Exercises for the course

Machine Learning 2

Summer semester 2017

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 11

In this exercise sheet, we refer to the sections of the paper "*Methods for interpreting and understanding deep neural networks*" linked via ISIS.

Exercise 1: Experts and Prototypes (40 P)

Consider the linear model $y = \boldsymbol{w}^\top \boldsymbol{x} + b$ mapping some input $\boldsymbol{x}$ to an output y. We would like to interpret the output y by building a prototype $\boldsymbol{x}^\star$ in the input domain following the activation maximization techniques outlined in Section 3.

(a) *Find* the prototype $\boldsymbol{x}^\star$ obtained by activation maximization as formulated in Section 3.1.

(b) *Find* the prototype $\boldsymbol{x}^\star$ obtained by activation maximization as formulated in Section 3.2. We assume that the data is represented by the Gaussian expert $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ and Σ are the mean and covariance.

(c) *Find* the prototype $\boldsymbol{x}^\star$ obtained by activation maximization as formulated in Section 3.3. The data is generated as (i) $\boldsymbol{z} \sim \mathcal{N}(0, I)$ and (ii) $\boldsymbol{x} = A\boldsymbol{z} + \boldsymbol{c}$, where A and $\boldsymbol{c}$ are the parameters of the generator.

(d) *Relate* the prototypes obtained for the three approaches above, in particular under which regularizers, experts and generator, the found prototypes are mutually equivalent.

Exercise 2: Sensitivity Analysis and Taylor Decomposition (30 P)

Let us consider a data point $\boldsymbol{x}$ and its prediction by a homogeneous linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. We would like to explain the prediction using the methods described in Section 4.

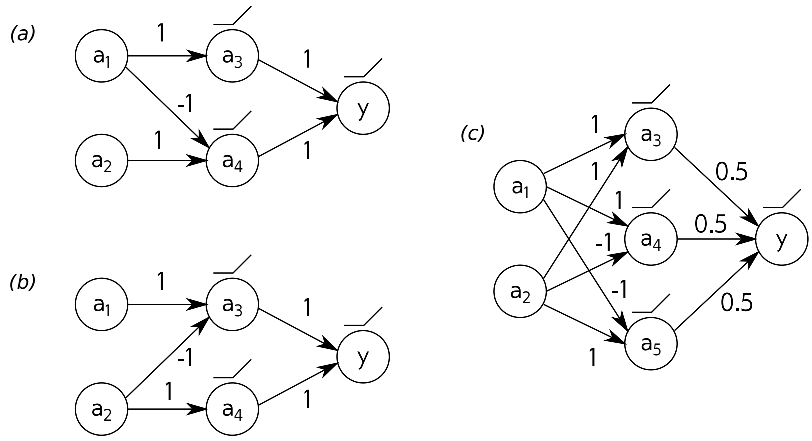
(a) *Compute* the explanation for the prediction $f(\boldsymbol{x})$ using sensitivity analysis as described in Section 4.1.

(b) *Compute* the explanation for the prediction $f(\boldsymbol{x})$ using Taylor decomposition (Section 4.2) at root point $\widetilde{\boldsymbol{x}} = \mathbf{0}$.

(c) *Compute* the explanation for the prediction $f(\boldsymbol{x})$ using Taylor decomposition at root point $\widetilde{\boldsymbol{x}}$ chosen to be the nearest (in the Euclidean sense) from $\boldsymbol{x}$. (Hint: You can use the Lagrange multipliers to find this root point.)

Exercise 3: Layer-Wise Relevance Propagation (30 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 $y = \max(a_1, a_2)$, where $a_1, a_2 \in \mathbb{R}^+$ are the input activations. This function can be implemented as a ReLU network in multiple ways. Three examples are given below.



Because of the positive activations, an appropriate rule for both layers is LRP-$\alpha_1\beta_0$ defined in Section 5.1.

(a) *Give* for each network an analytic solution for the obtained scores R_1 and R_2 obtained by application this propagation rule at each layer.

(b) *Discuss* which implementation of the "max" function (a, b, or c) gives the most intuitive explanations.