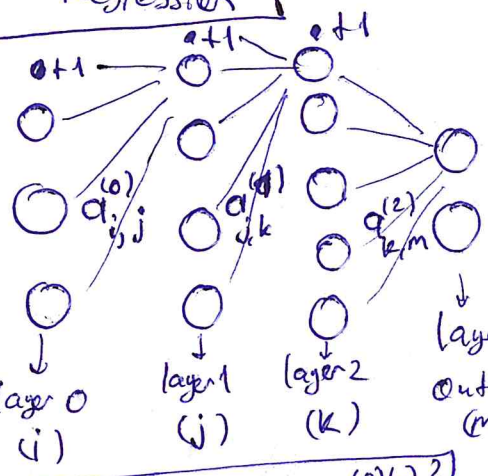
9.11

# MLP Regression

$q_{i,j}^{(l)}$ — $l$: layer number; $i$: left side node number; $j$: right side node number; $q$: weight

$X_i^{(l)t}$ — $l$: layer number; $t$: $t$th sample is input; $i$: node number; $X$: node output




$$X_i^{(0)t} = [1 \;\; x_1^t \;\; x_2^t \;\; x_3^t]$$

$$O_j^{(1)t} = \sum_{i=0} a_{i,j}^{(0)} \cdot X_i^{(0)t}$$

$$X_j^{(1)t} = \sigma(O_j^{(1)t})$$

$$O_k^{(2)t} = \sum_{j=0} a_{j,k}^{(1)} \cdot X_j^{(1)t}$$

$$X_k^{(2)t} = \sigma(O_k^{(2)t})$$

$$O_m^{(3)t} = \sum_{k=0} a_{k,m}^{(2)} X_k^{(2)t}$$

$t$ → $t$th sample's

$m$ → output node number

label

$$\sigma(x) = \frac{1}{1-e^{-x}}$$

$$E(X^{(0)}) = \frac{1}{2} \sum_t \sum_m (r_m^t - O_m^{(3)t})^2$$

$$w_{t+1} = w_t + -\eta \cdot \frac{\partial E}{\partial w}$$

$$\Delta w = -\eta \frac{\partial E}{\partial w}$$

$$\frac{\partial E}{\partial a_{k,m}^{(2)}} = \frac{\partial \left(\frac{1}{2} \sum_t \sum_m (r_m^t - O_m^{(3)t})^2\right)}{\partial a_{k,m}^{(2)}}$$

$$\Delta a_{k,m}^{(2)} = -\eta \cdot \frac{\partial E}{\partial a_{k,m}^{(2)}}$$

$$= \sum_t \sum_m (r_m^t - O_m^{(3)t}) \cdot -1 \cdot \frac{\partial O_m^{(3)t}}{\partial a_{k,m}^{(2)}}$$

$$\frac{\partial O_m^{(3)t}}{\partial a_{k,m}^{(2)}} = \frac{\partial \left(\sum_{k=0} a_{k,m}^{(2)} \cdot X_k^{(2)t}\right)}{\partial a_{k,m}^{(2)}}$$

$$= X_k^{(2)t}$$

$$\Delta a_{k,m}^{(2)} = \sum_t \sum_m (r_m^t - O_m^{(3)t}) \cdot X_k^{(2)t}$$

$$X_0^{(0)} = X_0^{(1)} = X_0^{(2)} = +1$$

$$a^{(0)} = \begin{bmatrix} a_{0,1}^{(0)} & a_{0,2}^{(0)} & a_{0,3}^{(0)} & a_{0,4}^{(0)} \\ a_{1,1}^{(0)} & a_{1,2}^{(0)} & a_{1,3}^{(0)} & a_{1,4}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} & a_{2,3}^{(0)} & a_{2,4}^{(0)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} & a_{3,3}^{(0)} & a_{3,4}^{(0)} \end{bmatrix}$$

$$X^{(l)} = [X_0^{(l)} \; X_1^{(l)} \; X_2^{(l)} \; X_3^{(l)} \; \cdots]$$

$$O^{(l)} = [\cancel{+1} \; O_1^{(l)} \; O_2^{(l)} \; O_3^{(l)} \; \cdots]$$

added after

$$O_s^{(l)t} = \sum_\alpha a_{\alpha,s}^{(l-1)} \cdot X_\alpha^{(l-1)t} \rightarrow \text{expand } \alpha \rightarrow$$

$$= \begin{bmatrix} a_{0,s}^{(l-1)} \\ a_{1,s}^{(l-1)} \\ a_{2,s}^{(l-1)} \\ a_{3,s}^{(l-1)} \end{bmatrix} \cdot [X_0 \; X_1 \; X_2 \; X_3]$$

$$= \begin{bmatrix} a_{0,0}^{(l-1)} \\ a_{1,0}^{(l-1)} \\ a_{2,0}^{(l-1)} \\ a_{3,0} \end{bmatrix} [X_0 \; X_1 \; X_2 \; X_3] \quad \begin{bmatrix} a_{0,1}^{(l-1)} \\ a_{1,1}^{(l-1)} \\ a_{2,1}^{(l-1)} \\ a_{3,1}^{(l-1)} \end{bmatrix} [X_0 \; X_1 \; X_2 \; X_3] \quad \begin{bmatrix} a_{0,s}^{(l-1)} \\ a_{1,s}^{(l-1)} \\ a_{2,s}^{(l-1)} \\ a_{3,s}^{(l-1)} \end{bmatrix} \cdots$$

scalar

$$O^{(l)t} = X^{(l-1)t} \cdot a^{(l-1)}$$

matmul

matrix form

matrix multiplication

$$[X_0 \; X_1 \; X_2 \; X_3] \cdot \begin{bmatrix} a_{0,0}^{(l-1)} & a_{0,1}^{(l-1)} & a_{0,2}^{(l-1)} \\ a_{1,0}^{(l-1)} & a_{1,1}^{(l-1)} & a_{1,2}^{(l-1)} \\ a_{2,0}^{(l-1)} & a_{2,1}^{(l-1)} & a_{2,2}^{(l-1)} \\ a_{3,0} & a_{3,1} & a_{3,2} \end{bmatrix}$$

$$= [+1 \; O_1^{(l)} \; O_2^{(l)} \; O_3^{(l)} \; \cdots]$$

2.) $\Delta a^{(2)}_{k,m} = \sum_t \sum_m (r_m{}^t - O_m^{(3)t}) \cdot X_k^{(2)t}$ $\quad r_m{}^t - O_m^{(3)t} = e_m^t$

$= \sum_t \sum_m e_m{}^t \cdot X_k^{(2)t}$ → open m index → depending on m,k shown $\sum_t [e_0{}^t \; e_1{}^t] \cdot X_k^{(2)t}$

$\Delta a_k^{(2)} = \sum_t [e_0{}^t \; e_1{}^t] \cdot X_k^{(2)t}$

$\Delta a^{(2)} = \begin{bmatrix} \sum_t [e_0{}^t \; e_1{}^t] \cdot X_0^{(2)t} \\ \sum_t [e_0{}^t \; e_1{}^t] \cdot X_1^{(2)t} \\ \sum_t [e_0{}^t \; e_1{}^t] \cdot X_2^{(2)t} \\ \sum_t [e_0{}^t \; e_1{}^t] \cdot X_3^{(2)t} \\ \vdots \end{bmatrix}$ ⇔ to matrix multiplication $\begin{bmatrix} X_0^{(2)t} \\ X_1^{(2)t} \\ X_2^{(2)t} \\ \vdots \end{bmatrix} \cdot [e_0{}^t \; e_1{}^t]$

→ $X^{(2)t\,T} \cdot err$

$\Delta a^{(2)} = X^{(2)t\,T} \cdot err^t$ → matrix multiplication

$err^t = [e_0^t \; e_1{}^t]$

$\Delta a^{(1)}_{j,k} = -\eta \frac{\partial E}{\partial a^{(1)}_{j,k}}$ $\quad \frac{\partial E}{\partial a^{(1)}_{j,k}} = \partial\left(\frac{1}{2}\sum_t \sum_m (r_m{}^t - O_m^{(3)t})^2\right)$

$= \sum_t \sum_m (r_m{}^t - O_m^{(3)t}) \cdot -1 \cdot \frac{\partial O_m^{(3)t}}{\partial a^{(1)}_{j,k}}$ $\quad \frac{\partial O_m^{(3)t}}{\partial a^{(1)}_{j,k}} = \frac{\partial\left(\sum_{k=1} a^{(2)}_{k,m} \cdot X_k^{(2)t}\right)}{\partial a^{(1)}_{j,k}}$

$= a^{(2)}_{k,m} \cdot \frac{\partial X_k^{(2)t}}{\partial a^{(1)}_{j,k}}$ $\quad \frac{\partial X_k^{(2)t}}{\partial a^{(1)}_{j,k}} = \frac{\partial(\sigma(O_k^{(2)t}))}{\partial a^{(1)}_{j,k}} = \sigma(O_k^{(2)t})(1-\sigma(O_k^{(2)t})) \frac{\partial O_k^{(2)t}}{\partial a^{(1)}_{j,k}}$

$\frac{\partial O_k^{(2)t}}{\partial a^{(1)}_{j,k}} = \frac{\partial\left(\sum_{j=1} a^{(1)}_{j,k} \cdot X_j^{(1)t}\right)}{\partial a^{(1)}_{j,k}} = X_j^{(1)t}$

$\Delta a^{(1)}_{j,k} = +\eta \sum_t \sum_m e_m{}^t \cdot +1 \cdot a^{(2)}_{k,m} \cdot \sigma(O_k^{(2)t})(1-\sigma(O_k^{(2)t})) \cdot X_j^{(1)t}$

$\alpha(O_k^{(2)t}) = \sigma(O_k^{(2)t})(1-\sigma(O_k^{(2)t}))$

$\Delta a^{(1)}_{j,k} = \eta \sum_t \sum_m e_m{}^t \cdot a^{(2)}_{k,m} \alpha(O_k^{(2)t}) \cdot X_j^{(1)t}$

↳ open m in vector form

$\Delta a^{(1)}_{j,k} = \eta \cdot \sum_t [e_0{}^t \; e_1{}^t] [a^{(2)}_{k,1} \; a^{(2)}_{k,2}] \cdot \alpha(O_k^{(2)t}) \cdot X_j^{(1)t}$

→ $[a^{(2)}_{0,1} \; a^{(2)}_{1,2}]$ → excluded !

release k ↳ $= \eta \sum_t [e_0 \; e_1] \begin{bmatrix} a^{(2)}_{1,1} & a^{(2)}_{1,2} \\ a^{(2)}_{2,1} & a^{(2)}_{2,2} \\ a^{(2)}_{3,1} & a^{(2)}_{3,2} \\ a^{(2)}_{4,1} & a^{(2)}_{4,2} \\ a^{(2)}_{5,1} & a^{(2)}_{5,2} \end{bmatrix} \cdot [\alpha(O_1^{(2)t}) \; \alpha(O_2^{(2)t}) \; \alpha(O_3^{(2)t}) \; O_5^{(2)t}] \cdot X_j^{(1)t}$

→ $a^{(1)}$ → 5 × 5 matrix

j is the row number width must be 5.

this vector already in dimension 5

This part should be similar

$\Delta a^{(1)}_{j,k} = \eta \sum_t err^t \cdot a^{(2)T} \odot \alpha(O^{(2)t}) \cdot X_j^{(1)t}$

$\eta \cdot \sum_t \left(\underbrace{err}_{\text{so}} \cdot a^{(2)T}\right) \odot [\alpha(O_1^{(2)t}) \; \alpha(O_2^{(2)t}) \cdots O_5^{(2)t}] \cdot X_j^{(1)t}$

↓ matrix multiplication → (1,5) ; ⊙ → pairwise ; → 1,5

$$\Delta a^{(1)}_{j,k} = \eta \sum_t \sum_m e_m^t \cdot a^{(2)}_{k,m} \alpha(O_k^{(2)t}) \cdot X_j^{(1)t}$$

↳ vectorize with m

$$\Delta a^{(1)}_{j,k} = \eta \sum_t [e_0^t \; e_1^t] [a^{(2)}_{k,0} \; a^{(2)}_{k,1}] \alpha(O_k^{(2)t}) \cdot X_j^{(1)t}$$

$[a^{(2)}_{0,0} \; a^{(2)}_{0,1}]$ → excluded

vectorize k

$$\Delta a^{(1)}_j = \eta \sum_t [e_0 \; e_1^t] \cdot \begin{bmatrix} a^{(2)}_{1,0} & a^{(2)}_{1,1} \\ a^{(2)}_{2,0} & a^{(2)}_{2,1} \\ a^{(2)}_{3,0} & a^{(2)}_{3,1} \\ a^{(2)}_{4,0} & a^{(2)}_{4,1} \\ a^{(2)}_{5,0} & a^{(2)}_{5,1} \end{bmatrix} [\alpha(O_1^{(2)t} \; \alpha(O_2^{(2)t}) \ldots \alpha(O_5^{(2)t}]] \cdot X_j^{(1)t}$$

↳ this a vector and row dimension is 5

whole expression here must be a row colum

summation could be obta

$$(e_0^t \cdot a^{(2)t}_{1,0} + e_1^t a^{(2)}_{1,1}) \cdot \alpha(O_1^{(2)t})$$

$[e_0^t \; e_1^t]$ pairwisely could be multiplied to $a^{(2)}$ but error could be an matrix like

$$\begin{bmatrix} e_0^0 & e_1^0 \\ e_0^1 & e_1^1 \\ e_0^2 & e_1^2 \end{bmatrix}$$ → so pairwise multiplication could not be used

So matrix multiplication should be used.

↳ Instead let's look $err.a^{(2)T} \; [e_0^t \; e_1^t] \begin{bmatrix} a_{1,0} & a_{2,0} & a_{3,0} & a_{4,0} & a_{5,0} \\ a_{1,1} & a_{2,1} & a_{3,1} & a_{4,1} & a_{5,1} \end{bmatrix}$

↳ this produces 5 × sample count

↳ $\alpha(O^{(2)t})$ → is also 5 × sample count therefore pairwisely could be done.

$$\Delta a^{(1)}_j = err.a^{(2)T} \odot \alpha(O^{(2)t}) \cdot X_j^{(1)t}$$

$$\Delta a^{(1)}_j = [(e_0^t \cdot a^{(2)}_{1,0} + e_1^t a^{(2)}_{1,1}) \cdot \alpha(O_1^{(2)t}) \quad (e_0^t a_{2,0} + e_1^t a^{(2)}_{2,1}) \alpha(O_2^{(2)t}) \; \ldots]$$

vectorize j

$$\Delta a^{(1)} = \begin{bmatrix} err.a^{(2)T} \odot \alpha(O^{(2)t}) \cdot X_0^{(1)t} \\ err.a^{(2)T} \odot \alpha(O^{(2)t}) \cdot X_1^{(1)t} \\ err.a^{(2)T} \odot \alpha(O^{(2)t}) \cdot X_2^{(1)t} \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} X_0^{(1)t} \\ X_1^{(1)t} \\ X_2^{(1)t} \end{bmatrix} \begin{bmatrix} (err.a^{(2)T}) \odot \alpha(O^{(2)t}) \end{bmatrix}$$

$$\Delta a^{(1)} = \sum_t X^{(1)T} \cdot ((err.a^{(2)T}) \odot \alpha(O^{(2)t}))$$

$q^{(2)} \cdot err^T$ could be done

$$\begin{bmatrix} a^{(2)}_{1,0} & a^{(2)}_{1,1} \\ a^{(2)}_{2,0} & a^{(2)}_{2,1} \\ a^{(2)}_{3,0} & a^{(2)}_{3,1} \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} e_0^0 & e_0^1 \ldots \\ e_1^0 & e_1^1 \end{bmatrix}$$

↳ this produ a matrix

from example

$[\;]_{5 \times \text{sample number}}$

but $\alpha(O_2^{(2)})$ ↳ is in dimens 5 × samp coun

↙ So to multiply b extra transpose operations shou be performed

1.)

$$\Delta a^{(1)}_{j,k} = \eta \sum_t \sum_m e_m^t \cdot a^{(2)}_{k,m} \, \alpha(O_k^{(2)t}) \cdot X_j^{(1)t}$$

$$\Delta a^{(1)}_{0,1} = \sum_t \left(e_0^t \cdot a^{(2)}_{1,0} + e_1^t a^{(2)}_{1,1}\right) \cdot \alpha(O_1^{(2)t}) \cdot X_0^{(1)t}$$

$$\Delta a^{(1)}_{0,2} = \sum_t \left(e_0^t \cdot a^{(2)}_{2,0} + e_1^t \cdot a^{(2)}_{2,1}\right) \cdot \alpha(O_2^{(2)t}) \cdot X_0$$

$$\Delta a^{(1)}_{0,3} = \sum_t \left(e_0 \cdot a^{(2)}_{3,0} + e_1^t \cdot a^{(2)}_{3,1}\right) \cdot \alpha(O_3^{(2)t}) \cdot X_0$$

$$\vdots$$

$$\Delta a^{(1)}_{1,1} = \sum_t \left(e_0^t \cdot a^{(2)}_{1,0} + e_1^t \cdot a^{(2)}_{1,1}\right) \alpha(O_0^{(2)}) \cdot X_1^{(1)t}$$

$$\Delta a^{(1)}_{1,2} = \sum \left(e_0^t \cdot a^{(2)}_{2,0} + e_1^t \cdot a^{(2)}_{2,1}\right) \alpha(O_2^{(2)t}) \cdot X_1^{(1)t}$$

$$\Delta a^{(1)} = \begin{bmatrix} \Delta a^{(1)}_{0,1} & \Delta a^{(1)}_{0,2} & \Delta a^{(1)}_{0,3} & \Delta a^{(1)}_{0,4} \\ \Delta a^{(1)}_{1,1} & a^{(1)}_{1,2} & \Delta a^{(1)}_{1,3} & \Delta a^{(1)}_{1,3} \\ \Delta a^{(1)}_{2,1} & \Delta a^{(1)}_{2,2} & \Delta a^{(1)}_{2,3} & \Delta a^{(1)}_{2,3} \\ \cdot & & & \end{bmatrix}$$

$$\sum_t \left(e_0^t \cdot a^{(2)}_{1,0} + e_1^t \cdot a^{(2)}_{1,1}\right) \cdot \alpha(O_1^{(2)t}) \cdot X_0^{(1)t} \quad \sum_t \left(e_0^t \cdot a^{(2)}_{2,0} + e_1^t a^{(2)}_{2,1}\right) \alpha(O_2^{(2)t}) \cdot X_0^{(1)t} \quad \cdots$$

$$\sum_t \left(e_0^t a^{(2)}_{1,0} + e_1^t \cdot a^{(2)}_{1,1}\right) \alpha(O_1^{(2)t}) \cdot X_1^{(1)t} \quad \sum \left(e_0^t \cdot a^{(1)}_{2,0} + e_1^t \cdot a^{(2)}_{2,1}\right) \cdot \alpha(O_2^{(2)t}) \cdot X_1^{(1)t} \quad \cdots$$

$$\vdots \qquad\qquad \vdots$$

$$\begin{bmatrix} e^t \cdot a_1^{(2)T} \alpha(O_1^{(2)t}) & e^t \cdot a_2^{(2)T} \alpha(O_2^{(2)t}) & e^t \cdot a_3^{(2)T} \alpha(O_3^{(2)t}) & \cdots \\ e^t \cdot a_1^{(2)T} \alpha(O_1^{(2)t}) & e^t \cdot a_2^{(2)T} \alpha(O_2^{(2)t}) & e^t a_3^{(2)T} \cdot \alpha(O_3^{(2)t}) & \cdots \\ \vdots & & & \end{bmatrix}$$

$$\cdot \left(e^t a^{(2)T} \odot \alpha(O^{(2)})\right) X_{\cdot j}^{(1)t}$$

$$\Delta a^{(1)}_{j,k} = \begin{bmatrix} \left(e^t \cdot a^{(2)T}\right) \odot \alpha(O^{(2)t}) \cdot X_0^{(1)t} \\ \quad \cdot X_1^{(1)t} \\ \quad \cdot X_2^{(1)t} \\ \vdots \end{bmatrix} \rightarrow \sum_t \begin{bmatrix} X_0^{(1)t} \\ X_1^{(1)t} \\ X_2^{(1)t} \\ X_3^{(1)t} \\ \vdots \end{bmatrix} \left[ \left(e^t \cdot a^{(2)T}\right) \odot \alpha(O^{(2)t}) \right]$$

$$\Delta a^{(1)} = \sum_t X^{(1)tT} \cdot \left( \left( e^t \cdot a^{(2)T} \right) \odot \alpha(O^{(2)t}) \right)$$

$$\Delta a_{i,j}^{(0)} = \eta \frac{\partial E}{\partial a_{i,j}^{(0)}} = \eta \frac{\partial \left( \frac{1}{2} \sum_t \sum_m (e_m^t - o_m^t)^2 \right)}{\partial a_{i,j}^{(0)}} = \eta \sum_t \sum_m (e_m^t - o_m^t) \frac{\partial o_m^t}{\partial a_{i,j}^{(0)}}$$

$$\frac{\partial o_m^{(2)t}}{\partial a_{i,j}^{(0)}} = \frac{\partial \left( \sum_{k=1} a_{k,m}^{(2)} \cdot x_k^{(2)t} \right)}{\partial a_{i,j}^{(0)}} = \sum_{k=1} a_{k,m}^{(2)} \cdot \frac{\partial x_k^{(2)t}}{\partial a_{i,j}^{(0)}} \quad \frac{\partial x_k^{(2)t}}{\partial a_{i,j}^{(0)}} = \frac{\partial (\sigma(o_k^{(2)t})}{\partial a_{i,j}^{(0)}}$$

$$\frac{\partial \sigma(o_k^{(2)t})}{\partial a_{i,j}^{(0)}} = \sigma(o_k^{(2)t})(1 - \sigma(o_k^{(2)t})) \frac{\partial o_k^{(2)t}}{\partial a_{i,j}^{(0)}} \quad \frac{\partial o_k^{(2)t}}{\partial a_{i,j}^{(0)}} = \frac{\partial \left( \sum_j a_{j,k}^{(1)} \cdot x_j^{(1)t} \right)}{\partial a_{i,j}^{(0)}}$$

$$= a_{j,k}^{(1)} \cdot \frac{\partial x_j^{(1)t}}{\partial a_{i,j}^{(0)}} \quad \frac{\partial x_j^{(1)t}}{\partial a_{i,j}^{(0)}} = \sigma(o_j^{(1)t})(1 - \sigma(o_j^{(1)t})) \cdot \frac{\partial o_j^{(1)t}}{\partial a_{i,j}^{(0)}} \quad \frac{\partial o_j^{(1)t}}{\partial a_{i,j}^{(0)}} = \frac{\partial \left( \sum_i a_{i,j}^{(0)} \cdot x_i^{(0)t} \right)}{\partial a_{i,j}^{(0)}}$$

$$= x_i^{(0)t}$$

$$\Delta a_{i,j}^{(0)} = \eta \sum_t \sum_m e_m^t \cdot \sum_{k=1} a_{k,m}^{(2)} \cdot \alpha(o_k^{(2)t}) \cdot a_{j,k}^{(1)} \cdot \alpha(o_j^{(1)t}) \cdot x_i^{(0)t}$$

vectorize $\Delta a_{i,j}^{(0)} = \eta \sum_t [e_0^t \; e_1^t] \sum_{k=1} \begin{bmatrix} a_{k,0}^{(2)} & a_{k,1}^{(2)} \end{bmatrix} \cdot \alpha(o_k^{(2)t}) \cdot a_{j,k}^{(1)} \cdot \alpha(o_j^{(1)t}) \cdot x_i^{(0)t}$

$$\Delta a_{i,j}^{(0)} = \eta \sum_t [e_0^t \; e_1^t] \Big[ [a_{1,0}^{(2)} \; a_{1,1}^{(2)}] \cdot \alpha(o_1^{(2)t}) \cdot a_{j,1} + [a_{2,0}^{(2)} \; a_{2,1}^{(2)}] \cdot \alpha(o_2^{(2)t}) \cdot a_{j,2}^{(2)}$$

excluded $+ [a_{3,0}^{(2)} \; a_{3,1}^{(2)}] \cdot \alpha(o_3^{(2)t}) \cdot a_{j,3}^{(2)} + [a_{4,0}^{(2)} \; a_{4,1}^{(2)}] \cdot \alpha(o_4^{(2)t}) \cdot a_{j,4} + [a_{5,0}^{(2)} \; a_{5,1}^{(2)}] \cdot \alpha(o_5^{(2)t}) \cdot a_{j,5}^{(2)} \Big]$

$\cdot \alpha(o_j^{(1)t})$

$\cdot x_i^{(0)t}$

$a_{0,0} \quad a_{0,1}$

$$a^{(2)} = \begin{bmatrix} a_{1,0} & a_{1,1} \\ a_{2,0} & a_{2,1} \\ a_{3,0} & a_{3,1} \\ a_{4,0} & a_{4,1} \\ a_{5,0} & a_{5,1} \end{bmatrix} \qquad a^{(1)} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} \end{bmatrix}$$

~~$\Delta a_{i,j}^{(0)} = [e_0^t \; e_1^t] [a^{(2)T} \cdot \alpha(o^{(2)t}) \cdot a$~~

$$\Delta a_{i,j}^{(0)} = \sum_t [e_0^t \; e_1^t] \cdot \begin{bmatrix} a_{1,0}^{(2)} & a_{2,0}^{(2)} & a_{3,0}^{(2)} & a_{4,0}^{(2)} & a_{5,0}^{(2)} \\ a_{1,1}^{(2)} & a_{2,1} & a_{3,1}^{(2)} & a_{4,1}^{(2)} & a_{5,1} \end{bmatrix} \odot \alpha(o^{(2)}) \cdot \begin{bmatrix} a_{j,1}^{(1)} \\ a_{j,2}^{(1)} \\ a_{j,3}^{(2)} \\ a_{j,4}^{(2)} \\ a_{j,5}^{(2)} \end{bmatrix} \cdot \alpha(o_j^{(1)t}) \cdot x_i^{(0)t}$$

rebe j

$$\Delta a_{i,j}^{(0)} = \sum_t [e_0^t \; e_1^t] \begin{bmatrix} a_{1,0}^{(2)} & a_{2,0}^{(2)} & a_{3,0}^{(2)} & a_{4,0}^{(2)} & a_{5,0}^{(2)} \\ a_{1,1}^{(2)} & a_{2,1}^{(2)} & a_{3,1}^{(2)} & a_{4,1}^{(2)} & a_{5,1}^{(2)} \end{bmatrix} \odot \alpha(o^{(2)}) \cdot \begin{bmatrix} a_{1,1}^{(1)} & a_{2,1}^{(1)} & a_{3,1}^{(1)} & a_{4,1}^{(1)} \\ a_{1,2}^{(1)} & a_{2,2}^{(1)} & a_{3,2}^{(1)} & a_{4,2}^{(1)} \\ a_{1,3}^{(1)} & a_{2,3}^{(1)} & a_{3,3}^{(1)} & a_{4,3}^{(1)} \\ a_{1,4}^{(1)} & a_{2,4}^{(1)} & a_{3,4}^{(1)} & a_{4,4}^{(1)} \\ a_{1,5}^{(1)} & a_{2,5}^{(1)} & a_{3,5}^{(1)} & a_{4,5}^{(1)} \end{bmatrix} [\alpha(o_1^{(1)t} \; \alpha(o_2^{(1)t}) \cdot \ldots \alpha(o_4^{(1)t}$$

$\cdot x_i^{(0)t}$

$$\begin{bmatrix} a_{0,1}^{(1)} \\ a_{0,2}^{(1)} \\ a_{0,3}^{(1)} \\ a_{0,4}^{(1)} \\ a_{0,5}^{(1)} \end{bmatrix}$$ → column is excluded !

$$\Delta a_i^{(0)} = \sum_t err \cdot a^{(2)T} \cdot \odot \alpha(o^{(2)t}) \cdot a^{(1)T} \cdot \odot \alpha(o_1^{(1)t}) \cdot x_1^{(0)t}$$

$$\Delta a^{(0)} = \sum_t \begin{bmatrix} err \cdot a^{(2)T} \odot \alpha(o^{(2)t}) \; a^{(1)T} \odot \alpha(o^{(1)t}) \cdot x_0^{(0)t} \\ \quad \cdot x_1^{(0)t} \\ \quad \cdot x_2^{(0)t} \\ \quad \cdot x_3^{(0)t} \end{bmatrix}$$

$$\Delta a^{(0)} = \sum_t x^{(0)tT} \cdot \left( \left( (err \cdot a^{(2)T}) \odot \alpha(o^{(2)t}) \right) \cdot a^{(1)T} \right) \odot \alpha(o^{(1)t}) \Big)$$

6.

$$\Delta a^{(2)} = \eta \cdot X^{(2)t^T} \cdot Err^t$$ → only for one sample

$$\Delta a^{(1)} = \eta \cdot X^{(1)t^T} \cdot \left( \left( Err^t \cdot a^{(2)T} \right) \odot \alpha'(O^{(2)t}) \right)$$

$$\Delta a^{(0)} = \eta \sum_t X^{(0)t^T} \cdot \left( \left( \left( \left( err \cdot a^{(2)T} \right) \odot \alpha'(O^{(2)t}) \right) \cdot a^{(1)T} \right) \odot \alpha'(O^{(1)t}) \right)$$

$Err^t \rightarrow$ Output layer error$^t$

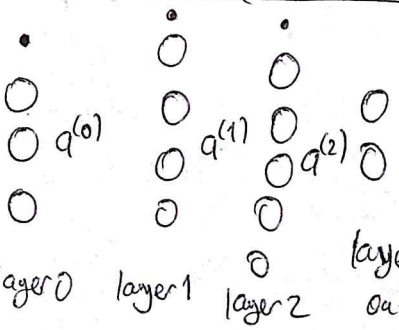
$d_0^{(2)}$ → is excluded

Layer 2 error → $\left( \left( \text{Output layer error}^t \right) \cdot a^{(2)T} \right) \odot \alpha'(O^{(2)t})$

Layer 1 error → $\left( \left( \text{Layer 2 error}^t \right) \cdot a^{(1)T} \right) \odot \alpha'(O^{(1)t})$

$a_0^{(1)}$ → is excluded

$$\boxed{Err^{(\ell)t} = \left( Err^{(\ell+1)t} \cdot a^{(\ell)T} \right) \odot \alpha'(O^{(\ell)t})}$$




$$\boxed{\Delta a^{(\ell)} = \eta \cdot X^{(\ell)t} \cdot Err^{(\ell+1)t}}$$

To deal with more than one sample we need to construct a matrix.

$$X = \begin{bmatrix} +1 & X_1^0 & X_2^0 & X_3^0 \\ +1 & X_1^1 & X_2^1 & X_3^1 \\ +1 & X_1^2 & X_2^2 & X_3^2 \end{bmatrix}$$

$X \cdot a^{(0)} =$ (Later added)

$$\begin{bmatrix} +1 & O_1^{(1)0} & O_2^{(1)0} & O_3^{(1)0} & O_4^{(1)0} \\ +1 & O_1^{(1)1} & O_2^{(1)1} & O_3^{(1)1} & O_4^{(1)1} \\ +1 & O_1^{(1)2} & O_2^{(1)2} & O_3^{(1)2} & O_4^{(1)2} \end{bmatrix} \rightarrow \text{Layer 1 Output (L1\_Output)}$$

$Err = r - O \rightarrow \begin{bmatrix} e_0^0 & e_1^0 \\ e_0^1 & e_1^1 \\ e_0^2 & e_1^2 \end{bmatrix}$

$\Delta a^{(2)} = \eta X^{(2)t} \cdot Err^t$ → for a given sample

$$\begin{bmatrix} +1 & X_1^{(2)0} & X_2^{(2)0} & X_3^{(2)0} & X_4^{(2)0} & X_5^{(2)0} \\ +1 & X_1^{(2)1} & X_2^{(2)1} & X_3^{(2)1} & X_4^{(2)1} & X_5^{(2)1} \\ +1 & X_1^{(2)2} & X_2^{(2)2} & X_3^{(2)2} & X_4^{(2)2} & X_5^{(2)2} \end{bmatrix} \rightarrow X^{(2)} \text{ for given 3 samples}$$

$L1\_Output \cdot a^{(1)} =$ ← Layer 2 output

$$\begin{bmatrix} +1 & O_1^{(2)0} & O_2^{(2)0} & O_3^{(2)0} & O_4^{(2)0} & O_5^{(2)0} \\ +1 & O_1^{(2)1} & O_2^{(2)1} & O_3^{(2)1} & O_4^{(2)1} & O_5^{(2)1} \\ +1 & O_1^{(2)2} & O_2^{(2)2} & O_3^{(2)2} & O_4^{(2)2} & O_5^{(2)2} \end{bmatrix}$$

Layer 2 Output $\cdot a^{(2)} =$

$$\begin{bmatrix} O_0^{(3)0} & O_1^{(3)0} \\ O_0^{(3)1} & O_1^{(3)1} \\ O_0^{(3)2} & O_1^{(3)2} \end{bmatrix} \rightarrow \text{Output layer}$$

$L1\_Output = \sigma(L1\_Output)$

$L2\_Output = \sigma(L2\_Output)$

$$a^{(2)} = \begin{bmatrix} +1 \\ X_1^{(2)0} \\ X_2^{(2)0} \\ X_3^{(2)0} \\ X_4^{(2)0} \\ X_5^{(2)0} \end{bmatrix} \begin{bmatrix} e_0^0 & e_1^0 \end{bmatrix} + \begin{bmatrix} 1 \\ X_1^{(2)1} \\ X_2^{(2)1} \\ X_3^{(2)1} \\ X_4^{(2)1} \\ X_5^{(2)1} \end{bmatrix} \begin{bmatrix} e_0^1 & e_1^1 \end{bmatrix} + \begin{bmatrix} 1 \\ X_1^{(2)2} \\ X_2^{(2)2} \\ X_3^{(2)2} \\ X_4^{(2)2} \\ X_5^{(2)2} \end{bmatrix} \begin{bmatrix} e_0^2 & e_1^2 \end{bmatrix}$$

$$r = \begin{bmatrix} r_0^0 & r_1^0 \\ r_0^1 & r_1^1 \\ r_0^2 & r_1^2 \end{bmatrix}$$

Output layer = Softmax( . Output layer).

$$\Delta a^{(2)} = \begin{bmatrix} +1 \\ X_1^{(2)0} \\ X_2^{(2)0} \\ X_3^{(2)0} \\ X_4^{(2)0} \\ X_5^{(2)0} \end{bmatrix} \begin{bmatrix} e_0^0 & e_1^0 \end{bmatrix} + \begin{bmatrix} +1 \\ X_1^{(2)1} \\ X_2^{(2)1} \\ X_3^{(2)1} \\ X_4^{(2)1} \\ X_5^{(2)1} \end{bmatrix} \begin{bmatrix} e_0^1 & e_1^1 \end{bmatrix} + \begin{bmatrix} +1 \\ X_1^{(2)2} \\ X_2^{(2)2} \\ X_3^{(2)2} \\ X_4^{(2)2} \\ X_5^{(2)2} \end{bmatrix} \begin{bmatrix} e_0^2 & e_1^2 \end{bmatrix}$$

$$\Delta a^{(2)} = \sum_t X^{(2)t T} \cdot err^t$$

→ instead multiplying individual input and one error we can combine them to matrix

equals

$$\begin{bmatrix} 1 & 1 & 1 \\ X_1^{(2)0} & X_1^{(2)1} & X_1^{(2)2} \\ X_2^{(2)0} & X_2^{(2)1} & X_2^{(2)2} \\ X_3^{(2)0} & X_3^{(2)1} & X_3^{(2)2} \\ X_4^{(2)0} & X_4^{(2)1} & X_4^{(2)2} \\ X_5^{(2)0} & X_5^{(2)1} & X_5^{(2)2} \end{bmatrix} \begin{bmatrix} e_0^0 & e_1^0 \\ e_0^1 & e_1^1 \\ e_0^2 & e_1^2 \end{bmatrix} = \begin{bmatrix} (e_0^0 + e_0^1 + e_0^2) & (e_1^0 + e_1^1 + e_1^2) \\ (X_1^{(2)0} \cdot e_0^0 + X_1^{(2)1} \cdot e_0^1 + X_1^{(2)2} \cdot e_0^{(2)}) & (X_1^{(2)0} e_0^0 + X_1^{(2)1} \cdot e_0^1 + X_2^{(2)2} e_0^{(2)}) \\ (X_2^{(2)0} \cdot e_0^0 + X_2^{(2)1} e_0^1 + X_2^{(2)2} e_0^2) & (X_2^{(2)0} e_1^0 + X_2^{(2)1} \cdot e_1^1 + X_2^{(2)2} \cdot e_1^2) \\ \vdots & \\ \cdot & \\ \cdot & \\ \cdot & \end{bmatrix}$$

$$\Delta a^{(1)} = \sum_t X^{(1)t T} \cdot \left( \left( Err^t \cdot a^{(2)T} \right) \odot \alpha(O^{(2)t}) \right)$$

$$e^t = [e_0^t \; e_1^t] \qquad Err = \begin{bmatrix} e^0 \\ e^1 \\ e^2 \end{bmatrix} \qquad a^{(2)} = \begin{bmatrix} a_1^{(1)} \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} \qquad a_i^{(l)} = [a_{i,1}^{(l)} \; a_{i,2}^{(l)} \; \dots]$$

$$\begin{bmatrix} -e^0- \\ -e^1- \\ -e^2- \\ \cdots \end{bmatrix} \cdot \begin{bmatrix} a_1^{1(2)} & a_2^{1(2)} & a_3^{1(2)} & a_4^{1(2)} & a_5^{1(2)} \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \odot \begin{bmatrix} \alpha(O^{(2)})^0 \\ \alpha(O^{(2)})^1 \\ \alpha(O^{(2)})^3 \end{bmatrix}$$

$$\begin{bmatrix} (e^0 a_1^{(2)}) & (e^0 a_2^{(2)}) & (e^0 a_3^{(2)}) & (e^0 a_4^{(2)}) & (e^0 a_5) \\ e^1 a_1^{(2)} & e^1 a_2^{(2)} & e^1 a_3^{(2)} & e^1 a_4^{(2)} & e^1 a_5^{(2)} \\ e^2 a_1^{(2)} & e^2 a_2^{(2)} & e^2 a_3^{(2)} & e^2 a_4^{(2)} & e^2 a_5^{(2)} \end{bmatrix} \odot \begin{bmatrix} \alpha(O^{(2)})^0 \\ \alpha(O^{(2)})^1 \\ \alpha(O^{(2)})^2 \end{bmatrix}$$

$$X^{(1)T} = \begin{bmatrix} | & | & | \\ X^{(1)0} & X^{(1)1} & X^{(1)2} \\ | & | & | \end{bmatrix}$$

$$\Delta a^{(1)} = X^{(1)T} \cdot \begin{bmatrix} (e^0 \cdot a_1^{(2)} \cdot \alpha(O_1^{(2)0})) & e^0 a_2^{(2)} \cdot \alpha(O_2^{(2)0}) & e^0 a_3^{(2)} \cdot \alpha(O_3^{(2)0}) & e^0 a_4^{(2)} \alpha(O_4^{(2)0}) & (e^0 a_5^{(2)} \cdot \alpha O_5^{(2)0}) \\ e^1 a_1^{(2)} \cdot \alpha(O_1^{(2)1}) & e^1 \cdot a_2^{(2)} \cdot \alpha(O_2^{(2)1}) & e^0 a_3^{(2)} \cdot \alpha(O_3^{(2)1}) & (e^0 a_4^{(2)} \alpha(O_4^{(2)1})) & (e^0 \cdot a_5^{(2)} \cdot \alpha O^{(2)1}) \\ e^2 a_4^{(2)} \alpha(O_4^{(2)2}) & e^2 \cdot a_2^{(2)} \alpha(O_2^{(2)2}) & e^2 a_3^{(2)} \cdot \alpha(O_3^{(2)2}) & e^2 a_4^{(2)} \cdot \alpha(O_4^{(2)2}) & e^0 a_5^{(2)} \alpha(O^{(2)2}) \end{bmatrix}$$

↑ layer 2 error

$$\beta_x^t = e^t a_x^{(2)} \alpha(O x^{(2)t})$$

$$\Delta a^{(1)} = \begin{bmatrix} (\beta_1^0 \cdot 1 + \beta_1^1 \cdot 1 + \beta_1^2) & (\beta_2^0 + \beta_2^1 + \beta_2^2) & (\beta_3^0 + \beta_3^1 \beta_3^2) & (\beta_4^0 + \beta_4^1 + \beta_4) & (\beta_5^0 + \beta_5^1 + \beta_5^2) \\ (\beta_1^0 X_1^{(1)0} + \beta_1^1 \cdot X_1^{(1)1} + \beta_1^2 X_1^{(1)2}) & (\beta_2^0 X_1^{(1)0} + \beta_2^1 X_1^{(1)1} + \beta_2^2 X_1^{(1)2}) & \cdots & \cdots & \cdots \\ (\beta_1^0 X_2^{(1)0} + \beta_1^1 X_2^{(1)1} + \beta_1^2 X_2^{(1)2}) & \beta_2^0 \cdot X_2^{(1)0} + \beta_2^1 \cdot X_2^{(1)1} + \beta_2^2 X_2^{(1)2}) & \cdots & \cdots & \cdots \\ (\beta_1^0 X_3^{(1)0} + \beta_1^1 X_3^{(1)1} + \beta_1^2 X_3^{(1)2}) & \vdots & \cdots & \cdots & \cdots \end{bmatrix}$$

5.)

$$\Delta a_{j,k}^{(1)} = \eta \sum_t \sum_m e_m^t \cdot a_{k,m}^{(2)} \cdot \alpha(O_k^{(2)t}) \cdot X_j^{(1)t}$$

$$= \eta \sum_t \underbrace{\left(e^t \cdot a_k^{(2)}\right) \cdot \alpha(O_k^{(2)t})}_{\text{defined previously } \beta_k^t} \cdot X_j^{(1)t}$$

$$\Delta a_{ij}^{(1)} = \eta \cdot \begin{bmatrix} \sum_t \beta_0^t \cdot X_0^{(1)t} & \sum_t \beta_2^t \cdot X_0^{(1)t} & \sum_t \beta_3^t X_0^{(1)t} & \sum_t \beta_4^t \cdot X_0^{(1)t} & \sum_t \beta_5^t X_0^{(1)t} \\ \sum_t \beta_1^t \cdot X_1^{(1)t} & \sum_t \beta_2^t \cdot X_1^{(1)t} & \sum_t \beta_3^t X_4^{(1)t} & & \\ \sum_t \beta_1^t \cdot X_2^{(1)t} & \sum_t \beta_2^t \cdot X_2^{(1)t} & \sum_t \beta_3^t X_2^{(1)t} & & \\ \sum_t \beta_1^t \cdot X_3^{(1)t} & & & & \end{bmatrix}$$

Same thing represented on previous page.

---

$$\Delta a_{i,j}^{(0)} = \eta \sum_t \underbrace{\sum_m e_m^t \cdot \sum_{k=1} a_{k,m}^{(2)} \cdot \alpha(O_k^{(2)t})}_{\text{layer 2 error}} \cdot a_{j,k}^{(1)} \cdot \alpha(O_j^{(1)t}) \cdot X_i^{(0)t}$$

this will produce

If we represent layer 2 error $\rightarrow \begin{bmatrix} e'^0 \\ e'^1 \\ e'^2 \end{bmatrix}$

$$\Delta a^{(0)} = \sum_t X^{(0)T} \left( \left( (\text{err} \cdot a^{(2)T}) \odot \alpha(O^{(2)t}) \right) \cdot a^{(1)T} \right) \odot \alpha(O^{(1)t})$$

$$\downarrow \begin{bmatrix} -e'^0- \\ -e'^1- \\ -e'^2- \end{bmatrix}$$

$$\Delta a^{(0)} = \sum_t X^{(0)T} \cdot \left( \underbrace{\text{err of prev layer}}_{?} \cdot a^{(1)T} \right) \odot \alpha(O^{(1)t})$$

$$\begin{bmatrix} -e'^0- \\ e'^1 \\ -e'^2- \end{bmatrix} \begin{bmatrix} a_1^{(1)} & a_2^{(1)} & a_3^{(1)} & a_4^{(1)} \end{bmatrix} \odot \begin{bmatrix} -\alpha(O^{(1)0})- \\ -\alpha(O^{(1)1})- \\ -\alpha(O^{(1)2})- \end{bmatrix}$$

$$X^{(0)T} = \begin{bmatrix} | & | & | \\ X^{(0)0} & X^{(0)1} & X^{(0)2} \\ | & | & | \end{bmatrix} \begin{bmatrix} (e'^0 a_1^{(1)}) & e'^0 a_2^{(1)} & e'^0 a_3^{(1)} & e'^0 a_4^{(1)} \\ (e'^1 a_1^{(1)}) & e'^1 a_2^{(1)} & e'^1 a_3^{(1)} & e'^1 a_4^{(1)} \\ (e'^2 a_1^{(1)}) & e'^2 a_2^{(1)} & e'^2 a_3^{(1)} & e'^2 a_4^{(1)} \end{bmatrix} \odot \begin{bmatrix} -\alpha(O^{(1)0})- \\ -\alpha(O^{(1)1})- \\ -\alpha(O^{(1)2})- \end{bmatrix}$$

$$X^{(0)T} \cdot \begin{bmatrix} e'^0 a_1^{(1)} \cdot \alpha(O_1^{(1)0}) & e'^0 a_2^{(1)} \cdot \alpha(O_2^{(1)0}) & e'^0 a_3^{(1)} \cdot \alpha(O_3^{(1)0}) & e'^0 a_4^{(1)} \cdot \alpha(O_4^{(1)0}) \\ e'^1 a_1^{(1)} \alpha(O_1^{(1)1}) & e'^1 a_2^{(1)} \alpha(O_2^{(1)1}) & e'^1 a_3^{(1)} \cdot \alpha(O_3^{(1)1}) & e'^1 a_4^{(1)} \alpha(O_4^{(1)1}) \\ e'^2 a_1^{(1)} \alpha(O_1^{(1)2}) & e'^2 a_2^{(1)} \alpha(O_2^{(1)2}) & e'^2 a_3^{(1)} \alpha(O_3^{(1)2}) & e'^2 a_4^{(1)} \alpha(O_4^{(1)2}) \end{bmatrix}$$

Again $\bar{\beta}_x^t = e'^t a_x^{(1)} \alpha(O_x^{(1)t})$

$$\begin{bmatrix} (\bar{\beta}_1^0 + \bar{\beta}_1^1 + \bar{\beta}_1^2) & (\bar{\beta}_2^0 + \bar{\beta}_2^1 + \bar{\beta}_2^2) & (\bar{\beta}_3^0 + \bar{\beta}_3^1 + \bar{\beta}_3^2) & (\bar{\beta}_4^0 + \bar{\beta}_4^1 + \bar{\beta}_4^2) \\ (\bar{\beta}_1^0 \cdot X_1^{(0)0} + \bar{\beta}_1^1 X_1^{(0)1} + \bar{\beta}_1^2 \cdot X_1^{(0)2}) & (\bar{\beta}_2^0 \cdot X_1^{(0)0} + \bar{\beta}_2^1 X_1^{(0)1} + \bar{\beta}_2^2 \cdot X_1^{(0)2}) & & \\ (\bar{\beta}_1^0 X_2^{(0)0} + \bar{\beta}_1^1 \cdot X_2^{(0)1} + \bar{\beta}_2^2 X_2^{(0)2}) & \bar{\beta}_2^0 X_2^{(0)0} + \bar{\beta}_2^1 X_2^{(0)1} + \bar{\beta}_2^2 \cdot X_2^{(0)2} & & \\ \bar{\beta}_1^0 X_3^{(0)0} + \bar{\beta}_1^1 X_3^{(0)1} + \bar{\beta}_2 X_3^{(0)2}) & & & \end{bmatrix}$$

3.] $\Delta a_{i,j}^{(0)} = \eta \sum_t \sum_k e_k^t \, a_{j,k} \, \alpha(g^{(1)t}) \cdot X_i^{(0)t}$

$\Delta a_{i,j}^{(0)} = \eta \sum_t \underbrace{(e^t \cdot a_j^{(1)}) \, \alpha(O_j^{(1)t})}_{\beta_j^t} \cdot X_i^{(0)t}$

$$\Delta a_i^{(0)} = \eta \cdot \begin{bmatrix} \sum_t \bar{\beta}_0^t X_0^{(0)t} & \sum_t \bar{\beta}_2^t X_0^{(0)t} & \sum_t \bar{\beta}_3^t X_0^{(0)t} & \sum_t \bar{\beta}_4^t X_0^{(0)t} \\ \sum_t \bar{\beta}_0^t X_1^{(0)t} & \sum_t \bar{\beta}_2^t \cdot X_1^{(0)t} & \sum_t \bar{\beta}_3^t X_1^{(0)t} & \sum_t \bar{\beta}_4^t X_1^{(0)t} \\ \sum_t \bar{\beta}_0^t X_2^{(0)t} & \sum_t \bar{\beta}_2^t \cdot X_2^{(0)t} & \sum_t \bar{\beta}_3^t X_2^{(0)t} & \sum_t \bar{\beta}_4^t X_2^{(0)t} \end{bmatrix}$$

Same as on prev. page

It can also be shown that the same procedure can be applied to inner layers. By stacking matrix of inputs and labels Batch of samples could be used to update weight matrices simultaneously in one pass with matrix multiplications.

In previous we have shown that

$$\Delta a^{(2)} = \eta \, X^{(2)tT} \cdot Err^t$$
$$\Delta a^{(1)} = \eta \, X^{(1)tT} \cdot ((Err^t \cdot a^{(2)T}) \odot \alpha(O^{(2)t}))$$
$$\Delta a^{(0)} = \eta \cdot X^{(0)tT} \cdot (((((err\, a^{(2)T}) \odot \alpha(O^{(2)t})) \cdot a^{(1)T}) \odot \alpha(O^{(1)t}))$$

→ There are for one sample

So as we have shown previously we can stack samples in a matrix

$$X = \begin{bmatrix} -X^{(0)0}- \\ -X^{(0)1}- \\ -X^{(0)2}- \\ -X^{(0)3}- \end{bmatrix} \quad \text{where } X^{(0)t} = [X_0^{(0)t} \; X_1^{(0)t} \; X_2^{(0)t} \; X_3^{(0)t}]$$

$$r = \begin{bmatrix} -r^0- \\ -r^1- \\ -r^2- \\ -r^3- \end{bmatrix} \rightarrow \text{labels} \qquad r^t = [r_0^t \; r_1^t]$$

So. for batch

$$\Delta a^{(2)} = \eta \; X^{(2)T} \cdot Err$$
$$\Delta a^{(0)} = \eta \cdot X^{(0)T} \cdot (((Err \cdot a^{(2)T}) \odot \alpha(O^{(2)})) \cdot a^{(1)T}) \odot \alpha(O^{(1)}))$$
$$\Delta a^{(1)} = \eta \cdot X^{(1)T} \cdot ((Err \cdot a^{(2)T}) \odot \alpha(O^{(2)}))$$

$$\boxed{Err = O^{(3)} - r}$$

Output error $= O^{(3)} - r$

layer 2 error $= ((\text{Output Error}) \cdot a^{(2)T}) \odot \alpha(O^{(2)})$

layer 1 error $= ((\text{layer 2 error}) \cdot a^{(1)T}) \odot \alpha(O^{(1)})$

$\text{Error}^{(e)} = (Err^{(e+1)} \cdot a^{(e)T}) \odot \alpha(O^{(e)})$

$\Delta a^{(2)} = \eta \, X^{(2)T} \cdot$ Output Error

$\Delta a^{(1)} = \eta \cdot X^{(1)T} \cdot$ layer 2 error

$\Delta a^{(0)} = \eta \cdot X^{(0)T} \cdot$ layer 1 error

$\Delta a^{(e)} = \eta \, X^{(e)T} \cdot Err^{(e+1)}$

10.1

• → operation is used for matrix multiplication, scalar and matrix multiplication, scalar-scalar multiplication

(*) → operation is used for matrix-matrix, vector-vector elementwise multiplication

$$X=\begin{bmatrix}1&2\\3&4\end{bmatrix} \quad Y=\begin{bmatrix}5&6\\7&8\end{bmatrix} \quad X.Y=\begin{bmatrix}19&22\\43&50\end{bmatrix} \quad X \odot Y=\begin{bmatrix}1*5&2*6\\3*7&4*8\end{bmatrix}=\begin{bmatrix}5&12\\21&32\end{bmatrix}$$

$$\Delta a^{(2)} = \eta \cdot X^{(2)T} . ErrT$$

($\eta$: scalar matrix multiplication; $X^{(2)T}$ . ErrT: matrix multiplication)

$$\Delta a^{(1)} = \eta \cdot X^{(1)T} . ((Err . a^{(2)T}) \odot \alpha(O^{(2)}))$$

(matrix multiplication; pairwise multiplication)

- Parenthesis has the most high priority in calculation. First the terms inside the parenthesis are handled.

$$\Delta a^{(0)} = \eta \cdot X^{(0)T} . ((((Err . a^{(2)T}) \odot \alpha(O^{(2)}) . a^{(1)T}) \odot \alpha(O^{(1)}))$$

(scalar-matrix multiplication; matrix multiplication; pairwise multiplication)

- sigmoid(x) = $\sigma(x)$ $\quad \sigma(x) = \frac{1}{1+e^{-x}}$

$$\sigma(x)' = \left((1+e^{-x})^{-1}\right)' = -1 \cdot \frac{(1+e^{-x})'}{(1+e^{-x})^2} = \frac{+1.(e^{-x}.+1)}{(1+e^{-x})^2}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad \sigma(x) + \sigma(x).e^{-x} = 1 \qquad e^{-x} = \frac{1-\sigma(x)}{\sigma(x)}$$

$$\sigma(x)' = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)^2 . e^{-x} = \sigma(x)^2 . \frac{(1-\sigma(x))}{\sigma(x)} = \sigma(x).(1-\sigma(x))$$

$$\boxed{\sigma(x)' = \sigma(x)(1-\sigma(x))}$$

## Classification cost function Derivation

Consider we have n number of class in the given data set. At the output of the network we use softmax to obtain the probability distribution of each class. for each class our network a probability $p_i$ for ~~each class~~ We can think of this distribution as multinomial density.

$p_i = p(k_i \mid \theta)$ → represents our network parameters (our network). So we want our network to produce maximum likelihood estimation for each class in dataset. (Maximum Likelihood Estimation.)

$$p(c_1, c_2, c_3, \dots c_t) \prod_{i=1}^{t} p_i^{t c_i}$$

probability of whole data set classes.

→ because each class probability is independent.

$c_i$ = { 1 if current label is i; 0 otherwise }

11.) $P(c_1, c_2, c_2, \dots c_t) = \prod_{t=1}^{t} \prod_{i=1}^{k} p_i^{t c_i} = \prod_{t=1}^{t} \prod_{i=1}^{k} P(k_i|\theta)^{c_i}$

so we are looking for the optimum $\theta$ value which maximizes the probability. We could take log of this expression because max value location will not change.

$$P(\text{All set}) = \log \prod_{t=1}^{t} \prod_{i=1}^{k} P(k_i|\theta)^{c_i} = \sum_{t} \sum_{i=1}^{k} c_i \log P(k_i|\theta)$$

this maximization problem but lay flipping the graph we can change maximization problem to minimization problem so

$$-\sum_{t} \sum_{i=1}^{k} c_i \log(P(k_i|\theta))$$

$$E = J = -\sum_{t=1} \sum_{m} r_m^t \cdot \log(y_m^t)$$

$y_m^t$ → neural network output.

$P(k_i|\theta)$ represents the probability of any class at that given sample data. $c_i$ represents which class is the correct one.

Simply our network output a probability distribution

[0.1 0.2 0.3 0.4]
class A class B class C class D

$c_i$ represents if given data is class A
$c_i = [1\ 0\ 0\ 0]$
if class B
$c_i = [0\ 1\ 0\ 0]$
if class C
$c_i = [0\ 0\ 1\ 0]$
if D
$c_i = [0\ 0\ 0\ 1]$

when correct class' probability is less this causes error to be high. for ex.

[0.9 0.09 0.009 0.001]
and correct class be [0 0 0 1]

$c_i \cdot \log([0.9\ 0.09\ 0.009\ 0.001])$
↓
$[0\ 0\ 0\ 1] \cdot \log([0.9\ 0.09\ 0.009\ 0.001])$
↘
$\log(0.001)$
→ this is a negative but at front there is also one $-1$. so the error becomes positive and high.

$$\Delta w_{km}^{(2)} = -\eta \frac{\partial E}{\partial w_{km}^{(2)}} \qquad \frac{\partial E}{\partial w_{km}^{(2)}} = \left(-\sum_{t} \sum_{m} r_m^t \cdot \log(y_m^t)\right)$$

$$= \sum_{t} \sum_{m} -r_m^t \frac{\partial(\log(y_m^t))}{\partial w_{km}^{(2)}}$$

$$\frac{\partial \log(y_m^t)}{\partial w_{km}^{(2)}} = \frac{1}{y_m} \cdot \frac{\partial y_m^t}{\partial w_{km}^{(2)}}$$

$$y_m^t = \frac{e^{o_m^{(3)t}}}{\sum_{n}^{N} e^{o_n^{(3)}}}$$

$$y_m = \text{softmax}(o_m^{(3)})$$

$$\text{softmax}(i) = \frac{e^i}{\sum_{n}^{N} e^n}$$

$$\frac{d\,\text{softmax}(i)}{d\,i} = \frac{1}{\sum_{n}^{N} e^n} \cdot e^i - \frac{1}{\left(\sum_{n}^{N} e^n\right)^2} \cdot e^i \cdot e^i$$

$$= \frac{e^i \sum_{n}^{N} e^n - e^{2i}}{\left(\sum_{n}^{N} e^n\right)^2}$$

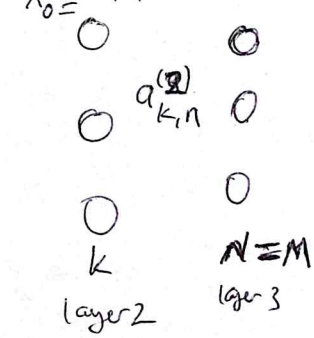
$$= \frac{e^i \cdot \left(\sum_{n}^{N} e^n - e^i\right)}{\left(\sum_{n}^{N} e^n\right)^2}$$

$$= \frac{e^i}{\left(\sum_{n}^{N} e^n\right)} \cdot \left(\frac{\sum_{n}^{N} e^n}{\sum_{n}^{N} e^n} - \frac{e^i}{\sum_{n}^{N} e^n}\right) = \boxed{\text{softmax}(i) \cdot (1 - \text{softmax}(i))}$$

*) $\frac{d\,softmax(i)}{dj} = -\frac{e^j}{(\sum_n^N e^n)^2}\cdot e^i = -\frac{e^j}{\sum_n^N e^n}\cdot\frac{e^i}{\sum_n^N e^n} = -softmax(i)\cdot softmax(j)$

$$\frac{d\,softmax(i)}{dj} = \begin{cases} softmax(i)\cdot(1-softmax(j)) & \text{if } i=j \\ -softmax(i)\cdot softmax(j) & \end{cases}$$

$E=j = -\sum_t\sum_m r_m^t \cdot \log(y_m^t)$ $\quad \frac{\partial E}{\partial a_{k,n}^{(2)}} = ?$




$y_m^t = softmax(O_m^{(3)t})$

$O_m^{(3)} = \sum_{k=0}^{(2)} a_{k,m}^{(2)} \cdot x_k^{(2)t}$

$\frac{\partial E}{\partial a_{k,n}^{(2)}} = -\sum_t\sum_m r_m^t \frac{\partial \log(y_m^t)}{\partial a_{k,n}^{(2)}}$

$\frac{\partial \log(y_m^t)}{\partial a_{k,n}^{(2)}} = \frac{\partial \log y_m^t}{\partial y_m^t}\cdot\frac{\partial y_m^t}{\partial O_n^{(2)t}}\cdot\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$

$= \frac{1}{y_m^t}\cdot y_m^t(1\{m=n\}-y_n^t)\cdot\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$

→ $y_m^t$ is function of $O_n^{(2)t}$ at the same time. $y_m^t = softmax(O_m^{(3)t}) = \frac{e^{O_m^{(3)t}}}{\sum_n^N e^{O_n^{(3)t}}}$

n represents the output layer.

$\frac{\partial E}{\partial a_{k,n}^{(2)}} = -\sum_t\sum_m r_m^t \frac{1}{y_m^t}\cdot y_m^t(1\{m=n\}-y_n^t)\cdot\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$ $\quad \frac{\partial O_{x\neq n}^{(3)}}{\partial a_{k,n}^{(2)}} = 0$ because $a_{k,n}^{(2)}$ effects only $\partial O_n^{(2)t}$.

$= -\sum_t\sum_m r_m^t(1\{m=n\}-y_n^t)\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$

→ $1\{m=n\}$ is 1 if $m=n$, otherwise 0

$= -\sum_t r_{m=n}^t(1-y_{m=n}^t)\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}} + \sum_{m\neq n} r_m^t \cdot -y_n^t \cdot \frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$

$= -\sum_t r_{m=n}^t\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}} - r_{m=n}^t y_{m=n}^t\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}} + -y_n^t\sum_{m\neq n} r_m^t\cdot\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$

$= -\sum_t r_{m=n}^t\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}} - y_n^t\cdot\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}\cdot\underbrace{(+r_m^t + \sum_{m\neq n} r_m^t)}_{1}$ → $r^t=[0\ 0\ 0\ 1] = sum=1$ or $[0\ 1\ 0\ 0] = sum=1$

$$= -\sum_t (r_n^t - y_n^t)\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$$

$\Delta a_{k,n}^{(2)} = -\eta\frac{\partial E}{\partial a_{k,n}^{(2)}} = \eta\sum_t(r_n^t - y_n^t)\frac{\partial O_n^{(2)t}}{\partial a_{k,n}^{(2)}}$

In regression $E=j = \sum_t\sum_m (r_m^t - O_m^{(3)t})^2$ → $O_m^{(3)t} = \sum_K a_{k,m}^{(2)} x_m^{(2)t}$

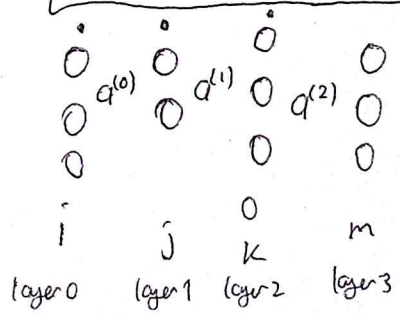
$\Delta a_{k,n}^{(2)} = -\eta\frac{\partial E}{\partial a_{k,n}^{(2)}}$ $\quad \frac{\partial E}{\partial a_{k,n}^{(2)}} = \sum_t\sum_m (r_m^t - O_m^{(3)t})\cdot -1\cdot\frac{\partial O_m^{(3)t}}{\partial a_{k,n}^{(2)}}$

$$\Delta a_{k,n}^{(2)} = +\eta\sum_t\sum_m (r_m^t - O_m^{(3)t})\cdot +1\cdot\frac{\partial O_m^{(3)t}}{\partial a_{k,n}^{(2)}}$$

in regression and classification loss functions' derivations of back propagation algorithm are same.

13.

## Implementation Considerations




In forward pass.

$$\begin{bmatrix} +1 \\ +1 \end{bmatrix} \begin{bmatrix} X_0^{(0)0} & X_2^{(0)0} & X_3^{(0)0} \\ X_1^{(0)1} & X_2^{(0)1} & X_3^{(0)1} \end{bmatrix} \cdot \begin{bmatrix} a_{0,1}^{(0)} & a_{0,2}^{(0)} \\ a_{1,1}^{(0)} & a_{1,2}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} \end{bmatrix}$$

this part should be inserted (column) and they correspond to bias value multiplier. (+1)

the result of this multiplication

$$\begin{matrix} +1 \\ +1 \end{matrix} \begin{bmatrix} O_1^{(1)0} & O_2^{(1)0} \\ O_1^{(1)1} & O_2^{(1)1} \end{bmatrix} \cdot \begin{bmatrix} a_{0,1}^{(1)} & a_{0,2}^{(1)} & a_{0,3}^{(1)} & a_{0,4}^{(1)} \\ a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & a_{1,4}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & a_{2,3}^{(1)} & a_{2,4}^{(1)} \end{bmatrix}$$

for every weight and input multiplication we are applying a linear transformation

for this multiplication to be refered this column vector should be added. And these are multiplier of biases (+1).

this results in

$$\begin{matrix} +1 \\ +1 \end{matrix} \begin{bmatrix} O_1^{(2)0} & O_2^{(2)0} & O_3^{(2)0} & O_4^{(2)0} \\ O_1^{(2)1} & O_2^{(2)1} & O_3^{(2)1} & O_4^{(2)1} \end{bmatrix} \cdot$$

to perform this operation we must add a column vector

$$\begin{bmatrix} a_{0,0}^{(2)} & a_{0,1}^{(2)} & a_{0,2}^{(2)} \\ a_{1,0}^{(2)} & a_{1,1}^{(2)} & a_{1,2}^{(2)} \\ a_{2,0}^{(2)} & a_{2,1}^{(2)} & a_{2,2}^{(2)} \\ a_{3,0}^{(2)} & a_{3,1}^{(2)} & a_{3,2}^{(2)} \\ a_{4,0}^{(2)} & a_{4,1}^{(2)} & a_{4,2}^{(2)} \end{bmatrix}$$

$$X = \begin{bmatrix} +1 & X_1 & X_2 & X_3 \\ X_0^{(0)1} & X_1^{(0)1} & X_2^{(0)1} & X_3^{(0)1} \end{bmatrix}$$

↓ input form

$$O_j^{(1)t} = \sum_{i=0} a_{i,j}^{(0)} X_i^{(0)t}$$

$$O^{(1)} = X_{\text{input}}^{(0)} \cdot a^{(0)}$$

$$a^{(0)} = \begin{bmatrix} a_{0,1}^{(0)} & a_{0,2}^{(0)} \\ a_{1,1}^{(0)} & a_{1,2}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} \end{bmatrix}$$

$$X_j^{(1)t} = \sigma(O_j^{(1)t})$$

$$\sigma(O^{(1)}) = X^{(1)}$$

$$O_k^{(2)t} = \sum_{j=0} a_{j,k}^{(1)} X_j^{(1)t}$$

$$O^{(2)} = X^{(1)} \cdot a^{(1)}$$

$$a^{(1)} = \begin{bmatrix} a_{0,1}^{(1)} & a_{0,2}^{(1)} & a_{0,3}^{(1)} & a_{0,4}^{(1)} & a_{0,5}^{(1)} \\ a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & a_{1,4}^{(1)} & a_{1,5}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & a_{2,3}^{(1)} & a_{2,4}^{(1)} & a_{2,5}^{(1)} \end{bmatrix}$$

$$X^{(2)} = \sigma(O^{(2)})$$

$$O_m^{(3)t} = \sum_{k=0} a_{k,m}^{(2)} X_k^{(2)t}$$

$$\text{Output} = O^{(3)} = X^{(2)} \cdot a^{(2)}$$ for regression

$$a^{(2)} = \begin{bmatrix} a_{0,0}^{(2)} & a_{0,1}^{(2)} & a_{0,2}^{(2)} \\ a_{1,0}^{(2)} & a_{1,1}^{(2)} & a_{1,2}^{(2)} \\ a_{2,0}^{(2)} & a_{2,1}^{(2)} & a_{2,2}^{(2)} \\ a_{3,0}^{(2)} & a_{3,1}^{(2)} & a_{3,2}^{(2)} \\ a_{4,0}^{(2)} & a_{4,1}^{(2)} & a_{4,2}^{(2)} \end{bmatrix}$$

In classification

$$\text{Output} = \text{softmax}(O^{(3)})$$

So to get rid of computation overhead of column insertion, we can seperate each weight matrice's first row. Because they represent the bias of weights.

$$a^{(0)} = \begin{bmatrix} a_{0,1}^{(0)} & a_{0,2}^{(0)} \\ a_{1,1}^{(0)} & a_{1,2}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} \end{bmatrix} \rightarrow \begin{bmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} \end{bmatrix} \text{ and } \begin{bmatrix} a_{0,1}^{(0)} & a_{0,1}^{(0)} \end{bmatrix}$$

new $a^{(0)}$ ; $\downarrow$ $b^{(0)}$

4.) $q^{(1)} = \begin{bmatrix} a_{0,1}^{(1)} & a_{0,2}^{(1)} & a_{0,3}^{(1)} & a_{0,4}^{(1)} & a_{0,5}^{(1)} \\ a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & a_{1,4}^{(1)} & a_{1,5}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & a_{2,3}^{(1)} & a_{2,4}^{(1)} & a_{2,5}^{(1)} \end{bmatrix} \rightarrow \begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & a_{1,4}^{(1)} & a_{1,5}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & a_{2,3}^{(1)} & a_{2,4}^{(1)} & a_{2,5}^{(1)} \end{bmatrix}$ and $\begin{bmatrix} a_{0,1}^{(1)} & a_{0,2}^{(1)} & a_{0,3}^{(1)} & a_{0,4}^{(1)} & a_{0,5}^{(1)} \end{bmatrix}$

new $a^{(1)}$ ; $b^{(1)}$

$a^{(2)} = \begin{bmatrix} a_{0,0}^{(2)} & a_{0,1}^{(2)} & a_{0,2}^{(2)} \\ a_{1,0}^{(2)} & a_{1,1}^{(2)} & a_{1,2}^{(2)} \\ a_{2,0}^{(2)} & a_{2,1}^{(2)} & a_{2,2}^{(2)} \\ a_{3,0}^{(2)} & a_{3,1}^{(2)} & a_{3,2}^{(2)} \\ a_{4,0}^{(2)} & a_{4,1}^{(2)} & a_{4,2}^{(2)} \end{bmatrix} \rightarrow \begin{bmatrix} a_{1,0}^{(2)} & a_{1,1}^{(2)} & a_{1,2}^{(2)} \\ a_{2,0}^{(2)} & a_{2,1}^{(2)} & a_{2,2}^{(2)} \\ a_{3,0}^{(2)} & a_{3,1}^{(2)} & a_{3,2}^{(2)} \\ a_{4,0}^{(2)} & a_{4,1}^{(2)} & a_{4,2}^{(2)} \end{bmatrix}$ and $\begin{bmatrix} a_{0,0}^{(2)} & a_{0,1}^{(2)} & a_{0,2}^{(2)} \end{bmatrix}$

new $a^{(2)}$ ; $b^{(2)}$

So without column addition.

original input $\begin{bmatrix} X_1^{(0)0} & X_2^{(0)0} & X_3^{(0)0} \\ X_1^{(0)1} & X_2^{(0)1} & X_3^{(0)1} \end{bmatrix} \cdot \begin{bmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} \end{bmatrix}$

$\begin{bmatrix} X_1^{(0)0} \cdot a_{1,1}^{(0)} + X_2^{(0)0} \cdot a_{2,1}^{(0)} + X_3^{(0)0} \cdot a_{3,1}^{(0)} & X_1^{(0)0} \cdot a_{1,2}^{(0)} + X_2^{(0)0} \cdot a_{2,2}^{(0)} + X_3^{(0)0} \cdot a_{3,2}^{(0)} \\ X_1^{(0)1} a_{1,1}^{(0)} + X_2^{(0)1} \cdot a_{2,1}^{(0)} + X_3^{(0)1} \cdot a_{3,1}^{(0)} & X_1^{(0)1} \cdot a_{1,2}^{(0)} + X_2^{(0)1} a_{2,2}^{(0)} + X_3^{(0)1} \cdot a_{3,2}^{(0)} \end{bmatrix}$ → this operation lacks the bias addition.

layer 1 1th node ; layer 1 2nd node

$\begin{bmatrix} a_{0,1}^{(0)} & a_{0,2}^{(0)} \end{bmatrix}$

↓ first node bias ; ↓ second node bias.

$\begin{bmatrix} X_0^{(0)0} \cdot a_{0,1}^{(0)} & X_0^{(0)0} \cdot a_{0,2}^{(0)} \\ X_0^{(0)1} a_{0,1}^{(0)} & X_0^{(0)1} \cdot a_{0,2}^{(0)} \end{bmatrix}$ → this matrix should be added to prev calculation.

if can be represented $\begin{bmatrix} +1 \\ +1 \end{bmatrix} \cdot \begin{bmatrix} a_{0,1}^{(0)} & a_{0,2}^{(0)} \end{bmatrix}$

$X^{(0)}$ → are bias weights.

→ this matrix's size must be equal to input matrix's row number. If there are 5 samples it must be (5,1)

if 10 → (10,1) → $\begin{bmatrix} +1 \\ +1 \\ +1 \\ +1 \\ +1 \end{bmatrix}$

So for all weights it can be shown that → seperated matrix

$O^{(1)} = X^{(0)} \cdot a^{(0)} + \begin{bmatrix} +1 \\ +1 \\ \vdots \end{bmatrix} \cdot b^{(0)}$

$X^{(1)} = \sigma(O^{(1)})$

$O^{(2)} = X^{(1)} \cdot a^{(1)} + \begin{bmatrix} +1 \\ +1 \\ \vdots \end{bmatrix} \cdot b^{(1)}$

$X^{(2)} = \sigma(O^{(2)})$

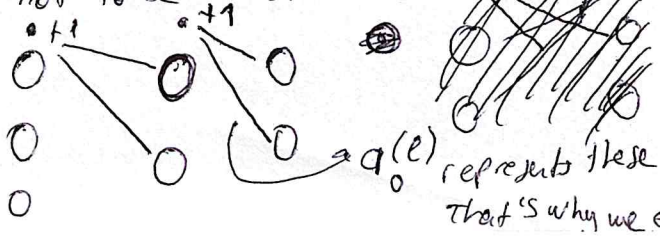
$O^{(3)} = X^{(2)} \cdot a^{(2)} + \begin{bmatrix} +1 \\ +1 \\ \vdots \end{bmatrix} \cdot b^{(2)}$ → for regression it is output

for classification softmax $(O^{(3)})$. ← output

## Back Propagation

Previous backpropagation derivation is based on +1 column vector addition.

In this configuration, we have to cut $a_0^{(e)}$ → row vector because they represent the bias weights. while transferring the error to previous layer these weights need not to be used.



while updating the weights bias weights are updated also in the calculation.

$$\Delta a^{(2)} = \eta\, X^{(2)T} \cdot Err$$

for ex. $Err = \begin{bmatrix} r_0^0 & r_1^0 & r_2^0 \\ r_0^1 & r_1^1 & r_2^1 \end{bmatrix}$

$$X^{(2)} = \begin{bmatrix} +1 & x_1^{(2)0} & x_2^{(2)0} & x_3^{(2)0} & x_4^{(2)0} \\ +1 & x_1^{(2)1} & x_2^{(2)1} & x_3^{(2)1} & x_4^{(2)1} \end{bmatrix}$$

$$\Delta a^{(2)} = \eta \begin{bmatrix} +1 & +1 \\ x_1^{(2)0} & x_1^{(2)1} \\ x_2^{(2)0} & x_2^{(2)1} \\ x_3^{(2)0} & x_3^{(2)1} \\ x_4^{(2)0} & x_4^{(2)1} \end{bmatrix} \cdot \begin{bmatrix} r_0^0 & r_1^0 & r_2^0 \\ r_0^1 & r_1^1 & r_2^1 \end{bmatrix} = \begin{bmatrix} \Delta a_{0,0}^{(2)} & \Delta a_{0,1}^{(2)} & \Delta a_{0,2}^{(2)} \\ \Delta a_{1,0}^{(2)} & \Delta a_{1,1}^{(2)} & \Delta a_{1,2}^{(2)} \\ \Delta a_{2,0}^{(2)} & \Delta a_{2,1}^{(2)} & \Delta a_{2,2}^{(2)} \\ \Delta a_{3,0}^{(2)} & \Delta a_{3,1}^{(2)} & \Delta a_{3,2}^{(2)} \\ \Delta a_{4,0}^{(2)} & \Delta a_{4,1}^{(2)} & \Delta a_{4,2}^{(2)} \end{bmatrix}$$

so to update $b^{(0)}$ we can get it from above

$$\eta \begin{bmatrix} +1 & +1 \end{bmatrix} \cdot \begin{bmatrix} r_0^0 & r_1^0 & r_2^0 \\ r_0^1 & r_1^1 & r_2^1 \end{bmatrix} \rightarrow \begin{bmatrix} \Delta a_{0,0}^{(2)} & \Delta a_{0,1}^{(2)} & \Delta a_{0,2}^{(2)} \end{bmatrix}$$

So if we seperate $a^{(2)}$ and don't use colum addition we will have

$$Err = \begin{bmatrix} r_0^0 & r_1^0 & r_2^0 \\ r_0^1 & r_1^1 & r_2^1 \end{bmatrix}$$

$$X^{(2)} = \begin{bmatrix} x_1^{(2)0} & x_2^{(2)0} & x_3^{(2)0} & x_4^{(2)0} \\ x_1^{(2)1} & x_2^{(2)1} & x_3^{(2)1} & x_4^{(2)1} \end{bmatrix}$$

$$= \eta \cdot \begin{bmatrix} x_1^{(2)0} & x_1^{(2)1} \\ x_2^{(2)0} & x_2^{(2)1} \\ x_3^{(2)0} & x_3^{(2)1} \\ x_4^{(2)0} & x_4^{(2)1} \end{bmatrix} \cdot \begin{bmatrix} r_0^0 & r_1^0 & r_2^0 \\ r_0^1 & r_1^1 & r_2^1 \end{bmatrix} \rightarrow \begin{bmatrix} \Delta a_{1,0}^{(2)} & \Delta a_{1,1}^{(2)} & \Delta a_{1,2}^{(2)} \\ \Delta a_{2,0}^{(2)} & \Delta a_{2,1}^{(2)} & \Delta a_{2,2}^{(2)} \\ \Delta a_{3,0}^{(2)} & \Delta a_{3,1}^{(2)} & \Delta a_{3,2}^{(2)} \\ \Delta a_{4,0}^{(2)} & \Delta a_{4,1}^{(2)} & \Delta a_{4,2}^{(2)} \end{bmatrix}$$

So similar to what we have done for forward backpropagation, while updating biases we need to use a row vector of ones.

$$\Delta a^{(1)} = \eta\, X^{(1)T} \cdot \text{layer 2 error}$$

previously $X^{(1)T} \rightarrow \eta \cdot \begin{bmatrix} +1 & +1 \\ x_1^{(1)0} & x_1^{(1)1} \\ x_2^{(1)0} & x_2^{(1)1} \end{bmatrix} \cdot \begin{bmatrix} r_0^0 & r_1^0 & r_2^0 & r_3^0 & r_4^0 \\ r_0^1 & r_1^1 & r_2^1 & r_3^1 & r_4^1 \end{bmatrix} \rightarrow \begin{bmatrix} \Delta a_{0,1}^{(1)} & \Delta a_{0,2}^{(1)} & \Delta a_{0,3}^{(1)} & \Delta a_{0,4}^{(1)} \\ \Delta a_{1,1}^{(1)} & \Delta a_{1,2}^{(1)} & \Delta a_{1,3}^{(1)} & \Delta a_{1,4}^{(1)} \\ \Delta a_{2,1}^{(1)} & \Delta a_{2,2}^{(1)} & \Delta a_{2,3}^{(1)} & \Delta a_{2,4}^{(1)} \end{bmatrix}$

in new configuration $X^{(1)T} \rightarrow \begin{bmatrix} x_1^{(1)0} & x_1^{(1)1} \\ x_2^{(1)0} & x_2^{(1)1} \end{bmatrix} \cdot \begin{bmatrix} r_1^0 & r_2^0 & r_3^0 & r_4^0 \\ r_1^1 & r_2^1 & r_3^1 & r_4^1 \end{bmatrix} \rightarrow \begin{bmatrix} \Delta a_{1,1}^{(1)} & \Delta a_{1,2}^{(1)} & \Delta a_{1,3}^{(1)} & \Delta a_{1,4}^{(1)} \\ a_{2,1}^{(1)} & \Delta a_{2,2}^{(1)} & \Delta a_{2,3}^{(1)} & \Delta a_{2,4}^{(1)} \end{bmatrix}$

$\Delta a^{(1)}$ so for $\Delta b^{(1)}$ $\eta \cdot [+1 \; +1] \cdot$ layer 2 error

So for any bias weight, while updating bias these weights should be multiplied with a row of one. Its size should be as input matrix row count.

$\Delta a^{(2)} = \eta\, X^{(2)T} \cdot$ Output Error $\quad \Delta b^{(2)} = \eta [+1 \; +1 \; \cdots] \cdot$ output Error

$\Delta a^{(1)} = \eta \cdot X^{(1)T} \cdot$ Layer 2 Error $\quad \Delta b^{(1)} = \eta [+1 \; +1 \; \cdots] \cdot$ Layer 2 Error

$\Delta a^{(0)} = \eta \cdot X^{(0)T} \cdot$ Layer 1 Error $\quad \Delta b^{(0)} = \eta [+1 \; +1 \; \cdots] \cdot$ Layer 1 Error

matrix multiplication

$$\Delta a^{(e)} = \eta\, X^{(e)T} \cdot Err^{(e+1)} \qquad \Delta b^{(e)} = \eta [+1 \; +1 \; \cdots] \cdot Err^{(e+1)}$$