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 σ 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

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

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 $\{(\mathbf{x}_i, y_i)\}_{i \in \{1,\dots,N\}}$, consider that the loss C given by the log-likelihood of the probability model (1) has the form

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

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

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

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).

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

where pos is an elementwise function defined as follows.

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

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).

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