Exercises for the course
Machine Learning 1
Winter semester 2021/22

Abteilung 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 $\widetilde{\boldsymbol{x}} = \mathbf{0}$, 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

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

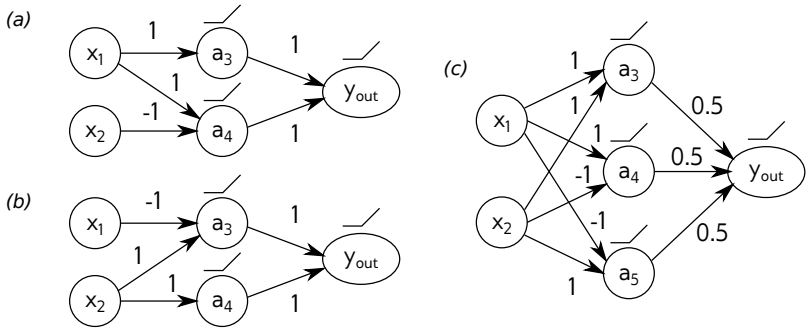
with $\theta > 0$. For the purpose of extracting an explanation, we would like to build a first-order Taylor expansion of the function at some root point $\widetilde{\boldsymbol{x}}$. We choose this root point to be taken on the segment connecting $\boldsymbol{\mu}$ and $\boldsymbol{x}$ (we assume that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

Sheet 14

1 The prototype can be found by setting the derivative of the objective to zero:

$$\max_x \left[f(x) + \Omega(x) \right] = \frac{\partial}{\partial x} \left(f(x) + \log p(x) \right)$$

$$= \frac{\partial}{\partial x} \left[w^T x + b + \log\left(a e^{-\frac{(x-\mu)^2}{2\Sigma^2}} \right) \right]$$

$$= \frac{\partial}{\partial x} \left[w^T x + b + \log(a) - \frac{(x-\mu)^2}{2\Sigma^2} \right]$$

$$= w^T - \frac{x(x-\mu)}{x\Sigma^2} \overset{!}{=} 0$$

$$\Leftrightarrow \quad w^T = \frac{x-\mu}{\Sigma^2} \quad \Leftrightarrow \quad w^T \Sigma^2 = x - \mu$$

$$\Leftrightarrow \quad x^* = w^T \Sigma^2 + \mu$$

2 The Shapley value ϕ_i is given by

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

For the Shapley value ϕ_1 there are four possible features:
$S_1: \emptyset$, $S_2: \{2\}$, $S_3: \{3\}$, $S_4: \{2,3\}$. Therefore, the value of S is $S_1 = 0$, $S_2 = 1$, $S_3 = 1$, $S_4 = 2$. The dimension d of $f(x_1, x_2, x_3)$ is $d = 3$. This yields for α_{S_1}:

$$\alpha_{S_1} = \frac{|S|!\,(d-|S|-1)!}{d!} = \frac{|0|!\,(3-|0|-1)!}{3!} = \frac{1 \cdot 2!}{3!} = \frac{1}{3}$$

And $\alpha_{S_2} = \frac{|1|!\,(3-|1|-1)!}{3!} = \frac{1}{6}$

And $\alpha_{S_3} = \frac{1}{6}$

And $\alpha_{S_4} = \frac{|2|!\,(3-|2|-1)!}{3!} = \frac{1}{3}$

To obtain the corresponding values ΔS_i, the function $f(x_1, x_2, x_3)$ needs to be calculated for all features.

In the first case, $S: \emptyset$ features $x_2 = x_3 = 0$:

$S_1: \emptyset \to f(0,0,0) = \min(0, \text{more}(0,0)) = 0 = \Delta S_1$

For $S_2: \{2\}$, feature $x_2 = 1$:

$\Delta S_2 = f(1,1,0) = \min(1, \text{more}(1,0)) = 1$

For $S_3: \{3\}$, feature $x_3 = 1$:

$\Delta S_3 = f(1,0,1) = 1$

For $S_4: \{2,3\}$, features $x_2 = x_3 = 1$:

$\Delta S_4 = f(1,1,1) = 1$

Therefore, $\phi_1 = \sum_{S_i \not\ni 3} \alpha_{S_i} \cdot \Delta S_i$

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

Since $\sum \phi_i = \phi_1 + \phi_2 + \phi_3 = \frac{2}{3} + \phi_2 + \phi_3 \overset{!}{=} 1$, we know

$\Leftrightarrow \phi_2 + \phi_3 = \frac{1}{3}$.

The function $f(x_1, x_2, x_3) = f(x_1, x_3, x_2)$ since $\text{more}(x_2, x_3) = \text{more}(x_3, x_2)$. For this reason, we conclude $\phi_2 = \phi_3 \Leftrightarrow \phi_2 = \frac{1}{6}, \phi_3 = \frac{1}{6}$

This yields $\phi_1 = \frac{2}{3}$, $\phi_2 = \frac{1}{6}$, $\phi_3 = \frac{1}{6}$.

3 The function is given as $f(\vec{x}) = \|\vec{x} - \vec{\mu}\| - \theta$

The Taylor expansion is $f(\vec{x}) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\tilde{\vec{x}})}{n!} (x - \tilde{\vec{x}})^n$

$f^{(0)}(\tilde{\vec{x}}) = f(\tilde{\vec{x}}) = \|\tilde{\vec{x}} - \mu\| - \theta$

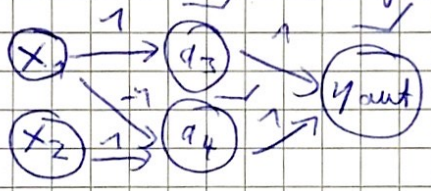
1) $f^{(1)}(x) = f'(\tilde{\vec{x}})(x - \tilde{\vec{x}}) = \frac{x - \mu}{\|\vec{x} - \vec{\mu}\|} (\vec{x} - \tilde{\vec{x}})$

$\phi_i = \frac{(x_i - \mu_i)^2}{\|\vec{x} - \vec{\mu}\|^2} (\|x - \mu\| - \theta)$ can be obtained

by setting $\tilde{\vec{x}} = \mu + \theta \frac{x - \mu}{\|\vec{x} - \vec{\mu}\|}$ in 1).

4 We calculate the relevance scores R_i for all nodes of the networks using $R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$

a)



$y_{out} = q_3 + q_4$

$R_5 = R_{out} = y_{out}$

$R_4 = \sum_{j=3}^{4} \frac{a_4 \cdot w_{44}}{\sum_{j=3}^{4} a_j w_{j4}} y_{out} = \frac{a_4 \cdot w_{44}}{a_3 \cdot w_{34} + a_4 w_{44}} y_{out} = \frac{q_4}{q_3 + q_4} y_{out}$

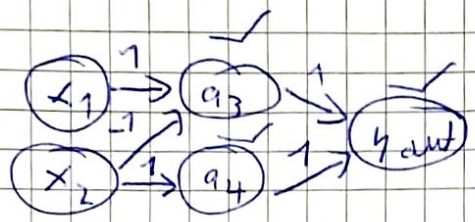
$R_3 = \sum_{j=3}^{4} \frac{a_3 w_{34}}{\sum_{j=3}^{4} a_j w_{j4}} y_{out} = \frac{a_3}{a_3 + a_4} y_{out} = q_3$

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

$= \frac{x_2 \cdot 0}{x_1 \cdot 0 + x_2 \cdot 0} R_3 + \frac{x_2 \cdot 1}{x_1 \cdot 0 + x_2 \cdot 1} R_4 = \frac{x_2}{x_2} R_4 = R_4$

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

b)



$a_3 + a_4 = y_{out}$

$R_5 = y_{out}$

$$R_4 = \sum_{j=1}^{1} \frac{a_4 w_{45}^+}{a_4 w_{45}^+ + a_3 w_{35}^+} \quad R_5 = \frac{a_4 \cdot 1}{a_4 \cdot 1 + a_3 \cdot 1} R_5 = \frac{a_4}{a_3 + a_4} y_{out} = a_4$$

$$R_3 = \sum_{j=1}^{1} \frac{a_3 w_{35}^+}{a_3 w_{35}^+ + a_4 w_{45}^+} \quad R_5 = \frac{a_3}{a_3 + a_4} y_{out} = a_3$$

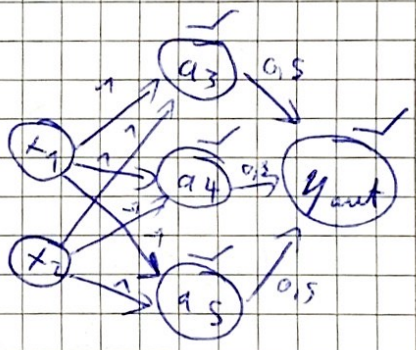
$$R_2 = \sum_{h=3}^{4} \frac{a_2 w_{2h}^+}{\sum_{j=1}^{2} a_j w_{jh}^+} R_h = \frac{a_2 w_{23}^+ R_3}{a_1 w_{13}^+ + a_2 w_{23}^+} + \frac{a_2 w_{24}^+ R_4}{a_1 w_{14}^+ + a_2 w_{24}^+} = R_4$$

($a_2 w_{23}^+ = 0$, $a_2 w_{24}^+ = 1$, $a_1 w_{14}^+ = 0$)

$$R_1 = \sum_{h=3}^{4} \frac{a_1 w_{1h}^+}{\sum_{j=1}^{2} a_j w_{jh}^+} R_h = \frac{a_1 w_{13}^+ R_3}{a_1 w_{13}^+ + a_2 w_{23}^+} + \frac{a_1 w_{14}^+ R_4}{a_1 w_{14}^+ + a_2 w_{24}^+} = R_3$$

($a_1 w_{13}^+ = 1$, $a_2 w_{23}^+ = 0$, $a_1 w_{14}^+ = 0$)

c) $R_6 = y_{out}$



$$R_5 = \sum_{j=5}^{1} \frac{a_5 w_{5y}^+ R_y}{\sum_{j=5}^{5} a_j w_{jy}^+} = \frac{a_5 \cdot 0{,}5 \cdot y_{out}}{a_3 \cdot 0{,}5 + a_4 \cdot 0{,}5 + a_5 \cdot 0{,}5} = 0{,}5 \cdot a_5$$

$$R_4 = \frac{a_4 \cdot 0{,}5 \cdot y_{out}}{a_3 \cdot 0{,}5 + a_4 \cdot 0{,}5 + a_5 \cdot 0{,}5} = 0{,}5 \cdot a_4 \qquad R_3 = 0{,}5 \cdot a_3$$

$$R_2 = \sum_{h=3}^{5} \frac{a_2 w_{2h}^+ R_h}{\sum_{j=1}^{2} a_j w_{jh}^+} = \frac{a_2 w_{23}^+ R_3}{a_1 w_{13}^+ + a_2 w_{23}^+} + \frac{a_2 w_{24}^+ R_4}{a_1 w_{14}^+ + a_2 w_{24}^+} + \frac{a_2 w_{25}^+ R_5}{a_1 w_{15}^+ + a_2 w_{25}^+}$$

($a_2 w_{23}^+ = 1$, $a_1 w_{13}^+ = 1$, $a_2 w_{24}^+ = 0$, $a_1 w_{15}^+ = 0$, $a_2 w_{25}^+ = 1$)

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

$$R_1 = \frac{a_1 w_{13}^+ R_3}{a_1 w_{13}^+ + a_2 w_{23}^+} + \frac{a_1 w_{14}^+ R_4}{a_1 w_{14}^+ + a_2 w_{24}^+} + \frac{a_1 w_{15}^+ R_5}{a_1 w_{15}^+ + a_2 w_{25}^+} = \frac{x_1}{x_1 + x_2} R_3 + R_4$$

($a_1 w_{13}^+ = 1$, $a_2 w_{23}^+ = 1$, $a_1 w_{14}^+ = 1$, $a_2 w_{24}^+ = 0$, $a_1 w_{15}^+ = 0$, $a_2 w_{25}^+ = 1$)