Assignment 1: Solution

Emanuele Sansone*

December 7, 2016

The following is the objective function we want to minimize:

$$J(\boldsymbol{\theta})=\sum_{j=1}^{K}\left\{-\frac{\pi_j}{|D_p^j|}\sum_{\boldsymbol{x}_i\in D_p^j} f_j(\boldsymbol{x}_i)+\frac{1}{|D_n^j|}\sum_{\boldsymbol{x}_i\in D_n^j}\max\left\{f_j(\boldsymbol{x}_i),\max\left\{0,\frac{1}{2}+\frac{f_j(\boldsymbol{x}_i)}{2}\right\}\right\}\right\}+\lambda\|\boldsymbol{\theta}\|_2^2 \tag{1}$$

A local minimum can be obtained by using any gradient-based optimization algorithm. Therefore, we have to compute the gradient of the objective function with respect to all network parameters. We consider a neural network characterized by 3 layers, namely one output and two hidden layers.

Third Layer

We use indices μ,ν to identify any neuron in the second and the third layer, respectively. Consequently,

$$\begin{aligned}
\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_{\mu\nu}^{(3)}}&=\sum_{j=1}^{K}\left\{-\frac{\pi_j}{|D_p^j|}\sum_{\boldsymbol{x}_i\in D_p^j}\frac{\partial f_j(\boldsymbol{x}_i)}{\partial\theta_{\mu\nu}^{(3)}}+\frac{1}{|D_n^j|}\sum_{\boldsymbol{x}_i\in D_n^j}\frac{\partial\max\left\{f_j(\boldsymbol{x}_i),\max\left\{0,\frac{1}{2}+\frac{f_j(\boldsymbol{x}_i)}{2}\right\}\right\}}{\partial\theta_{\mu\nu}^{(3)}}\right\}+2\lambda\theta_{\mu\nu}^{(3)}\\
&=\sum_{j=1}^{K}\left\{-\frac{\pi_j}{|D_p^j|}\sum_{\boldsymbol{x}_i\in D_p^j}\frac{\partial f_j(\boldsymbol{x}_i)}{\partial\theta_{\mu\nu}^{(3)}}+\frac{1}{|D_n^j|}\sum_{\boldsymbol{x}_i\in D_n^j}\alpha_{ji}\frac{\partial f_j(\boldsymbol{x}_i)}{\partial\theta_{\mu\nu}^{(3)}}\right\}+2\lambda\theta_{\mu\nu}^{(3)}\\
&=-\frac{\pi_\nu}{|D_p^\nu|}\sum_{\boldsymbol{x}_i\in D_p^\nu}\hat{y}_{\mu i}^{(2)}+\frac{1}{|D_n^\nu|}\sum_{\boldsymbol{x}_i\in D_n^\nu}\alpha_{\nu i}\hat{y}_{\mu i}^{(2)}+2\lambda\theta_{\mu\nu}^{(3)}\\
&=\sum_{\boldsymbol{x}_i\in D_p^\nu}-\frac{\pi_\nu}{|D_p^\nu|}\hat{y}_{\mu i}^{(2)}+\sum_{\boldsymbol{x}_i\in D_n^\nu}\frac{\alpha_{\nu i}}{|D_n^\nu|}\hat{y}_{\mu i}^{(2)}+2\lambda\theta_{\mu\nu}^{(3)}\\
&=\sum_{\boldsymbol{x}_i\in D}\delta_{\nu i}^{(3)}\hat{y}_{\mu i}^{(2)}+2\lambda\theta_{\mu\nu}^{(3)}
\end{aligned} \tag{2}$$

where

$$\alpha_{ji}=\begin{cases}1 & f_j(\boldsymbol{x}_i)>1\\ \frac{1}{2} & -1\leq f_j(\boldsymbol{x}_i)\leq 1\\ 0 & f_j(\boldsymbol{x}_i)<-1\end{cases}$$

$$\frac{\partial f_j(\boldsymbol{x}_i)}{\partial\theta_{\mu\nu}^{(3)}}=\begin{cases}0 & j\neq\nu\\ \frac{\partial f_j(\boldsymbol{x}_i)}{\partial z_{ji}^{(3)}}\frac{\partial z_{ji}^{(3)}}{\partial\theta_{\mu\nu}^{(3)}}=\hat{y}_{\mu i}^{(2)} & j=\nu\end{cases}$$

$$\delta_{\nu i}^{(3)}=\begin{cases}-\frac{\pi_\nu}{|D_p^\nu|} & \boldsymbol{x}_i\in D_p^\nu\\ \frac{\alpha_{\nu i}}{|D_n^\nu|} & \boldsymbol{x}_i\in D_n^\nu\\ 0 & \boldsymbol{x}_i\notin D_p^\nu\bigcup D_n^\nu\end{cases}$$

*https://emsansone.github.io/

Second Layer

We use indices μ, ν to identify any neuron in the first and the second layer, respectively (whereas j identifies any neuron in the third layer). Consequently,

$$\begin{aligned}
\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_{\mu\nu}^{(2)}} &= \sum_{j=1}^{K} \left\{ -\frac{\pi_j}{|D_p^j|} \sum_{\boldsymbol{x}_i \in D_p^j} \frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(2)}} + \frac{1}{|D_n^j|} \sum_{\boldsymbol{x}_i \in D_n^j} \frac{\partial \max\left\{ f_j(\boldsymbol{x}_i), \max\left\{0, \frac{1}{2} + \frac{f_j(\boldsymbol{x}_i)}{2}\right\}\right\}}{\partial \theta_{\mu\nu}^{(2)}} \right\} + 2\lambda \theta_{\mu\nu}^{(2)} \\
&= \sum_{j=1}^{K} \left\{ -\frac{\pi_j}{|D_p^j|} \sum_{\boldsymbol{x}_i \in D_p^j} \frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(2)}} + \frac{1}{|D_n^j|} \sum_{\boldsymbol{x}_i \in D_n^j} \alpha_{ji} \frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(2)}} \right\} + 2\lambda \theta_{\mu\nu}^{(2)} \\
&= \sum_{j=1}^{K} \left\{ -\frac{\pi_j}{|D_p^j|} \sum_{\boldsymbol{x}_i \in D_p^j} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \hat{y}_{\mu i}^{(1)} + \frac{1}{|D_n^j|} \sum_{\boldsymbol{x}_i \in D_n^j} \alpha_{ji} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \hat{y}_{\mu i}^{(1)} \right\} + 2\lambda \theta_{\mu\nu}^{(2)} \\
&= \sum_{j=1}^{K} \left\{ \sum_{\boldsymbol{x}_i \in D_p^j} -\frac{\pi_j}{|D_p^j|} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \hat{y}_{\mu i}^{(1)} + \sum_{\boldsymbol{x}_i \in D_n^j} \frac{\alpha_{ji}}{|D_n^j|} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \hat{y}_{\mu i}^{(1)} \right\} + 2\lambda \theta_{\mu\nu}^{(2)} \\
&= \sum_{j=1}^{K} \left\{ \sum_{\boldsymbol{x}_i \in D} \delta_{ji}^{(3)} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \hat{y}_{\mu i}^{(1)} \right\} + 2\lambda \theta_{\mu\nu}^{(2)} \\
&= \sum_{\boldsymbol{x}_i \in D} \left\{ \sum_{j=1}^{K} \delta_{ji}^{(3)} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \right\} \hat{y}_{\mu i}^{(1)} + 2\lambda \theta_{\mu\nu}^{(2)} \\
&= \sum_{\boldsymbol{x}_i \in D} \delta_{\nu i}^{(2)} \hat{y}_{\mu i}^{(1)} + 2\lambda \theta_{\mu\nu}^{(2)} \qquad (3)
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(2)}} &= \frac{\partial f_j(\boldsymbol{x}_i)}{\partial z_{ji}^{(3)}} \frac{\partial z_{ji}^{(3)}}{\partial \hat{y}_{\nu i}^{(2)}} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \frac{\partial z_{\nu i}^{(2)}}{\partial \theta_{\mu\nu}^{(2)}} = \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}} \hat{y}_{\mu i}^{(1)} \\
\delta_{\nu i}^{(2)} &= \sum_{j=1}^{K} \delta_{ji}^{(3)} \theta_{\nu j}^{(3)} \frac{\partial \hat{y}_{\nu i}^{(2)}}{\partial z_{\nu i}^{(2)}}
\end{aligned}$$

First Layer

We use indices μ, ν, ρ to identify any neuron in the input, in the first and in the second layer, respectively (whereas j identifies any neuron in the output layer). Consequently,

$$
\begin{aligned}
\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_{\mu\nu}^{(1)}} &= \sum_{j=1}^{K} \left\{ -\frac{\pi_j}{|D_p^j|} \sum_{\boldsymbol{x}_i \in D_p^j} \frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(1)}} + \frac{1}{|D_n^j|} \sum_{\boldsymbol{x}_i \in D_n^j} \frac{\partial \max\left\{ f_j(\boldsymbol{x}_i), \max\left\{0, \frac{1}{2} + \frac{f_j(\boldsymbol{x}_i)}{2}\right\}\right\}}{\partial \theta_{\mu\nu}^{(1)}} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{j=1}^{K} \left\{ -\frac{\pi_j}{|D_p^j|} \sum_{\boldsymbol{x}_i \in D_p^j} \frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(1)}} + \frac{1}{|D_n^j|} \sum_{\boldsymbol{x}_i \in D_n^j} \alpha_{ji} \frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(1)}} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{j=1}^{K} \left\{ -\frac{\pi_j}{|D_p^j|} \sum_{\boldsymbol{x}_i \in D_p^j} \sum_{\rho=1}^{n_2} \theta_{\rho j}^{(3)} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{\mu i} + \frac{1}{|D_n^j|} \sum_{\boldsymbol{x}_i \in D_n^j} \alpha_{ji} \mathrm{IDEM} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{j=1}^{K} \left\{ \sum_{\boldsymbol{x}_i \in D_p^j} -\frac{\pi_j}{|D_p^j|} \sum_{\rho=1}^{n_2} \theta_{\rho j}^{(3)} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{\mu i} + \sum_{\boldsymbol{x}_i \in D_n^j} \frac{\alpha_{ji}}{|D_n^j|} \mathrm{IDEM} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{j=1}^{K} \left\{ \sum_{\boldsymbol{x}_i \in D} \delta_{ji}^{(3)} \sum_{\rho=1}^{n_2} \theta_{\rho j}^{(3)} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{\mu i} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{j=1}^{K} \left\{ \sum_{\boldsymbol{x}_i \in D} \sum_{\rho=1}^{n_2} \delta_{ji}^{(3)} \theta_{\rho j}^{(3)} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{\mu i} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{\boldsymbol{x}_i \in D} \left\{ \sum_{\rho=1}^{n_2} \left\{ \sum_{j=1}^{K} \delta_{ji}^{(3)} \theta_{\rho j}^{(3)} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \right\} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{\mu i} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{\boldsymbol{x}_i \in D} \left\{ \sum_{\rho=1}^{n_2} \delta_{\rho i}^{(2)} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{\mu i} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{\boldsymbol{x}_i \in D} \left\{ \left\{ \sum_{\rho=1}^{n_2} \delta_{\rho i}^{(2)} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} \right\} x_{\mu i} \right\} + 2\lambda \theta_{\mu\nu}^{(1)} \\
&= \sum_{\boldsymbol{x}_i \in D} \delta_{\nu i}^{(1)} x_{i\mu} + 2\lambda \theta_{\mu\nu}^{(1)}
\end{aligned}
$$

where

$$
\frac{\partial f_j(\boldsymbol{x}_i)}{\partial \theta_{\mu\nu}^{(1)}} = \frac{\partial f_j(\boldsymbol{x}_i)}{\partial z_{ji}^{(3)}} \sum_{\rho=1}^{n_2} \frac{\partial z_{ji}^{(3)}}{\partial \hat{y}_{\rho i}^{(2)}} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \frac{\partial z_{\rho i}^{(2)}}{\partial \hat{y}_{\nu i}^{(1)}} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} \frac{\partial z_{\nu i}^{(1)}}{\partial \theta_{\mu\nu}^{(1)}} = \sum_{\rho=1}^{n_2} \theta_{\rho j}^{(3)} \frac{\partial \hat{y}_{\rho i}^{(2)}}{\partial z_{\rho i}^{(2)}} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}} x_{i\mu}
$$

$$
\delta_{\nu i}^{(1)} = \sum_{\rho=1}^{n_2} \delta_{\rho i}^{(2)} \theta_{\nu\rho}^{(2)} \frac{\partial \hat{y}_{\nu i}^{(1)}}{\partial z_{\nu i}^{(1)}}
$$

Extension to the General Case

Based on the previous computation, we can obtain the following general formulas:

$$
\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_{\mu\nu}^{(\ell)}} = \sum_{\boldsymbol{x}_i \in D} \delta_{\nu i}^{(\ell)} \hat{y}_{\mu i}^{(\ell)} + 2\lambda \theta_{\mu\nu}^{(\ell)}
$$

where, if ℓ identifies the output layer, then

$$
\delta_{\nu i}^{(\ell)} = \begin{cases} -\frac{\pi_\nu}{|D_p^\nu|} & \boldsymbol{x}_i \in D_p^\nu \\ \frac{\alpha_{\nu i}}{|D_n^\nu|} & \boldsymbol{x}_i \in D_n^\nu \\ 0 & \boldsymbol{x}_i \notin D_p^\nu \bigcup D_n^\nu \end{cases}, \quad \alpha_{\nu i} = \begin{cases} 1 & f_\nu(\boldsymbol{x}_i) > 1 \\ \frac{1}{2} & -1 \le f_\nu(\boldsymbol{x}_i) \le 1 \\ 0 & f_\nu(\boldsymbol{x}_i) < -1 \end{cases}
$$

while, if ℓ identifies any other layer (viz. any hidden layer), then

$$
\delta_{\nu i}^{(\ell)} = \sum_{\rho=1}^{n_{\ell+1}} \delta_{\rho i}^{(\ell+1)} \theta_{\nu\rho}^{(\ell+1)} \frac{\partial \hat{y}_{\nu i}^{(\ell)}}{\partial z_{\nu i}^{(\ell)}}
$$