Exercises
Deep Learning
Fall 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Fernando Pérez-Cruz
Web http://www.da.inf.ethz.ch/teaching/2018/DeepLearning/

Series Monday, Oct 8, 2018
(Deep Learning, Exercise series 2)

Problem 1 (Activation functions):

1. Consider the sigmoid function $s(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$ acting element-wise on a vector x, which is a common activation function used in neural networks. Prove that

$$\nabla_x s(x) = s(x)(1 - s(x)).$$

2. Similarly, derive the gradients of the following functions:

$$\begin{aligned}
\mathsf{ReLU}(x) &= \max(0, x) \\
\mathsf{tanh}(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} \\
\mathsf{softmax}(x)_i &= \frac{e^{x_i}}{\sum_k e^{x_k}} \\
\mathsf{tanh}\left(\mathsf{softmax}(x)_i\right)
\end{aligned}$$

Problem 2 (Cross-entropy):

Consider a binary classification problem for which we are given a set of input vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ and we want to predict an output variable $y_i = \{0, +1\}$ for each input $\boldsymbol{x}_i$. We use a neural network with parameters $\boldsymbol{w}$ and a cross-entropy loss function defined as

$$H(\boldsymbol{w}) = \sum_{i=1}^{n} -y_i \log \hat{y}_i(\boldsymbol{x}_i; \boldsymbol{w}) - (1 - y_i) \log(1 - \hat{y}_i(\boldsymbol{x}_i; \boldsymbol{w})), \tag{1}$$

where $\hat{y}_i(\boldsymbol{w}) = \Pr(\boldsymbol{x}_i; \boldsymbol{w})$ is the output of the neural network.

1. Show that maximizing the log-likelihood is equivalent to minimizing $H(\boldsymbol{w})$. Recall that the likelihood is defined as:

$$\mathcal{L}(\boldsymbol{w}) = \prod_{i=1}^{n} \Pr(x_i; \boldsymbol{w})^{y_i}((1 - \Pr(x_i; \boldsymbol{w}))^{1-y_i}) \tag{2}$$

Problem 3 (Finite differences):

1. One common way to check that the computation of a derivative is correct is to use finite differences. Considering only the i-th dimension of the parameter vector $\boldsymbol{w}$, show that finite difference yields an error $O(\epsilon)$, i.e.

$$\nabla f(w_i) = \frac{f(w_i + \epsilon) - f(w_i)}{\epsilon} + O(\epsilon), \tag{3}$$

   where $\epsilon \in \mathbb{R}^+ \leq 1$.

2. The accuracy of the finite difference method can be improved significantly by using symmetrical central differences.

$$\nabla f(w_i) \approx \frac{f(w_i + \epsilon) - f(w_i - \epsilon)}{2\epsilon} \tag{4}$$

   What approximation error do we get using Eq. 4?

Problem 4 (Deep linear networks):

In the lecture you have seen fully connected layers that combine a linear map with a non-linear activation function $\sigma(\cdot)$ e.g.

$$G_i(\boldsymbol{x}) = \sigma(\boldsymbol{W}_i \boldsymbol{x} + \mathbf{b}_i),$$

where $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{W}_i \in \mathbb{R}^{k \times d}, \mathbf{b}_i \in \mathbb{R}^k$.
Here we want to show that the expressiveness of linear fully connected layers (i.e. σ is the identity function), in contrast to non-linear ones, does not increase with depth.

Let $\mathcal{G}$ be the set of all such linear functions. For $g_1, g_2 \in \mathcal{G}$, show that $g_1 \circ g_2$ is equivalent to any function in $g \in \mathcal{G}$.