# Machine Learning HW#3

Evgeny Marshakov

## Problem 1

We have the following network

$$y_k(x,w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)}\sigma\left(\sum_{i=1}^{D} w_{ji}^{(1)}x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{1}$$

We want to find an equivalent network with $\tanh(x)$ as the hidden unit activation function. It can be easily seen that

$$2\sigma(a) - 1 = 2\frac{1}{1+e^a} - 1 = \frac{1-e^a}{1+e^a} = -\frac{e^{\frac{a}{2}} - e^{-\frac{a}{2}}}{e^{\frac{a}{2}} + e^{-\frac{a}{2}}} = -\tanh\left(\frac{a}{2}\right) \tag{2}$$

or

$$\sigma(a) = \frac{1-\tanh(\frac{a}{2})}{2} \tag{3}$$

Substituting this formula to the first layer of our network we obtain

$$y_k(x,w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)}\left[\frac{1}{2} - \frac{1}{2}\tanh\left(\sum_{i=1}^{D}\frac{w_{ji}^{(1)}}{2}x_i + \frac{w_{j0}^{(1)}}{2}\right)\right] + w_{k0}^{(2)}\right) = \tag{4}$$

$$= \sigma\left(\sum_{j=1}^{M}\tilde{w}_{kj}^{(2)}\tanh\left(\sum_{i=1}^{D}\tilde{w}_{ji}^{(1)}x_i + \tilde{w}_{j0}^{(1)}\right) + \tilde{w}_{k0}^{(2)}\right) \tag{5}$$

where

$$\tilde{w}_{ji}^{(1)} = \frac{1}{2} w_{ji}^{(1)} \tag{6}$$

$$\tilde{w}_{j0}^{(1)} = \frac{1}{2} w_{j0}^{(1)} \tag{7}$$

$$\tilde{w}_{kj}^{(2)} = -\frac{1}{2} w_{kj}^{(2)} \tag{8}$$

$$\tilde{w}_{k0}^{(2)} = w_{k0}^{(2)} + \frac{1}{2} \sum_{i=1}^{M} w_{kj}^{(2)} \tag{9}$$

Here we change only the hidden activation function. If we also want to change an output activation function then we obtain

$$y_k(x, w) = \frac{1}{2} - \frac{1}{2} \tanh \left( \sum_{j=1}^{M} \frac{\tilde{w}_{kj}^{(2)}}{2} \tanh \left( \sum_{i=1}^{D} \tilde{w}_{ji}^{(1)} x_i + \tilde{w}_{j0}^{(1)} \right) + \frac{\tilde{w}_{k0}^{(2)}}{2} \right) = \tag{10}$$

$$= \frac{1}{2} - \frac{1}{2} \tanh \left( \sum_{j=1}^{M} \hat{w}_{kj}^{(2)} \tanh \left( \sum_{i=1}^{D} \hat{w}_{ji}^{(1)} x_i + \hat{w}_{j0}^{(1)} \right) + \hat{w}_{k0}^{(2)} \right) \tag{11}$$

where

$$\hat{w}_{ji}^{(1)} = \frac{1}{2} w_{ji}^{(1)} \tag{12}$$

$$\hat{w}_{j0}^{(1)} = \frac{1}{2} w_{j0}^{(1)} \tag{13}$$

$$\hat{w}_{kj}^{(2)} = \frac{1}{2} \tilde{w}_{kj}^{(2)} = -\frac{1}{4} w_{kj}^{(2)} \tag{14}$$

$$\hat{w}_{k0}^{(2)} = \frac{1}{2} \tilde{w}_{k0}^{(2)} = \frac{1}{2} w_{k0}^{(2)} + \frac{1}{4} \sum_{i=1}^{M} w_{kj}^{(2)} \tag{15}$$

# Problem 2

We have i.i.d. data, so the likelihood function is the following

$$L(t_1, \ldots, t_N | w) = p(t_1, \ldots, t_N | w) = \prod_{i=1}^{N} p(t_i | w)$$

Since we have the classification problem, we consider Bernoulli distribution

$$p(t_i | y) = y^{t_i} (1 - y)^{1 - t_i}$$

Also we know that there is a probability $\epsilon$ that the class label on a training data point has been incorrectly set, so we have the following

$$p(t_i|w) = p(t_i|y_i) = (y_i(1-\epsilon) + (1-y_i)\epsilon)^{t_i} ((1-y_i)(1-\epsilon) + y_i\epsilon)^{1-t_i}$$

So calculating the likelihood function we obtain

$$L(w) = p(\mathcal{D}|w) = \prod_{i=1}^{N} p(t_i|y_i) = \prod_{i=1}^{N} (y_i(1-\epsilon) + (1-y_i)\epsilon)^{t_i} ((1-y_i)(1-\epsilon) + y_i\epsilon)^{1-t_i} \tag{16}$$

Hence the negative log-likelihood function is the following

$$E_\epsilon(w) := -\log L(w) = -\sum_{i=1}^{N} [t_i \ln (y_i(1-\epsilon) + (1-y_i)\epsilon) + (1-t_i) \ln ((1-y_i)(1-\epsilon) + y_i\epsilon)] \tag{17}$$

Finally, we substitute $\epsilon = 0$ and find that

$$E_0(w) = -\sum_{i=1}^{N} [t_i \ln y_i + (1-t_i) \ln(1-y_i)] = E(w)$$

# Problem 3

We want to find the Hessian matrix of the following function

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \|y(x_n, w) - t_n\|^2 = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^T (y(x_n, w) - t_n)$$

Let us calculate the first derivative with respect to $w_i$

$$\frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_{n=1}^{N} \left( \frac{\partial y(x_n, w)}{\partial w_i} \right)^T (y(x_n, w) - t_n) + (y(x_n, w) - t_n)^T \left( \frac{\partial y(x_n, w)}{\partial w_i} \right) = \tag{18}$$

$$= \sum_{n=1}^{N} \left( \frac{\partial y(x_n, w)}{\partial w_i} \right)^T (y(x_n, w) - t_n) \tag{19}$$

We assume that $\nabla y$Let us calculating the second derivative of $E$ with respect to $w_i, w_j$:

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \sum_{n=1}^{N} \left( \frac{\partial^2 y(x_n, w)}{\partial w_i \partial w_j} \right)^T (y(x_n, w) - t_n) + \left( \frac{\partial y(x_n, w)}{\partial w_i} \right)^T \left( \frac{\partial y(x_n, w)}{\partial w_j} \right) \tag{20}$$

If the network has been trained on the data set, and its outputs $y_n(x_n, w)$ happen to be very close to the target values $t_n$, then the second term in (20) will be small and can be neglected. Under this assumption we obtain that

$$\mathbf{H} = \sum_{n=1}^{N} \nabla y(x_n, w) \nabla y(x_n, w)^T$$

where

$$\{\nabla y(x_n, w)\}_{ij} = \frac{\partial y(x_n, w)_j}{\partial w_i}$$

# Problem 4

We have the following neural network

$$y_k(x, w) = a_k \left( \sum_{j=1}^{M} w_{kj}^{(2)} \tanh \left( \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} + \sum_{i=1}^{D} v_{ki} x_i \right) + b_k \tag{21}$$

here $v_{ki}$ are additional parameters corresponding to skip-layer connections. We consider the sum-of-squares errors loss function

$$E(w) = \frac{1}{2} \sum_{n=1}^{D} \|y(x_n, w) - t_n\|^2$$

Let us calculate the derivative of $E(w)$ w.r.t $v_{ki}$. Using the expression (19) we obtain

$$\frac{\partial E}{\partial v_{ki}} = \sum_{n=1}^{D} \left( \frac{\partial y(x_n, w)}{\partial v_{ki}} \right)^T (y(x_n, w) - t_n) \tag{22}$$

It is obvious that

$$\frac{\partial y(x_n, w)}{\partial v_{ki}} = (0, \ldots, 0, x_i, 0, \ldots, 0) \tag{23}$$

Where $x_i$ is in the position $k$. So we obtain

$$\frac{\partial E}{\partial v_{ki}} = \sum_{n=1}^{D} x_i \cdot [y_k(x_n, w) - t_n] \tag{24}$$

# Problem 5

We want to estimate the following function

$$\tilde{E} = \frac{1}{2} \iiint (y(x+\xi) - t)^2 p(t|x)p(x)p(\xi)d\xi dtdx \tag{25}$$

It is obvious that

$$\frac{\partial y(x+\xi)}{\partial \xi_i} = \frac{\partial y(x)}{\partial x_i} = \{\nabla y\}_i$$

$$\frac{\partial^2 y(x+\xi)}{\partial \xi_i \partial \xi_j} = \frac{\partial \{\nabla y\}_i}{\partial \xi_j} = \mathbf{H}_{ij}$$

Using Taylor series of $(y(x+\xi) - t)^2$ we have the following approximation for $\tilde{E}$

$$\tilde{E} = \frac{1}{2} \iiint (y(x) - t)^2 p(t|x)p(x)p(\xi)d\xi dtdx + \iiint (y(x) - t)(\nabla y)^T \xi p(t|x)p(x)p(\xi)d\xi dtdx + \tag{26}$$

$$+\frac{1}{2} \iiint \xi^T \left[(y(x) - t)\mathbf{H} + (\nabla y)(\nabla y)^T\right] \xi p(t|x)p(x)p(\xi)d\xi dtdx \tag{27}$$

The second term we can split into multiplication of two integrals

$$\iiint (y(x) - t)(\nabla y)^T \xi p(t|x)p(x)p(\xi)d\xi dtdx = \int \xi p(\xi)d\xi \iint (y(x) - t)(\nabla y)^T p(t|x)p(x)dtdx$$

One of them is equal to zero due to the fact that

$$\int \xi p(\xi)d\xi = \mathbb{E}[\xi] = 0$$

So the second term vanishes. We use the same argument as in the exercise 4. We think that our outputs are very close to the target values, so the term with $\mathbf{H}$ can be neglected. So we can approximate the last term of (27) as follows

$$\frac{1}{2} \iiint \xi^T \left[(y(x) - t)\mathbf{H} + (\nabla y)(\nabla y)^T\right] \xi p(t|x)p(x)p(\xi)d\xi dtdx = \tag{28}$$

$$= \frac{1}{2} \iint \xi^T \left[(\nabla y)(\nabla y)^T\right] \xi p(x)p(\xi)d\xi dtdx \tag{29}$$

It can be easily seen that

$$\xi^T A \xi = \sum_{ij} A_{ij}\xi_i\xi_j = \mathrm{tr}(\xi\xi^T A)$$

We know that $\mathbb{E}[\xi\xi^T] = I$ so it can be easily seen that

$$\frac{1}{2}\iint \{\nabla y\}_i \{\nabla y\}_j \xi_i \xi_j p(\xi) p(x) d\xi dx = \frac{1}{2}\int \delta_{ij} \{\nabla y\}_i \{\nabla y\}_j p(x) dx$$

so we can simplify the expression (29) as follows

$$\frac{1}{2}\iint \xi^T \left[(\nabla y)(\nabla y)^T\right] p(x)p(\xi) d\xi dx = \frac{1}{2}\iint \operatorname{tr}\left(\xi\xi^T (\nabla y)(\nabla y)^T\right) p(x)p(\xi) d\xi dx = \tag{30}$$

$$= \frac{1}{2}\int \operatorname{tr}\left((\nabla y)(\nabla y)^T\right) p(x) dx = \frac{1}{2}\int (\nabla y)^T (\nabla y) p(x) dx = \frac{1}{2}\int \|\nabla y\|^2 p(x) dx =: \Omega \tag{31}$$

So we see that

$$\tilde{E} \simeq E + \Omega$$