Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 9 - Bonus

Exercise 1: Analysis of a similarity models (0 P)

We consider here similarity models of type $y(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$ with the dot product on a feature map $\phi\colon \mathbb{R}^d \to \mathbb{R}^h$ and satisfying first-order positive homogeneity i.e. $\forall_{\boldsymbol{x}}, \forall_{t>0} : \phi(t\boldsymbol{x}) = t\phi(t\boldsymbol{x})$.
In the following we focus on Linear/ReLU layers:

$$
\begin{aligned}
a_k &= \left(\textstyle\sum_j a_j w_{jk}\right)^+ \\
a_{k'} &= \left(\textstyle\sum_{j'} a_{j'} w_{j'k'}\right)^+,
\end{aligned}
$$

with activations a_j and weights w_{jk} and $(\cdot)^+$ indicating the ReLU function.
Further assume root points $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') = (\varepsilon\widetilde{\boldsymbol{x}}, \varepsilon\widetilde{\boldsymbol{x}}')$ with ε almost zero.

(a) Write down the Taylor expansion of function $y(\boldsymbol{x}, \boldsymbol{x}')$ up to second-order terms.

$$
\begin{aligned}
y(\boldsymbol{x}, \boldsymbol{x}') = y(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') \\
&+ \textstyle\sum_i [\nabla y(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}')]_i \,(x_i - \widetilde{x}_i) \\
&\quad + \textstyle\sum_{i'} [\nabla y(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}')]_{i'} \,(x'_{i'} - \widetilde{x}'_{i'}) \\
&\qquad + \textstyle\sum_{ii'} [\nabla^2 y(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}')]_{ii'} \,(x_i - \widetilde{x}_i)\,(x'_{i'} - \widetilde{x}'_{i'})
\end{aligned}
$$

(b) Analyse zero-order terms. Why do they vanish?
With $\phi(t\boldsymbol{x}) = t\phi(t\boldsymbol{x})$, and $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') = (\varepsilon\widetilde{\boldsymbol{x}}, \varepsilon\widetilde{\boldsymbol{x}}')$ with ε almost zero we see that zero-order terms vanish and $y(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}') = 0$.
Now, assume the following propagation rule for the Linear/ReLU layer to identify relevant interaction between a pair of neurons j and j':

$$
\begin{aligned}
R_{jj'} &= \textstyle\sum_{kk'} R_{jj' \leftarrow kk'} \\
&= \sum_{kk'} \frac{a_j a_{j'} \rho(w_{jk}) \rho(w_{j'k'})}{\sum_{jj'} a_j a_{j'} \rho(w_{jk}) \rho(w_{j'k'})} R_{kk'}
\end{aligned}
$$

(i) If neurons (j, j') jointly activate, i.e. $a_j a_{j'}$ is non-zero.

(ii) If pairs of neurons in the layer above jointly react, i.e. $R_{kk'}$ is non-zero (or relevant).

(iii) If these reacting pairs are themselves relevant, i.e. the term $a_j a_{j'} \rho(w_{jk}) \rho(w_{j'k'})$ is non-zero.

(c) Show that $R_{jj'}$ factorizes as $R_{jj'} = \sum_{m=1}^h R_{jm} R_{j'm}$. Use the factorization of the subsequent layer $R_{kk'} = \sum_{m=1}^h R_{km} \cdot R_{k'm}$.

$$
\begin{aligned}
R_{jj'} &= \sum_{kk'} \frac{a_j a_{j'} \rho(w_{jk}) \rho(w_{j'k'})}{\sum_{jj'} a_j a_{j'} \rho(w_{jk}) \rho(w_{j'k'})} \sum_{m=1}^h R_{km} R_{k'm} \\
&= \sum_{m=1}^h \sum_{kk'} \frac{a_j \rho(w_{jk}) a_{j'} \rho(w_{j'k'})}{\sum_j a_j \rho(w_{jk}) \sum_{j'} a_{j'} \rho(w_{j'k'})} R_{km} R_{k'm} \\
&= \sum_{m=1}^h \Bigg(\underbrace{\sum_k \frac{a_j \rho(w_{jk})}{\sum_j a_j \rho(w_{jk})} R_{km}}_{R_{jm}} \Bigg) \cdot \Bigg(\underbrace{\sum_{k'} \frac{a_{j'} \rho(w_{j'k'})}{\sum_{j'} a_{j'} \rho(w_{j'k'})} R_{k'm}}_{R_{j'm}} \Bigg)
\end{aligned}
$$