Machine Learning I: Foundations
Exercise Sheet 4

Prof. Marius Kloft TA:Billy Joe Franks

19.05.2020

Deadline: 26.05.2020

1) (MANDATORY) 10 Points

In this exercise, we will be deriving everything necessary for the gradient of a neural network. Later in the lecture you will see these terms in action. Calculate the following derivatives for $a, y \in \mathbb{R}, \mathbf{x}, \mathbf{b}, \mathbf{w} \in \mathbb{R}^d, W \in \mathbb{R}^{d\times d}$:

a) $\frac{\partial}{\partial W_{i,j}}(W\mathbf{x}+\mathbf{b})$.

b) $\frac{\partial}{\partial b_i}(W\mathbf{x}+\mathbf{b})$.

c) $\frac{\partial}{\partial x_i}(W\mathbf{x}+\mathbf{b})$.

d) $\frac{\partial}{\partial a}\sigma(a) = \frac{\partial}{\partial a}\left(1+e^{-a}\right)^{-1}$.

e) $\frac{\partial}{\partial a}\log\left(1+e^{-ya}\right)$.

f) Use the above to calculate $\frac{\partial}{\partial W_{i,j}}\log\left(1+\exp(-y\mathbf{w}^T\hat{\sigma}(W\mathbf{x}+\mathbf{b}))\right)$. $\hat{\sigma}$ is the elementwise application of σ, i.e.

$$\hat{\sigma}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\right) = \begin{bmatrix}\sigma(x_1)\\\sigma(x_2)\end{bmatrix}$$

2) We will explore activation functions for neural networks. For each function do the following, plot it, state the range of its output (i.e., $f(\mathbb{R})$), derive its gradient, and state an advantage and disadvantage of using it.

a) **TanH**: $f(x) = \tanh(x)$

b) **Error function**: $f(x) = \operatorname{erf}(x)$

c) **ReLU**: $f(x) = \max(0, x)$

d) **Leaky-ReLU**: $f(x) = \max(0.01x, x)$

e) **PLU**: $f(x) = \max\left(\alpha(x+c)-c, \min\left[\alpha(x-c)+c, x\right]\right)$

f) **Sinusoid**: $f(x) = \sin(x)$

g) **Gaussian**: $f(x) = e^{-x^2}$

3) The aim of this question is to show the similarities between the hinge loss and logistic loss in optimization. For the SVM problem we considered using gradient descent in slide 10 (Lecture 3.3). We proposed two such methods one using the hinge loss $h(\mathbf{x}_i) = \max(0, 1 - y_i(\mathbf{w}^T \mathbf{x}_i))$ and the other using the logistic loss function:

$$l(x_i) = \ln(1 + e^{-\mathbf{w}^\top \Phi(\mathbf{x}_i)}),$$

where $\Phi(\mathbf{x}_i)$ is the output of the output layer.

a) Let $a_1, a_2, \cdots, a_n \in \mathbb{R}^+$, where $a_i \neq a_j$, $a^* = \max(\{a_1, a_2, \cdots, a_n\})$ and $a^{**} = \max(\{a_1, a_2, \cdots, a_n\} \backslash \{a^*\})$. Show that:

$$\lim_{(a^{**} - a^*) \to -\infty} \ln(e^{a_1} + e^{a_2} + \cdots + e^{a_n}) = \max(\{a_1, a_2, \cdots, a_n\}) = a^*$$

b) Using item (a), show that the hinge loss can be approximated asymptotically by the calculated limit.

c) You probably noticed that the hinge loss can be approximated by the function $h^*(t) = \ln(1 + e^{1-t})$, where $t = y_i(\mathbf{w}^\top \mathbf{x}_i)$. Compare $h^*(t)$ to $l(t) = \ln(1 + e^{-t})$ (logistic loss).

4) Solve programming task 4.