\overset{\shortparallel}{w_{23}^+}}}{a_1 \underset{1}{\underset{\shortparallel}{w_{13}^+}} + a_2 \underset{0}{\underset{\shortparallel}{w_{23}^+}}} + R_4 \cdot \frac{a_2 \overset{1}{\overset{\shortparallel}{w_{24}^+}}}{a_1 \underset{0}{\underset{\shortparallel}{w_{14}^+}} + a_2 \underset{1}{\underset{\shortparallel}{w_{24}^+}}} = R_4$$

$$R_1 = \sum_{k=3}^{4} \frac{a_1 w_{1k}^+}{\sum_{j=1}^{2} a_j w_{jk}^+} R_k = R_3 \frac{a_1 \overset{1}{\overset{\shortparallel}{w_{13}^+}}}{a_1 \underset{1}{\underset{\shortparallel}{w_{13}^+}} + a_2 \underset{0}{\underset{\shortparallel}{w_{23}^+}}} + R_4 \cdot \frac{a_1 \overset{0}{\overset{\shortparallel}{w_{14}^+}}}{a_1 \underset{0}{\underset{\shortparallel}{w_{14}^+}} + a_2 \underset{1}{\underset{\shortparallel}{w_{24}^+}}} = R_3$$

(b)



$$R_5 = R_{out} = Y$$

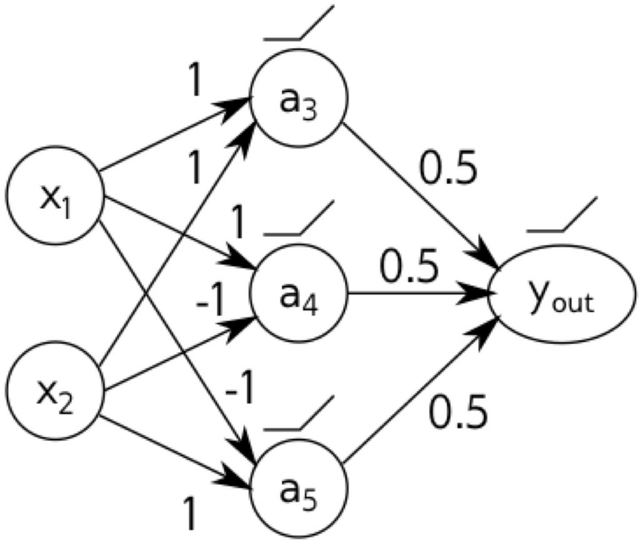
$$R_4 = \frac{a_4 w_{45}}{\sum_{k=3}^{4} a_k w_{k5}} \cdot R_5 = \frac{a_4}{a_3+a_4} Y$$

$$R_3 = \frac{a_3 V_3}{\sum_{k=3}^{4} a_k V_k} \cdot R_5 = \frac{a_3}{a_3+a_4} Y$$

$$R_2 = \sum_{k=3}^{4} \frac{a_2 w_{2k}^+}{\sum_{j=1}^{2} a_j w_{jk}^+} R_k = R_3 \cdot \frac{\overset{0}{\overset{\shortparallel}{a_2 w_{23}^+}}}{\underset{\underset{1}{\shortparallel}}{a_1 w_{13}^+} + \underset{\underset{0}{\shortparallel}}{a_2 w_{23}^+}} + R_4 \cdot \frac{\overset{1}{\overset{\shortparallel}{a_2 w_{24}^+}}}{\underset{\underset{0}{\shortparallel}}{a_1 w_{14}^+} + \underset{\underset{1}{\shortparallel}}{a_2 w_{24}^+}} = R_4$$

$$R_1 = \sum_{k=3}^{4} \frac{a_1 w_{1k}^+}{\sum_{j=1}^{2} a_j w_{jk}^+} R_k = R_3 \frac{\overset{1}{\overset{\shortparallel}{a_1 w_{13}^+}}}{\underset{\underset{1}{\shortparallel}}{a_1 w_{13}^+} + \underset{\underset{0}{\shortparallel}}{a_2 w_{23}^+}} + R_4 \cdot \frac{\overset{0}{\overset{\shortparallel}{a_1 w_{14}^+}}}{\underset{\underset{0}{\shortparallel}}{a_1 w_{14}^+} + \underset{\underset{1}{\shortparallel}}{a_2 w_{24}^+}} = R_3$$

(c)



$$R_6 = R_{out} = Y$$

$$R_5 = \frac{a_5 w_{56}^+}{\sum_{k=3}^{5} a_k w_{k6}} \cdot R_6 = \frac{0.5 \cdot a_5}{0.5 \cdot (a_3+a_4+a_5)} Y = \frac{a_5}{a_3+a_4+a_5} Y$$

$$R_4 = \frac{a_4 w_{46}^+}{\sum_{k=3}^{5} a_k w_{k6}} \cdot R_6 = \frac{0.5 \cdot a_4}{0.5 \cdot (a_3+a_4+a_5)} Y = \frac{a_4}{a_3+a_4+a_5} Y$$

$$R_3 = \frac{a_3 w_{36}^+}{\sum_{k=3}^{5} a_k w_{k6}} \cdot R_6 = \frac{0.5 \cdot a_3}{0.5 \cdot (a_3+a_4+a_5)} Y = \frac{a_3}{a_3+a_4+a_5} Y$$

$$R_2 = \sum_{k=3}^{5} \frac{a_2 w_{2k}^+}{\sum_{j=1}^{2} a_j w_{jk}^+} R_k = R_3 \cdot \frac{\overset{1}{\overset{\shortparallel}{a_2 w_{23}^+}}}{\underset{\underset{1}{\shortparallel}}{a_1 w_{13}^+} + \underset{\underset{1}{\shortparallel}}{a_2 w_{23}^+}} + R_4 \frac{\overset{0}{\overset{\shortparallel}{a_2 w_{24}^+}}}{\underset{\underset{1}{\shortparallel}}{a_1 w_{14}^+} + \underset{\underset{0}{\shortparallel}}{a_2 w_{24}^+}} + R_5 \frac{\overset{1}{\overset{\shortparallel}{a_2 w_{25}^+}}}{\underset{\underset{0}{\shortparallel}}{a_1 w_{15}^+} + \underset{\underset{1}{\shortparallel}}{a_2 w_{25}^+}}$$

$$= R_3 \cdot \frac{x_2}{x_1+x_2} + R_5$$

$$R_1 = \sum_{k=3}^{5} \frac{a_1 w_{1k}^+}{\sum_{j=1}^{2} a_j w_{jk}^+} R_k = R_3 \cdot \frac{\overset{1}{\overset{\shortparallel}{a_1 w_{13}^+}}}{\underset{\underset{1}{\shortparallel}}{a_1 w_{13}^+} + \underset{\underset{1}{\shortparallel}}{a_2 w_{23}^+}} + R_4 \frac{\overset{1}{\overset{\shortparallel}{a_1 w_{14}^+}}}{\underset{\underset{1}{\shortparallel}}{a_1 w_{14}^+} + \underset{\underset{0}{\shortparallel}}{a_2 w_{24}^+}} + R_5 \frac{\overset{0}{\overset{\shortparallel}{a_1 w_{15}^+}}}{\underset{\underset{0}{\shortparallel}}{a_1 w_{15}^+} + \underset{\underset{1}{\shortparallel}}{a_2 w_{25}^+}}$$

$$= R_3 \frac{x_1}{x_1+x_2} + R_4$$