G. Montavon
Institute of Computer Science
Department of Mathematics and Computer Science
Freie Universität Berlin

Exercise Sheet 12 (theory part)

Exercise 1: Activation Maximization (10 + 10 + 10 P)

Consider the linear model $f\colon \mathbb{R}^d \to \mathbb{R}$ with $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$. We would like to interpret the function f by building a prototype $\boldsymbol{x}^\star$ in the input domain which produces a large value of f. Activation maximization produces such prototype by computing

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} \big[f(\boldsymbol{x}) - \Omega(\boldsymbol{x})\big].$$

where Ω is a regularization function.

(a) *Find* the prototype $\boldsymbol{x}^\star$ obtained by activation maximization subject to the penalty $\Omega(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|^2$ where λ is a hyperparameter.

$$\frac{\partial}{\partial \boldsymbol{x}}\big(\boldsymbol{w}^\top \boldsymbol{x} + b - \lambda\|\boldsymbol{x}\|^2\big) = \boldsymbol{w} - 2\lambda\boldsymbol{x} = 0 \quad\Rightarrow\quad \boldsymbol{x}^\star = \frac{1}{2\lambda}\boldsymbol{w}$$

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

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{x}}\big(\boldsymbol{w}^\top \boldsymbol{x} + b - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \text{const.}\big)\\ &\qquad = \boldsymbol{w} - \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = 0 \quad\Rightarrow\quad \boldsymbol{x}^\star = \Sigma\boldsymbol{w} + \boldsymbol{\mu}\end{aligned}$$

(c) *Find* the prototype $\boldsymbol{x}^\star$ obtained when the data is generated as (i) $\boldsymbol{z} \sim \mathcal{N}(0, I)$ and (ii) $\boldsymbol{x} = A\boldsymbol{z} + \boldsymbol{c}$, with A and $\boldsymbol{c}$ the parameters of the generator. Here, we optimize f w.r.t. the code $\boldsymbol{z}$ subject to the penalty $\Omega(\boldsymbol{z}) = \lambda\|\boldsymbol{z}\|^2$.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{z}}\big(\boldsymbol{w}^\top(A\boldsymbol{z} + \boldsymbol{c}) + b - \lambda\|\boldsymbol{z}\|^2\big) = A^\top\boldsymbol{w} - 2\lambda\boldsymbol{z} = 0 \quad&\Rightarrow\quad \boldsymbol{z}^\star = \frac{A^\top}{2\lambda}\boldsymbol{w}\\ &\Rightarrow\quad \boldsymbol{x}^\star = \frac{AA^\top}{2\lambda}\boldsymbol{w} + \boldsymbol{c}\end{aligned}$$

Exercise 2: Attribution (10 + 10 P)

Consider the function $f : \mathbb{R}_+^3 \to \mathbb{R}_+$ with

$$f(\boldsymbol{x}) = \min(x_1, \max(x_2, x_3))$$

implementing some min-max pooling between three input features. For the data point $\boldsymbol{x} = (3, 2, 1)$ and its prediction $f(\boldsymbol{x}) = 2$, we would like to perform an attribution of the prediction to the input features. We investigate the Shapley value and Gradient $\times$ Input methods for attribution.

(a) Recall that the Shapley value method identifies the contributions $R_1, \dots, R_d$ of features $x_1, \dots, x_d$ as:

$$R_i = \sum_{\mathcal{S}\colon i \notin \mathcal{S}} \tfrac{|\mathcal{S}|!(d-|\mathcal{S}|-1)!}{d!}\big[f(\boldsymbol{x}_{\mathcal{S}\cup\{i\}}) - f(\boldsymbol{x}_{\mathcal{S}})\big]$$

where $(\boldsymbol{x}_{\mathcal{S}})_{\mathcal{S}}$ are all possible subsets of features contained in the input $\boldsymbol{x}$. *Compute* the Shapley values associated to the prediction above. (We assume a reference point $\widetilde{\boldsymbol{x}} = \mathbf{0}$, i.e. we set features to zero when removing them).

We first evaluate the function, and we get $f(3,2,1) = 2$.

Consider R_1, and analyze all coalitions that don't contain that feature:

$\mathcal{S}$	$\alpha_\mathcal{S}$	$f(\boldsymbol{x}_{\mathcal{S}\cup\{1\}}) - f(\boldsymbol{x}_\mathcal{S})$
$\varnothing$	1/3	0
$\{2\}$	1/6	2
$\{3\}$	1/6	1
$\{2,3\}$	1/3	2

Therefore $R_1 = 1/3 \cdot 0 + 1/6 \cdot 2 + 1/6 \cdot 1 + 1/3 \cdot 2 = 7/6$.

Consider now R_2, and analyze all coalitions that don't contain that feature:

$\mathcal{S}$	$\alpha_\mathcal{S}$	$f(\boldsymbol{x}_{\mathcal{S}\cup\{2\}}) - f(\boldsymbol{x}_\mathcal{S})$
$\varnothing$	1/3	0
$\{1\}$	1/6	2
$\{3\}$	1/6	0
$\{1,3\}$	1/3	1

Therefore $R_2 = 1/3 \cdot 0 + 1/6 \cdot 2 + 1/6 \cdot 1 + 1/3 \cdot 1 = 2/3$.

Consider now R_3, and analyze all coalitions that don't contain that feature:

$\mathcal{S}$	$\alpha_\mathcal{S}$	$f(\boldsymbol{x}_{\mathcal{S}\cup\{2\}}) - f(\boldsymbol{x}_\mathcal{S})$
$\varnothing$	1/3	0
$\{1\}$	1/6	1
$\{2\}$	1/6	0
$\{1,2\}$	1/3	0

Therefore $R_3 = 1/3 \cdot 0 + 1/6 \cdot 1 + 1/6 \cdot 1 + 1/3 \cdot 0 = 1/6$.

We can test our result by verifying the conservation property:

$$R_1 + R_2 + R_3 = 7/6 + 2/3 + 1/6 = 12/6 = f(\boldsymbol{x})$$

(b) The Gradient $\times$ Input method attributes to the input features according to the formula:

$$R_i = x_i \cdot [\nabla f(\boldsymbol{x})]_i$$

Compute the Gradient $\times$ Input attribution associated to the prediction above.

$$R_1 = 3 \cdot 0 = 0$$
$$R_2 = 2 \cdot 1 = 2$$
$$R_3 = 1 \cdot 0 = 0$$