TECHNISCHE UNIVERSITÄT MÜNCHEN
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Datenverarbeitung
PD Dr. Martin Kleinsteuber

Information Retrieval in High Dimensional Data

Lab #6: Theoretical Exercises, 14.06.2018

## Neural Networks

Task 1. Recall Logistic Regression. We determined the probability of a sample $\mathbf{x} \in \mathbb{R}^p$ belonging to class $y \in \{-1, 1\}$ as

$$Pr(Y = y|X = \mathbf{x}) = \sigma(y(\mathbf{w}^\top \mathbf{x} + b)),$$

where $\sigma$ denotes the sigmoid function and $\mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}$ are trainable parameters of the model. In the following, we will extend the model to

$$Pr(Y = y|X = \mathbf{x}) = \sigma(y f_L(\cdots \mathrm{relu}(f_2(\mathrm{relu}(f_1(\mathbf{x})))) \cdots)). \tag{1}$$

where the functions $f_l$ are affine, i.e. of the form

$$f_l(\mathbf{z}) = \mathbf{W}_l \mathbf{z} + \mathbf{b}_l,$$

with trainable parameters $\mathbf{W}_l \in \mathbb{R}^{m_l \times n_l}, \mathbf{b}_l \in \mathbb{R}^{m_l}$. The dimensions have the following properties. $n_1 = p$, $m_L = 1$ and $m_l = n_{l+1}$. Note that for $L = 1$, this is the classical logistic regression model. Given a training set $\{(\mathbf{x}_i, y_i)\}_{i \in \{1,\ldots,N\}}$, consider that the log-likelihood $C$ of the probability model (1) has the form

$$C = -\sum_{i=1}^{N} \log(1 + \exp(-y_i(f_L(\cdots \mathrm{relu}(f_2(\mathrm{relu}(f_1(\mathbf{x}_i)))) \cdots)))).$$

a) Let us denote by $z_l(\mathbf{x})$ the output of the $l$-th layer, i.e.

$$z_l(\mathbf{x}) = \begin{cases} \mathbf{x} & \text{if } l = 0 \\ \mathrm{relu}(f_l(z_{l-1}(\mathbf{x}))) & \text{if } 0 < l < L \end{cases}$$

Show that the derivative of $C$ w.r.t, the bias vector of a layer is given by the following equation (For $l = L$, the product is replaced by a 1).

$$\nabla_{\mathbf{b}_l} C = \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} \left( \prod_{k=0}^{L-l-1} \mathbf{W}_{L-k} \mathrm{diag}(\mathrm{step}(z_{L-k-1}(\mathbf{x}_i))) \right)^\top,$$

where step is an elementwise function defined as follows.

$$\mathrm{step}(\mathbf{x})_j = \begin{cases} 0 & \text{if } x_j \leq 0, \\ 1 & \text{otherwise.} \end{cases}$$

**Solution**: Let us denote the Jacobi matrix of a vector-valued function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ by

$$J_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}. \tag{2}$$

For $l = L$, we write

$$C = \sum_{i=1}^{N} \log(1 + \exp(-y_i(\mathbf{W}_L z_{L-1}(\mathbf{x}_i) + \mathbf{b}_L))) \tag{3}$$

and observe

$$J_C(\mathbf{b}_L) = \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i(\mathbf{W}_L z_{L-1}(\mathbf{x}_i) + \mathbf{b}_L))}. \tag{4}$$

For $l < L$, we write

$$C = \sum_{i=1}^{N} \log(1 + \exp(-y_i(f_L(\mathrm{relu}(f_{L-1}(\cdots f_{l+1}(\mathrm{relu} f_l(z_{l-1}(\mathbf{x}))))))))).$$

With

$$J_{f_l \circ \mathrm{relu}}(z_{l-1}(\mathbf{x})) = \mathbf{W}_l \mathrm{diag}(\mathrm{step}(z_{l-1}(\mathbf{x}))),\ 1 \leq l \leq L \tag{5}$$

and

$$J_{f_{l+1}}(\mathbf{b}_l) = \mathbf{I}_{m_l}$$

the chain rule yields

$$\begin{aligned} J_C(\mathbf{b}_l) &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} J_{f_L \circ \mathrm{relu}}(z_{L-1}(\mathbf{x})) \\ &\quad \cdots J_{f_{l+1} \circ \mathrm{relu}}(z_l(\mathbf{x})) J_{f_l}(\mathbf{b}_l) \\ &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} \mathbf{W}_L \mathrm{diag}(\mathrm{step}(z_{L-1}(\mathbf{x}_i)) \\ &\quad \cdots \mathbf{W}_{l+1} \mathrm{diag}(\mathrm{step}(z_l(\mathbf{x}_i)) \mathbf{I}_{m_l}, \end{aligned}$$

which is the transpose of $\nabla_{\mathbf{b}_l} C$.

b) Show that the derivative of C w.r.t. weight matrix of a layer is given by the following equation (For $l = L$, the product is replaced by a 1).

$$\nabla_{\mathbf{W}_l} C = \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} \cdot \left( \prod_{k=0}^{L-l-1} \mathbf{W}_{L-k} \operatorname{diag}(\operatorname{step}(z_{L-k-1}(\mathbf{x}_i))) \right)^\top (z_{l-1}(\mathbf{x}_i))^\top.$$

**Solution:** For $l = L$, we consider again (3). The weight matrix $\mathbf{W}_L$ has only one row, so the Jacobi matrix is well defined:

$$\begin{aligned} J_C(\mathbf{W}_L) &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} J_{f_L}(\mathbf{W}_L) \\ &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} (z_{l-1}(\mathbf{x}_i))^\top. \end{aligned}$$

Since $\mathbf{W}_l$ is a matrix with possibly several rows and columns for $l < L$, we can not write down the Jacobi matrix $J_C(\mathbf{W}_l)$for $l < L$ without either vectorizing $\mathbf{W}_l$ or using Tensor notation. An alternative is to consider the Jacobi matrix w.r.t. to the $k$-th row of $\mathbf{W}_l$ which we denote by $\mathbf{W}_l^k$. We get

$$\begin{aligned} J_C(\mathbf{W}_l^k) &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} J_{f_L \circ \mathrm{relu}}(z_{L-1}(\mathbf{x})) \\ &\quad \cdots J_{f_{l+1} \circ \mathrm{relu}}(z_l(\mathbf{x})) J_{f_l}(\mathbf{W}_l^k) \\ &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x}_i)))} \mathbf{W}_L \operatorname{diag}(\operatorname{step}(z_{L-1}(\mathbf{x}_i)) \\ &\quad \cdots \mathbf{W}_{l+1} \operatorname{diag}(\operatorname{step}(z_l(\mathbf{x}_i)) J_{f_l}(\mathbf{W}_l^k) \end{aligned}.$$

For the last factor, we observe

$$J_{f_l}(\mathbf{W}_l^k) = \mathbf{e}_k (z_{l-1}(\mathbf{x}_i))^\top,$$

where $\mathbf{e}_k$ denotes the $k$-th unit vector. This means that the $k$-th row of $\nabla_{\mathbf{W}_L} C$ is given by

$$\begin{aligned} J_C(\mathbf{W}_l^k) &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x})))} J_{f_L \circ \mathrm{relu}}(z_{L-1}(\mathbf{x})) \\ &\quad \cdots J_{f_{l+1} \circ \mathrm{relu}}(z_l(\mathbf{x})) J_{f_l}(\mathbf{W}_l^k) \\ &= \sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i f_L(z_{L-1}(\mathbf{x})))} \mathbf{W}_L \operatorname{diag}(\operatorname{step}(z_{L-1}(\mathbf{x})) \\ &\quad \cdots \mathbf{W}_{l+1} \operatorname{diag}(\operatorname{step}(z_l(\mathbf{x})) \mathbf{e}_k (z_{l-1}(\mathbf{x}_i))^\top \end{aligned}$$

Note that

$$\sum_{i=1}^{N} \frac{y_i}{1+\exp(y_i f_L(z_{L-1}(\mathbf{x})))} \mathbf{W}_L \mathrm{diag}(\mathrm{step}(z_{L-1}(\mathbf{x})) \cdots \mathbf{W}_{l+1} \mathrm{diag}(\mathrm{step}(z_l(\mathbf{x})) \tag{6}$$

is a row vector and the multiplication with $\mathbf{e}_k$ is simply extracting its $k$-th component. If $J_C(\mathbf{W}_l^k)$ is the $k$-th row of $\nabla_{\mathbf{W}_l} C$, then we can write $\nabla_{\mathbf{W}_l} C$ by transposing (6) and multiplying the result with $(z_{l-1}(\mathbf{x}_i))^\top$.