Deep Neural Networks

Exercise Sheet 1

---

by group

DLTMHK

in

SS 2018

May 27, 2018

Exercise 1.1. Log-Likelihood

1.1 (a).

The gradient for the loss function wrt μ is

$$\frac{\partial l(\mu, t)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\log(2\sigma) + \frac{|t-\mu|}{\sigma}\right) = -\frac{1}{\sigma}\operatorname{sign}(t-\mu).$$

1.1 (b).

With $\mu = w^T x$ it follows for the loss function $l(\mu, t)$

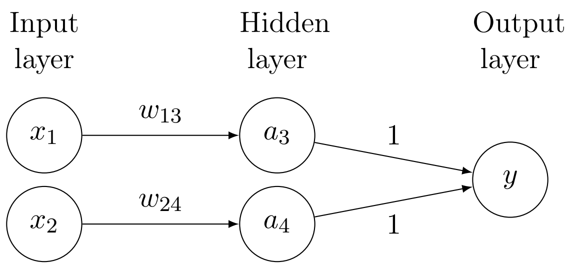
$$\frac{\partial l(\mu, t)}{\partial w} = \frac{\partial l(\mu, t)}{\partial \mu}\frac{\partial \mu}{\partial w} = -\frac{1}{\sigma}\operatorname{sign}(t - w^T x)x.$$

Since the loss function for the entire dataset $(x_i, t_i)_{i=1,\dots,N}$ is the sum of all individual loss functions $\frac{\partial l(\mu_i, t_i)}{\partial w}$

$$\frac{\partial l(\mu, t)}{\partial w} = \sum_{i=1}^{N}\frac{\partial l(\mu_i, t_i)}{\partial w} = -\frac{1}{\sigma}\sum_{i=1}^{N}\operatorname{sign}(t_i - w^T x_i)x_i.$$

Exercise 1.2. Shared Parameters

1.2 (a).



where $a_i = g(z_i), i = 3, 4$ with $g(x) = 0.5x^2$ and $z_3 = x_1 w_{13}, z_4 = x_2 w_{24}$ and $y = a_3 + a_4$.

1.2 (b).

According to the chain rule, the gradient of the loss function wrt to the weights w_{13}, w_{24} are

$$\frac{\partial l}{\partial w_{13}} = \frac{\partial l}{y}\frac{\partial y}{z_3}\frac{\partial z_3}{w_{13}} = (y-t)z_3 x_1, \qquad \frac{\partial l}{\partial w_{24}} = \frac{\partial l}{y}\frac{\partial y}{z_4}\frac{\partial z_4}{w_{24}} = (y-t)z_4 x_2.$$

1.2 (c).

With $w_{13} = \log(1+e^v), w_{23} = -\log(1+e^{-v})$ the gradient of the loss function for v is

$$\frac{\partial l}{\partial v} = \frac{\partial l}{\partial w_{13}}\frac{\partial w_{13}}{\partial v} + \frac{\partial l}{\partial w_{24}}\frac{\partial w_{24}}{\partial v} = (y-t)\left[g'(z_3)x_1\frac{e^v}{1+e^v} + g'(z_4)x_2\frac{e^{-v}}{1+e^{-v}}\right]$$

which follows with

$$\frac{\partial w_{13}}{\partial v} = \frac{e^v}{1+e^v}, \qquad \frac{\partial w_{24}}{\partial v} = \frac{e^{-v}}{1+e^{-v}}.$$

Exercise 1.3. Layered Networks

1.3 (a).

The gradient for the loss function l wrt $z_j^{(l+1)}$ is

$$\frac{\partial l}{\partial z_j^{(l+1)}} = \frac{\partial l}{\partial a_j^{(l+1)}} \frac{\partial a_j^{(l+1)}}{\partial z_j^{(l+1)}} = \frac{\partial l}{\partial a_j^{(l+1)}} a_j^{(l+1)}(1 - a_j^{(l+1)})$$

where

$$\frac{\partial a_j^{(l+1)}}{\partial z_j^{(l+1)}} = \frac{\partial}{\partial z_j^{(l+1)}} \frac{\exp(z_j^{(l+1)})}{1+\exp(z_j^{(l+1)})} = \frac{\exp(z_j^{(l+1)})}{1+\exp(z_j^{(l+1)})} \left(1 - \frac{\exp(z_j^{(l+1)})}{1+\exp(z_j^{(l+1)})}\right) = a_j^{(l+1)}(1 - a_j^{(l+1)}).$$

1.3 (b).

Furthermore it holds for all $i = 1, ..., d$

$$\frac{\partial l}{\partial a_i^{(l)}} = \sum_j \frac{\partial l}{\partial z_j^{(l+1)}} \underbrace{\frac{\partial z_j^{(l+1)}}{\partial a_i^{(l)}}}_{=w_{ij}^{(l)}} = \sum_j \underbrace{\frac{\partial l}{\partial a_j^{(l+1)}} a_j^{(l+1)}(1 - a_j^{(l+1)})}_{=:\tilde{a}_j^{(l+1)}} w_{ij}^{(l)} = \sum_j \tilde{a}_j^{(l+1)} w_{ij}^{(l)}$$

which is by definition of the inner product equal to

$$\frac{\partial l}{\partial a_i^{(l)}} = W\tilde{a}^{(l+1)},$$

where $W = [w_{ij}]_{\substack{i=1,\dots,d \\ j=1,\dots,h}}$ and

$$\tilde{a}^{(l+1)} = \begin{pmatrix} \tilde{a}_1^{(l+1)} \\ \vdots \\ \tilde{a}_h^{(l+1)} \end{pmatrix} = \begin{pmatrix} \frac{\partial l}{\partial a_1^{(l+1)}} a_1^{(l+1)}(1 - a_1^{(l+1)}) \\ \vdots \\ \frac{\partial l}{\partial a_h^{(l+1)}} a_h^{(l+1)}(1 - a_h^{(l+1)}) \end{pmatrix} = \frac{\partial l}{\partial a^{(l+1)}} \odot a^{(l+1)} \odot (1 - a^{(l+1)})$$

with $a^{(l+1)} = \left(a_1^{(l+1)}, ..., a_h^{(l+1)}\right)^T$.