1 Introduction

2 MLP for binary classification

2.1 Inplementing the gradient descent

2.1.1 Forward pass

First Layer

$$\begin{cases} a_{L,q}^{(1)} = (\boldsymbol{w}_{L,q}^{(1)})^T \cdot \boldsymbol{x}_L + b_{L,q}^{(1)},\ q = 1, ..., h_1 \\ a_{R,q}^{(1)} = (\boldsymbol{w}_{R,q}^{(1)})^T \cdot \boldsymbol{x}_R + b_{R,q}^{(1)},\ q = 1, ..., h_1 \end{cases}$$

Or in a more compact way:

$$\begin{cases} \boldsymbol{a}_L^{(1)} = \boldsymbol{W}_L^{(1)} \cdot \boldsymbol{x}_L + \boldsymbol{b}_L^{(1)} \\ \boldsymbol{a}_R^{(1)} = \boldsymbol{W}_R^{(1)} \cdot \boldsymbol{x}_R + \boldsymbol{b}_R^{(1)} \end{cases}$$

$$\begin{cases} \boldsymbol{z}_L^{(1)} = \boldsymbol{g}_1(\boldsymbol{a}_L^{(1)}) \\ \boldsymbol{z}_R^{(1)} = \boldsymbol{g}_1(\boldsymbol{a}_R^{(1)}) \end{cases}$$

Second Layer

$$a_{L,q}^{(2)} = (\boldsymbol{w}_{L,q}^{(2)})^T \cdot \boldsymbol{z}_L^{(1)} + b_{L,q}^{(2)},\ q = 1, ..., h_2$$

$$a_{LR,q}^{(2)} = (\boldsymbol{w}_{LR,q}^{(2)})^T \cdot \begin{bmatrix} \boldsymbol{z}_L^{(1)} \\ \boldsymbol{z}_R^{(1)} \end{bmatrix} + b_{L,q}^{(2)},\ q = 1, ..., h_2$$

$$a_{R,q}^{(2)} = (\boldsymbol{w}_{R,q}^{(2)})^T \cdot \boldsymbol{z}_R^{(1)} + b_{R,q}^{(2)},\ q = 1, ..., h_2$$

More compactly,

$$\boldsymbol{a}_L^{(2)} = \boldsymbol{W}_L^{(2)} \cdot \boldsymbol{z}_L^{(1)} + \boldsymbol{b}_L^{(2)}$$

$$\boldsymbol{a}_{LR}^{(2)} = \boldsymbol{W}_{LR}^{(2)} \cdot \begin{bmatrix} \boldsymbol{z}_L^{(1)} \\ \boldsymbol{z}_R^{(1)} \end{bmatrix} + \boldsymbol{b}_{LR}^{(2)}$$

$$\boldsymbol{a}_R^{(2)} = \boldsymbol{W}_R^{(2)} \cdot \boldsymbol{z}_R^{(1)} + \boldsymbol{b}_R^{(2)}$$

And,

$$\boldsymbol{z}^{(2)} = \boldsymbol{g}_2(\boldsymbol{a}_L^{(2)}, \boldsymbol{a}_{LR}^{(2)}, \boldsymbol{a}_R^{(2)})$$

Third Layer

$$a_q^{(3)} = (\boldsymbol{w}_q^{(3)})^T \cdot \boldsymbol{z}^{(2)} + b_q^{(3)},\ q = 1, ..., K$$

$$\boldsymbol{a}^{(3)} = \mathbf{W}^{(3)} \cdot \boldsymbol{z}^{(2)} + \boldsymbol{b}^{(3)}$$

2.2 backward pass

TODO

2.3 Results

(EXTREMELY good results on validation set, training set too easy)

2.4 Extrapolation

(creation of another binary test)

3 Multi-way classification

3.1 Linear multi-way classification

(squared error, tichinov regularizer, logistic error)

3.2 How to deal multis way with MLP?

3.3 Computation of gradient

3.3.1 Forward pass

The first and second layer or the forward bass are exactly the same as in the binary case. Only change the third laye r :

$$a_q^{(3)} = (\boldsymbol{w}_q^{(3)})^T \cdot \boldsymbol{z}^{(2)} + b_q^{(3)},\ q = 1, ..., K$$

$$\boldsymbol{a}^{(3)} = \mathbf{W}^{(3)} \cdot \boldsymbol{z}^{(2)} + \boldsymbol{b}^{(3)}$$

3.3.2 Backward pass

Third Layer The goal is the minimize $E_2(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{N} \left\| \boldsymbol{a}^{(3)}(\boldsymbol{x}_i) - \tilde{\boldsymbol{t}_i} \right\|^2$. We compute the gradient for one i that is we minimize $E_{2,i}$ that we will refer to as E_i

$$r_q^{(3)} = \frac{\partial E_i}{\partial a_q^{(3)}} = 2(a_q^{(3)} - \tilde{\boldsymbol{t}_q}),\ q = 1, ..., K$$

$$\boldsymbol{r}^{(3)} = [r_q^{(3)}] = 2(\boldsymbol{a}^{(3)} - \tilde{\boldsymbol{t}})$$

$$\begin{cases} \nabla_{\boldsymbol{W}^{(3)}} E_i = r^{(3)} \cdot \nabla_{\boldsymbol{W}^{(3)}} a^{(3)} = \boldsymbol{r}^{(3)} \cdot \left(\boldsymbol{z}^{(2)}\right)^T \\ \nabla_{\boldsymbol{b}^{(3)}} E_i = r^{(3)} \cdot \nabla_{\boldsymbol{b}^{(3)}} a^{(3)} = \boldsymbol{r}^{(3)} \end{cases}$$

Second Layer First, let's compute the derivatives (copy from mlp_implementation.docx)

$$\begin{aligned} r_{L,q}^{(2)} &= \frac{\partial E_i}{\partial a_{L,q}^{(2)}} = \sum_j^K \frac{\partial E_i}{\partial a_j^{(3)}} \cdot \frac{\partial a_j^{(3)}}{\partial a_{L,q}^{(2)}} \\ &= \sum_j^K r_j^{(3)} \frac{\partial [(\boldsymbol{w}_j^{(3)})^T \cdot \boldsymbol{g}_2(\boldsymbol{a}_L^{(2)}, \boldsymbol{a}_{LR}^{(2)}, \boldsymbol{a}_R^{(2)}) + b_j^{(3)}]}{\partial a_{L,q}^{(2)}} \\ &= \sum_j^K r_j^{(3)} \frac{\partial [\sum_{\mathbf{k}} \boldsymbol{w}_{j,k}^{(3)} g_2(\boldsymbol{a}_{L,k}^{(2)}, \boldsymbol{a}_{LR,k}^{(2)}, \boldsymbol{a}_{R,k}^{(2)})]}{\partial a_{L,q}^{(2)}} \\ &= \sum_j^K r_j^{(3)} \frac{\partial [\mathbf{w}_{j,q}^{(3)} g_2(\boldsymbol{a}_{L,q}^{(2)}, \boldsymbol{a}_{LR,q}^{(2)}, \boldsymbol{a}_{R,q}^{(2)})]}{\partial a_{L,q}^{(2)}} \\ &= \sum_j^K r_j^{(3)} w_{j,q}^{(3)} g_2'(\boldsymbol{a}_{L,q}^{(2)}, \boldsymbol{a}_{LR,q}^{(2)}, \boldsymbol{a}_{R,q}^{(2)}) \end{aligned}$$

Vectorizing gives:

$$\begin{cases} \boldsymbol{r}_L^{(2)} = \left(diag\left(g_2'(\boldsymbol{a}_L^{(2)}, \boldsymbol{a}_{LR}^{(2)}, \boldsymbol{a}_R^{(2)})\right)\right) \cdot (\boldsymbol{W}^{(3)})^T \cdot \boldsymbol{r}^{(3)} \\ \boldsymbol{r}_{LR}^{(2)} = \left(diag\left(g_2'(\boldsymbol{a}_L^{(2)}, \boldsymbol{a}_{LR}^{(2)}, \boldsymbol{a}_R^{(2)})\right)\right) \cdot (\boldsymbol{W}^{(3)})^T \cdot \boldsymbol{r}^{(3)} \\ \boldsymbol{r}_R^{(2)} = \left(diag\left(g_2'(\boldsymbol{a}_L^{(2)}, \boldsymbol{a}_{LR}^{(2)}, \boldsymbol{a}_R^{(2)})\right)\right) \cdot (\boldsymbol{W}^{(3)})^T \cdot \boldsymbol{r}^{(3)} \end{cases}$$

So now the gradients are:

$$\begin{cases} \nabla_{\boldsymbol{W}_L^{(2)}} E_i = \boldsymbol{r}_L^{(2)} \left(\nabla_{\boldsymbol{W}_L^{(2)}} a_L^{(2)}\right)^T = \boldsymbol{r}_L^{(2)} \cdot \left(\boldsymbol{z}_L^{(1)}\right)^T \\ \nabla_{\boldsymbol{W}_{LR}^{(2)}} E_i = \boldsymbol{r}_{LR}^{(2)} \left(\nabla_{\boldsymbol{W}_{LR}^{(2)}} a_{LR}^{(2)}\right)^T = \boldsymbol{r}_{LR}^{(2)} \cdot \begin{bmatrix} \boldsymbol{z}_L^{(1)} \\ \boldsymbol{z}_R^{(1)} \end{bmatrix}^T \\ \nabla_{\boldsymbol{W}_R^{(2)}} E_i = \boldsymbol{r}_R^{(2)} \left(\nabla_{\boldsymbol{W}_R^{(2)}} a_R^{(2)}\right)^T = \boldsymbol{r}_R^{(2)} \cdot \left(\boldsymbol{z}_R^{(1)}\right)^T \end{cases}$$

And for b:

$$\begin{cases} \nabla_{\boldsymbol{b}_L^{(2)}} E_i = \boldsymbol{r}_L^{(2)} \left(\nabla_{\boldsymbol{b}_L^{(2)}} a_L^{(2)}\right)^T = \boldsymbol{r}_L^{(2)} \\ \nabla_{\boldsymbol{b}_{LR}^{(2)}} E_i = \boldsymbol{r}_{LR}^{(2)} \left(\nabla_{\boldsymbol{b}_{LR}^{(2)}} a_{LR}^{(2)}\right)^T = \boldsymbol{r}_{LR}^{(2)} \\ \nabla_{\boldsymbol{b}_R^{(2)}} E_i = \boldsymbol{r}_R^{(2)} \left(\nabla_{\boldsymbol{b}_R^{(2)}} a_R^{(2)}\right)^T = \boldsymbol{r}_R^{(2)} \end{cases}$$

First Layer

$$\begin{aligned} r_{L,q}^{(1)} &= \frac{\partial E_i}{\partial a_{L,q}^{(1)}} = \sum_j^{h_2} \frac{\partial E_i}{\partial a_{L,j}^{(2)}} \cdot \frac{\partial a_{L,j}^{(2)}}{\partial a_{L,q}^{(1)}} + \sum_j^{h_2} \frac{\partial E_i}{\partial a_{LR,j}^{(2)}} \cdot \frac{\partial a_{LR,j}^{(2)}}{\partial a_{L,q}^{(1)}} \\ &= \sum_j^{h_2} r_{L,j}^{(2)} \frac{\partial a_{L,j}^{(2)}}{\partial a_{L,q}^{(1)}} + \sum_j^{h_2} r_{LR,j}^{(2)} \frac{\partial a_{LR,j}^{(2)}}{\partial a_{L,q}^{(1)}} \end{aligned}$$

$$\begin{aligned} r_{R,q}^{(1)} &= \frac{\partial E_i}{\partial a_{R,q}^{(1)}} = \sum_j^{h_2} \frac{\partial E_i}{\partial a_{R,j}^{(2)}} \cdot \frac{\partial a_{R,j}^{(2)}}{\partial a_{R,q}^{(1)}} + \sum_j^{h_2} \frac{\partial E_i}{\partial a_{LR,j}^{(2)}} \cdot \frac{\partial a_{LR,j}^{(2)}}{\partial a_{R,q}^{(1)}} \\ &= \sum_j^{h_2} r_{R,j}^{(2)} \frac{\partial a_{R,j}^{(2)}}{\partial a_{R,q}^{(1)}} + \sum_j^{h_2} r_{LR,j}^{(2)} \frac{\partial a_{LR,j}^{(2)}}{\partial a_{R,q}^{(1)}} \end{aligned}$$

So, for a (p, j) pair, we have

$$\begin{aligned} \frac{\partial a_{L,j}^{(2)}}{\partial a_{L,q}^{(1)}} &= \frac{\partial \left[\boldsymbol{w}_{L,j}^{(2)} \cdot \boldsymbol{g}_1(\boldsymbol{a}_L^{(1)}) + b_{L,j}^{(2)} \right]}{\partial a_{L,q}^{(1)}} \\ &= \frac{\partial \left[\sum_k w_{L,j,k}^{(2)} \cdot g_1(a_{L,k}^{(1)}) \right]}{\partial a_{L,q}^{(1)}} \\ &= \frac{\partial \left[w_{L,j,q}^{(2)} \cdot g_1(a_{L,q}^{(1)}) \right]}{\partial a_{L,q}^{(1)}} \\ &= w_{L,j,q}^{(2)} \cdot g_1'(a_{L,q}^{(1)}) \end{aligned}$$

$$\begin{aligned} \frac{\partial a_{LR,j}^{(2)}}{\partial a_{L,q}^{(1)}} &= \frac{\partial \left[\boldsymbol{w}_{LR,j}^{(2)} \cdot \begin{bmatrix} \boldsymbol{g}_1(\boldsymbol{a}_L^{(1)}) \\ \boldsymbol{g}_1(\boldsymbol{a}_R^{(1)}) \end{bmatrix} + b_{L,j}^{(2)} \right]}{\partial a_{L,q}^{(1)}} \\ &= \frac{\partial \left[\sum_{k=1}^{h1} w_{LR,j}^{(2)}(k) \cdot g_1(a_{L,k}^{(1)}) + \sum_{k=1}^{h1} w_{LR,j}^{(2)}(h_1 + k) \cdot g_1(a_{R,k}^{(1)}) \right]}{\partial a_{L,q}^{(1)}} \\ &= w_{LR,j}^{(2)}(q) \cdot g_1'(a_{L,q}^{(1)}) \end{aligned}$$

$$
\begin{aligned}
r_{L,q}^{(1)} &= \sum_{j}^{h_2} r_{L,j}^{(2)} \frac{\partial a_{L,j}^{(2)}}{\partial a_{L,q}^{(1)}} + \sum_{j}^{h_2} r_{LR,j}^{(2)} \frac{\partial a_{LR,j}^{(2)}}{\partial a_{L,q}^{(1)}} \\
&= \sum_{j}^{h_2} r_{L,j}^{(2)} \cdot w_{L,j,q}^{(2)} \cdot g_1'(a_{L,q}^{(1)}) + \sum_{j}^{h_2} r_{LR,j}^{(2)} \cdot w_{LR,j}^{(2)}(q) \cdot g_1'(a_{L,q}^{(1)}) \\
&= g_1'(a_{L,q}^{(1)}) \cdot \left(\boldsymbol{W}_L^{(2)}\right)_q^T \cdot \boldsymbol{r}_L^{(2)} + g_1'(a_{L,q}^{(1)}) \cdot \left(\boldsymbol{W}_{LR}^{(2)}\right)_q^T \cdot \boldsymbol{r}_{LR}^{(2)}
\end{aligned}
$$

In summary after vectorization we have:

$$
\begin{cases}
\boldsymbol{r}_L^{(1)} &= diag\left(g_1'(\boldsymbol{a}_L^{(1)})\right) \cdot \left(\boldsymbol{W}_L^{(2)}\right)^T \cdot \boldsymbol{r}_L^{(2)} + diag\left(g_1'(\boldsymbol{a}_L^{(1)})\right) \cdot \left(\boldsymbol{W}_{LR}^{(2)}(:, 1:h_1)\right)^T \cdot \boldsymbol{r}_{LR}^{(2)} \\
\boldsymbol{r}_R^{(1)} &= diag\left(g_1'(\boldsymbol{a}_R^{(1)})\right) \cdot \left(\boldsymbol{W}_R^{(2)}\right)^T \cdot \boldsymbol{r}_R^{(2)} + diag\left(g_1'(\boldsymbol{a}_R^{(1)})\right) \cdot \left(\boldsymbol{W}_{LR}^{(2)}(:, h_1+1:2\times h_1)\right)^T \cdot \boldsymbol{r}_{LR}^{(2)}
\end{cases}
$$

Now that we have computed the residual variables, we can easily find the gradients for the first layer. Indeed,

$$
\begin{cases}
\nabla_{\boldsymbol{W}_L^{(1)}} E_i = \boldsymbol{r}_L^{(1)} \cdot \left(\nabla_{\boldsymbol{W}_L^{(1)}} a_L^{(1)}\right)^T = \boldsymbol{r}_L^{(1)} \cdot (\boldsymbol{x}_L)^T \\
\nabla_{\boldsymbol{W}_R^{(1)}} E_i = \boldsymbol{r}_R^{(1)} \cdot \left(\nabla_{\boldsymbol{W}_R^{(1)}} a_R^{(1)}\right)^T = \boldsymbol{r}_R^{(1)} \cdot (\boldsymbol{x}_R)^T
\end{cases}
$$

And,

$$
\begin{cases}
\nabla_{\boldsymbol{b}_L^{(1)}} E_i = \boldsymbol{r}_L^{(1)} \cdot \left(\nabla_{\boldsymbol{b}_L^{(1)}} a_L^{(1)}\right)^T = \boldsymbol{r}_L^{(1)} \\
\nabla_{\boldsymbol{b}_R^{(1)}} E_i = \boldsymbol{r}_R^{(1)} \cdot \left(\nabla_{\boldsymbol{b}_R^{(1)}} a_R^{(1)}\right)^T = \boldsymbol{r}_R^{(1)}
\end{cases}
$$

3.4 Results and optimization

3.5 Discussion

4 Conclusion