Exercises for the course
Machine Learning 1
Winter semester 2024/25

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

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

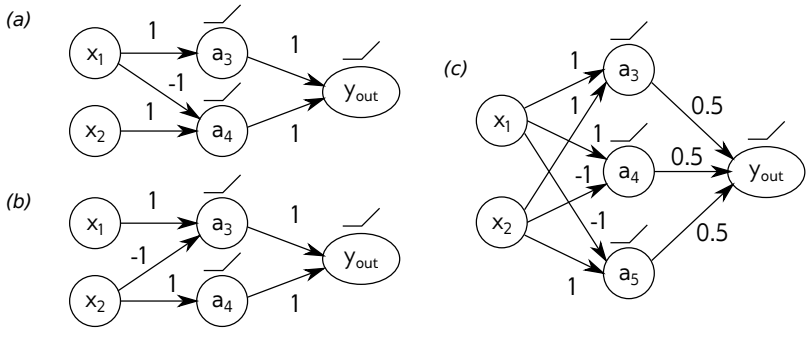
with $\theta > 0$. For the purpose of extracting an explanation, we would like to build a first-order Taylor expansion of the function at some root point $\widetilde{\boldsymbol{x}}$. We choose this root point to be taken on the segment connecting $\boldsymbol{\mu}$ and $\boldsymbol{x}$ (we assume that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