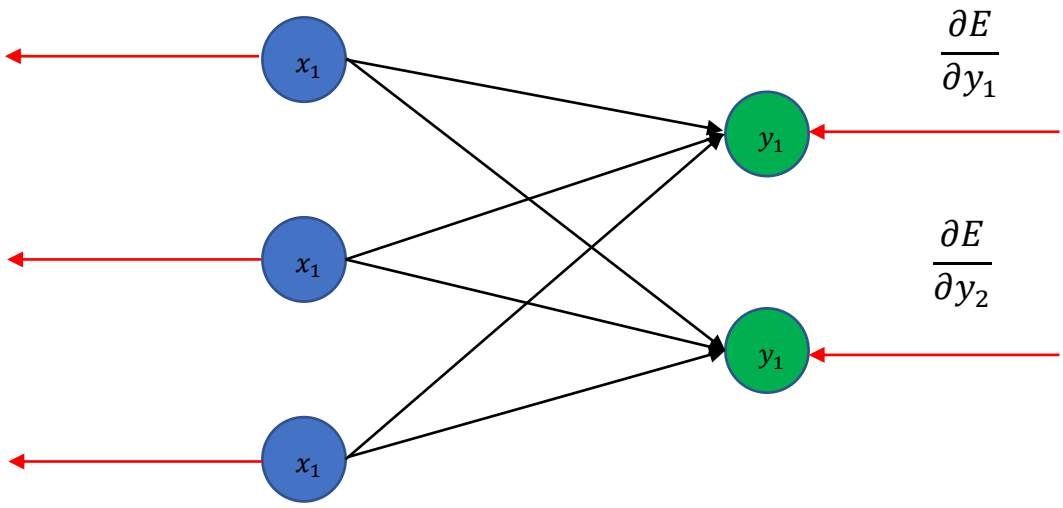
深度学习的反向传播

这篇文章主要讲解一下深度学习的方向传播理论。这个也是深度学习最重要的理论基础，同时也是深度学习中比较比较复杂的理论，需要对偏导数有较深的理解。很多人面对这个感觉无从下手，希望当看完我这篇文章后，会让大家能更容易地理解方向传播。本篇文章主要从数学的角度入手，会对反向传播中，对误差的求导，对权重 W 的求导，对偏移量 b 的求导这些求导公式做出详细的数学证明。其实这些求导公式都是基于矩阵的乘法和转置运算，乘法包括矩阵的点乘和按位乘。很多人看到这些矩阵相乘就能求出对应的导数干到很神奇，其实这些矩阵的乘法运算都是基于链式求导法则得到的，只是通过向量化，然后以矩阵相乘的形式给出，其实本质是链式求导法得到的结果。本文会给出详细的证明过程，甚至证明过程会过于啰嗦，为的是让大家很容易理解。好了言归正传，我们开始吧。

我们首先要规定一下数学符号的表达方式。这样方便我们更加容易的表达后面的各种数学证明。我们用$X^{\{1\}}$，$X^{\{2\}}\cdots\cdots X^{m}$来表达输入数据上标用大括号表示的是第几个样本，$m$表示的是样本数量大小。所以$X^{\{1\}}$就表示第一个样本，$X^{\{2\}}$表示第二个样本，以此类推。$X^{\{1\}}$是个向量，我们可以这样表达，$X^{\{1\}}=\left[x_1^{\{1\}},x_2^{\{1\}},x_3^{\{1\}}\cdots x_n^{\{1\}}\right]$,这个表达的是$X^{\{1\}}$是个 nx1 向量，它有$n$个维度。大括号表示的是第几个样本。我们用$W_{m\times n}^{[1]}=\begin{pmatrix} w_{11}^{[1]} & w_{12}^{[1]} & \cdots & w_{1n}^{[1]} \\ w_{21}^{[1]} & w_{22}^{[1]} & \cdots & w_{2n}^{[1]} \\ \vdots & \vdots & \cdots & \vdots \\ w_{m1}^{[1]} & w_{m2}^{[1]} & \cdots & w_{mn}^{[1]} \end{pmatrix}$来表示深度网络模型中的任意一层。中括号如[1],[2]等里面的数字表达的就是网络的第几层。$W_{m\times n}^{[1]}$表示的意思就是第一层权重，即是连接输入层和第一个隐藏层之间的权重矩阵，shape 是$m\times n$。$w_{11}^{[1]}$表达的$W_{m\times n}^{[1]}$的元素，下边的 11 表达式的是在$W_{m\times n}^{[1]}$的位置。这里是 2 维的，所以是行列的位置。如果不涉及层数的问题，我们就不需要使用中括号来表达层次关系。我们用向量$b_{n\times 1}^{[1]}=\begin{pmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \\ \vdots \\ b_n^{[1]} \end{pmatrix}$来表示偏移向量。下标 1,2,3,n 表达的是第几个个维度。我们用$\hat{y}$表示输出的预测值向量，y^{true}表达真实值向量。$\hat{y}$和y^{true}里面具体元素的表达方式和前面介绍的权重输入矩阵等类似。利用大括号，中括号，上标和下标的数字来表达各种含义。后面我们给出具体的表达方式，还会再次说明符号的含义，大家现在记不住也不要担心，这里是给大家一个总体概念上的认识。好了，我们数学符号的表达说完了，我们接着开始推导了。

我先给出个例子，会通过一个神经网络中的某一层来具体说明每个反向传播的公式是如何推导出的。我给出的是某个神经网络的某一层的反向传播截图。因为每一层的求导过程都是一样的，所以知道其中一层如何求导，其余的层就按部就班的照做，然后根据链式法则连乘就可以。我先把这个网络的结构展示出来，请看下图：



神经网络某一层的反向传播

单个样本情况下的反向传播:

我们具体解释一下这个图。首先我们先以单个样本为例子，所以x_1、x_2、x_3是输入样本数据$X_{3\times1}$的各个维度的元素。为了方便描述，我设置输入样本的维度是 3。设置的权重矩阵$W_{2\times3}$。该层的输出是$y_{2\times1} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$，这里输出的维度设置大小为 2。偏移量向量是$b_{2\times1} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$。这些维度设置的值都是为了方便描述和数学表达式的书写。那么这一层的前向传递的计算结果是如下表达式：

$$y_{2\times1} = W_{2\times3} \cdot X_{3\times1} + b_{2\times1}$$

$$\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \boldsymbol{b_1} \\ \boldsymbol{b_2} \end{pmatrix}$$

$$\downarrow$$

$$y_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1$$

$$y_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2$$

上面的这些计算表达式明确给出了$X_{3\times1}$, 权重$W_{2\times3}$, 偏移量$b_{2\times1}$和输出$y_{2\times1}$之间的关系。我们将反向传播传递过来的误差设置为$\frac{\partial E}{\partial y} = \begin{pmatrix} \frac{\partial E}{\partial y_1} \\ \frac{\partial E}{\partial y_2} \end{pmatrix}$，$E$是损失函数。根据链式求导法则传递到$y_{2\times1}$的误差就是$\frac{\partial E}{\partial y}$。下面我们就要计算当前这一层的相关导数。即是$\frac{\partial E}{\partial X}$、$\frac{\partial E}{\partial W}$、以及$\frac{\partial E}{\partial b}$。我们给出计算这三个导数的公式，同时给出相应的数学证明.

$$\frac{\partial E}{\partial X} = W^T \cdot \frac{\partial E}{\partial y} \qquad \frac{\partial E}{\partial W} = \frac{\partial E}{\partial y} \cdot X^T \qquad \frac{\partial E}{\partial b} = \frac{\partial E}{\partial y}$$

其中T是转置的意思。好了，公式我们已经给出了，那么下面我们就会证明这件事情了。我们首先证明最左边的公式。即$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial y} \cdot X^T$，通过链式法则证明如下：

$$\begin{aligned} \frac{\partial E}{\partial x_1} &= \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial x_1} + \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial x_1} \\ &= \frac{\partial E}{\partial y_1} \cdot w_{11} + \frac{\partial E}{\partial y_2} \cdot w_{21} \end{aligned}$$

$$\begin{aligned} \frac{\partial E}{\partial x_2} &= \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial x_2} + \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial x_2} \\ &= \frac{\partial E}{\partial y_1} \cdot w_{12} + \frac{\partial E}{\partial y_2} \cdot w_{22} \end{aligned}$$

$$\frac{\partial E}{\partial x_3} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial x_3} + \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial x_3}$$

$$= \frac{\partial E}{\partial y_1} \cdot w_{13} + \frac{\partial E}{\partial y_2} \cdot w_{23}$$

所以我们可以得到$\frac{\partial E}{\partial X}$的矩阵了。

$$\frac{\partial E}{\partial X} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \frac{\partial y}{\partial x_3} \end{pmatrix} = \begin{pmatrix} \frac{\partial E}{\partial y_1} \cdot w_{11} + \frac{\partial E}{\partial y_2} \cdot w_{21} \\ \frac{\partial E}{\partial y_1} \cdot w_{12} + \frac{\partial E}{\partial y_2} \cdot w_{22} \\ \frac{\partial E}{\partial y_1} \cdot w_{13} + \frac{\partial E}{\partial y_2} \cdot w_{23} \end{pmatrix}$$

$$= \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial E}{\partial y_1} \\ \frac{\partial E}{\partial y_2} \end{pmatrix}$$

$$= W^T \cdot \frac{\partial E}{\partial y}$$

证明完毕。对单个样本$\frac{\partial E}{\partial X} = W^T \cdot \frac{\partial E}{\partial y}$，我们接下来再来证明$\frac{\partial E}{\partial W}$.

$$\frac{\partial E}{\partial w_{11}} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial w_{11}} = \frac{\partial E}{\partial y_1} \cdot x_1$$

$$\frac{\partial E}{\partial w_{12}} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial w_{12}} = \frac{\partial E}{\partial y_1} \cdot x_2$$

$$\frac{\partial E}{\partial w_{13}} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial w_{13}} = \frac{\partial E}{\partial y_1} \cdot x_3$$

$$\frac{\partial E}{\partial w_{21}} = \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial w_{21}} = \frac{\partial E}{\partial y_2} \cdot x_1$$

$$\frac{\partial E}{\partial w_{22}} = \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial w_{22}} = \frac{\partial E}{\partial y_2} \cdot x_2$$

$$\frac{\partial E}{\partial w_{23}} = \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial w_{23}} = \frac{\partial E}{\partial y_2} \cdot x_3$$

因此矩阵$\frac{\partial E}{\partial W}$就可以得到了:

$$\frac{\partial E}{\partial W} = \begin{pmatrix} \frac{\partial E}{\partial w_{11}} & \frac{\partial E}{\partial w_{12}} & \frac{\partial E}{\partial w_{13}} \\ \frac{\partial E}{\partial w_{21}} & \frac{\partial E}{\partial w_{22}} & \frac{\partial E}{\partial w_{23}} \end{pmatrix} = \begin{pmatrix} \frac{\partial E}{\partial y_1} \cdot x_1 & \frac{\partial E}{\partial y_1} \cdot x_3 & \frac{\partial E}{\partial y_1} \cdot x_3 \\ \frac{\partial E}{\partial y_2} \cdot x_1 & \frac{\partial E}{\partial y_2} \cdot x_2 & \frac{\partial E}{\partial y_2} \cdot x_3 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial E}{\partial y_1} \\ \frac{\partial E}{\partial y_2} \end{pmatrix} \cdot (x_1 \quad x_2 \quad x_3) = \frac{\partial E}{\partial y} \cdot X^T$$

证明完毕。最后我们再把$\frac{\partial E}{\partial b}$的证明过程给出：

$$\frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial b_1} = \frac{\partial E}{\partial y_1} \cdot 1 = \frac{\partial E}{\partial y_1}$$

$$\frac{\partial E}{\partial b_2} = \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial b_2} = \frac{\partial E}{\partial y_2} \cdot 1 = \frac{\partial E}{\partial y_2}$$

$$\frac{\partial E}{\partial b} = \begin{pmatrix} \frac{\partial E}{\partial b_1} \\ \frac{\partial E}{\partial b_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial E}{\partial y_1} \\ \frac{\partial E}{\partial y_2} \end{pmatrix} = \frac{\partial E}{\partial y}$$

好了$\frac{\partial E}{\partial b}$也已经证明好了。可见我们给出的公式是没有问题的。其实根本原理就是复合函数求解导数的链式法则。深度学习的反向传播理论的基础也就是这个链式法则。每一层都针对一个链接。这样我们就对单个样本的情况下，反向传播计算公式的证明做出了充分的论证。下面我们在此基础上扩展一下，如果是批量样本如何处理呢。我们下面接着讨论。

多个个样本情况下的反向传播:

我们设置样本数量为$\boldsymbol{m}$个。那么相应的X，y则从原来的一个变成m个了。W和$\boldsymbol{b}$不变，依然保持我们之前设置的形状: $W_{2\times3}=\begin{pmatrix}w_{11} & w_{12} & w_{13}\\ w_{21} & w_{22} & w_{23}\end{pmatrix}\quad b=\begin{pmatrix}b_1\\ b_2\end{pmatrix}$

$$y_{2\times m}=W_{2\times3}\cdot X_{3\times m}+b$$

$$\begin{pmatrix}y_1^{\{1\}} & y_1^{\{2\}} & y_1^{\{3\}} & \cdots & y_1^{\{m\}}\\ y_2^{\{1\}} & y_2^{\{2\}} & y_2^{\{3\}} & \cdots & y_2^{\{m\}}\end{pmatrix}=\begin{pmatrix}b_1\\ b_2\end{pmatrix}+\begin{pmatrix}w_{11} & w_{12} & w_{13}\\ w_{21} & w_{22} & w_{23}\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{x_1^{\{1\}}} & \boldsymbol{x_1^{\{2\}}} & \boldsymbol{x_1^{\{3\}}} & \cdots & \boldsymbol{x_1^{\{m\}}}\\ \boldsymbol{x_2^{\{1\}}} & \boldsymbol{x_2^{\{2\}}} & \boldsymbol{x_2^{\{3\}}} & \cdots & \boldsymbol{x_2^{\{m\}}}\\ \boldsymbol{x_3^{\{1\}}} & \boldsymbol{x_3^{\{2\}}} & \boldsymbol{x_3^{\{3\}}} & \cdots & \boldsymbol{x_3^{\{m\}}}\end{pmatrix}$$

以上的给出的式子就是m个样本的时候得到的计算关系。还记得前面说的{}符号，表示第几个样本。和前面单样本相比，仅仅是增加了多个样本。计算的时候把批量考虑进去就好。我们给出m个样本的时候反向传播的计算公式：

$$\frac{\partial E}{\partial x}=w^T\cdot\frac{\partial E}{\partial y}\quad(1)$$

$$\frac{\partial E}{\partial W}=\frac{\partial E}{\partial y}\cdot X^T/m\quad(2)$$

$$\frac{\partial E}{\partial b}=sum\left(\frac{\partial E}{\partial y},axis=1\right)/m\quad(3)$$

我们给出了m个样本的反向传播的计算公式，我们发现当$m=1$的时候，这些公式就退化为单样本的计算公式了。$sum\left(\frac{\partial E}{\partial y}, axis=1\right)$的含义是对误差矩阵$\frac{\partial E}{\partial y}$的第二个轴，也就是列方向求和。如果$m=1$的时候，$\frac{\partial E}{\partial y}$就只有一列。所以矩阵$sum\left(\frac{\partial E}{\partial y}, axis=1\right)$就等于$\frac{\partial E}{\partial y}$。综上$m=1$的时候，给出的多个样本方向传播计算公式就转变为单样本计算公式。所以我们以后就使用多个样本的反向传播计算公式就可以了，因为它考虑了一个和多个两种情况了，我们可以不需要之前的单样本公式了。好了，我们下面开始推导这些公式吧。还是和之前推导单样本的时候一样，我们按照给出的顺序逐一推导。

$$\frac{\partial E}{\partial x_1^{\{1\}}}=\frac{\partial E}{\partial y_1^{\{1\}}}\cdot\frac{\partial y_1^{\{1\}}}{\partial x_1^{\{1\}}}+\frac{\partial E}{\partial y_2^{\{1\}}}\cdot\frac{\partial y_2^{\{1\}}}{\partial x_1^{\{1\}}}=\frac{\partial E}{\partial y_1^{\{1\}}}\cdot w_{11}+\frac{\partial E}{\partial y_2^{\{1\}}}\cdot w_{21}$$

$$\frac{\partial E}{\partial x_2^{\{1\}}}=\frac{\partial E}{\partial y_1^{\{1\}}}\cdot\frac{\partial y_1^{\{1\}}}{\partial x_2^{\{1\}}}+\frac{\partial E}{\partial y_2^{\{1\}}}\cdot\frac{\partial y_2^{\{1\}}}{\partial x_2^{\{1\}}}=\frac{\partial E}{\partial y_1^{\{1\}}}\cdot w_{12}+\frac{\partial E}{\partial y_2^{\{1\}}}\cdot w_{22}$$

$$\frac{\partial E}{\partial x_3^{\{1\}}}=\frac{\partial E}{\partial y_1^{\{1\}}}\cdot\frac{\partial y_1^{\{1\}}}{\partial x_3^{\{1\}}}+\frac{\partial E}{\partial y_2^{\{1\}}}\cdot\frac{\partial y_2^{\{1\}}}{\partial x_3^{\{1\}}}=\frac{\partial E}{\partial y_1^{\{1\}}}\cdot w_{13}+\frac{\partial E}{\partial y_2^{\{1\}}}\cdot w_{23}$$

$$\frac{\partial E}{\partial x_1^{\{2\}}}=\frac{\partial E}{\partial y_1^{\{2\}}}\cdot\frac{\partial y_1^{\{2\}}}{\partial x_1^{\{2\}}}+\frac{\partial E}{\partial y_2^{\{2\}}}\cdot\frac{\partial y_2^{\{2\}}}{\partial x_1^{\{2\}}}=\frac{\partial E}{\partial y_1^{\{2\}}}\cdot w_{11}+\frac{\partial E}{\partial y_2^{\{2\}}}\cdot w_{21}$$

$$\frac{\partial E}{\partial x_2^{\{2\}}}=\frac{\partial E}{\partial y_1^{\{2\}}}\cdot\frac{\partial y_1^{\{2\}}}{\partial x_2^{\{2\}}}+\frac{\partial E}{\partial y_2^{\{2\}}}\cdot\frac{\partial y_2^{\{2\}}}{\partial x_2^{\{2\}}}=\frac{\partial E}{\partial y_1^{\{2\}}}\cdot w_{12}+\frac{\partial E}{\partial y_2^{\{2\}}}\cdot w_{22}$$

$$\frac{\partial E}{\partial x_3^{\{2\}}}=\frac{\partial E}{\partial y_1^{\{2\}}}\cdot\frac{\partial y_1^{\{2\}}}{\partial x_3^{\{2\}}}+\frac{\partial E}{\partial y_2^{\{2\}}}\cdot\frac{\partial y_2^{\{2\}}}{\partial x_3^{\{2\}}}=\frac{\partial E}{\partial y_1^{\{2\}}}\cdot w_{13}+\frac{\partial E}{\partial y_2^{\{2\}}}\cdot w_{23}$$

规律可以找到了，我们可以推广到$\frac{\partial E}{\partial x^{\{m\}}}$

$$\frac{\partial E}{\partial x_1^{\{m\}}} = \frac{\partial E}{\partial y_1^{\{m\}}} \cdot \frac{\partial y_1^{\{m\}}}{\partial x_1^{\{m\}}} + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot \frac{\partial y_2^{\{m\}}}{\partial x_1^{\{m\}}} = \frac{\partial E}{\partial y_1^{\{m\}}} \cdot w_{11} + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot w_{21}$$

$$\frac{\partial E}{\partial x_2^{\{m\}}} = \frac{\partial E}{\partial y_1^{\{m\}}} \cdot \frac{\partial y_1^{\{m\}}}{\partial x_2^{\{m\}}} + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot \frac{\partial y_2^{\{m\}}}{\partial x_2^{\{m\}}} = \frac{\partial E}{\partial y_1^{\{m\}}} \cdot w_{12} + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot w_{22}$$

$$\frac{\partial E}{\partial x_3^{\{m\}}} = \frac{\partial E}{\partial y_1^{\{m\}}} \cdot \frac{\partial y_1^{\{m\}}}{\partial x_3^{\{m\}}} + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot \frac{\partial y_2^{\{m\}}}{\partial x_3^{\{m\}}} = \frac{\partial E}{\partial y_1^{\{m\}}} \cdot w_{13} + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot w_{23}$$

那么我们就可以得到$\frac{\partial E}{\partial X}$的求导结果了。我们写成如下的矩阵形式:$\boldsymbol{\frac{\partial E}{\partial X}} =$

$$\begin{pmatrix} \frac{\partial E}{\partial y_1^{\{1\}}} w_{11} + \frac{\partial E}{\partial y_2^{\{1\}}} w_{21} & \frac{\partial E}{\partial y_1^{\{2\}}} w_{11} + \frac{\partial E}{\partial y_2^{\{2\}}} w_{21} & \cdots & \frac{\partial E}{\partial y_1^{\{m\}}} w_{11} + \frac{\partial E}{\partial y_2^{\{m\}}} w_{21} \\ \frac{\partial E}{\partial y_1^{\{1\}}} w_{12} + \frac{\partial E}{\partial y_2^{\{1\}}} w_{22} & \frac{\partial E}{\partial y_1^{\{2\}}} w_{12} + \frac{\partial E}{\partial y_2^{\{2\}}} w_{22} & \cdots & \frac{\partial E}{\partial y_1^{\{m\}}} w_{12} + \frac{\partial E}{\partial y_2^{\{m\}}} w_{22} \\ \frac{\partial E}{\partial y_1^{\{1\}}} w_{13} + \frac{\partial E}{\partial y_2^{\{1\}}} w_{23} & \frac{\partial E}{\partial y_1^{\{2\}}} w_{13} + \frac{\partial E}{\partial y_2^{\{2\}}} w_{23} & \cdots & \frac{\partial E}{\partial y_1^{\{m\}}} w_{13} + \frac{\partial E}{\partial y_2^{\{m\}}} w_{23} \end{pmatrix}$$

$$= \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial E}{\partial y_1^{\{1\}}} & \frac{\partial E}{\partial y_1^{\{2\}}} & \frac{\partial E}{\partial y_1^{\{3\}}} & \cdots & \frac{\partial E}{\partial y_1^{\{m\}}} \\ \frac{\partial E}{\partial y_2^{\{1\}}} & \frac{\partial E}{\partial y_2^{\{2\}}} & \frac{\partial E}{\partial y_2^{\{3\}}} & \cdots & \frac{\partial E}{\partial y_2^{\{m\}}} \end{pmatrix}$$

$$= W^T \cdot \frac{\partial E}{\partial y}$$

那么第一个关于$\frac{\partial E}{\partial x}$的公式就证明完毕了。我们接着再证明下一个。

$$\frac{\partial E}{\partial w_{11}} = \frac{\partial E}{\partial y_1^{\{1\}}} \frac{\partial y_1^{\{1\}}}{\partial w_{11}} + \frac{\partial E}{\partial y_1^{\{2\}}} \frac{\partial y_1^{\{2\}}}{\partial w_{11}} + \frac{\partial E}{\partial y_1^{\{3\}}} \frac{\partial y_1^{\{3\}}}{\partial w_{11}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}} \frac{\partial y_1^{\{m\}}}{\partial w_{11}}$$

$$= \frac{\partial E}{\partial y_1^{\{1\}}} x_1^{\{1\}} + \frac{\partial E}{\partial y_1^{\{2\}}} x_1^{\{2\}} + \frac{\partial E}{\partial y_1^{\{3\}}} x_1^{\{3\}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}} x_1^{\{m\}}$$

$$\frac{\partial E}{\partial w_{12}} = \frac{\partial E}{\partial y_1^{\{1\}}}\frac{\partial y_1^{\{1\}}}{\partial w_{12}} + \frac{\partial E}{\partial y_1^{\{2\}}}\frac{\partial y_1^{\{2\}}}{\partial w_{12}} + \frac{\partial E}{\partial y_1^{\{3\}}}\frac{\partial y_1^{\{3\}}}{\partial w_{12}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}\frac{\partial y_1^{\{m\}}}{\partial w_{12}}$$
$$= \frac{\partial E}{\partial y_1^{\{1\}}}x_2^{\{1\}} + \frac{\partial E}{\partial y_1^{\{2\}}}x_2^{\{2\}} + \frac{\partial E}{\partial y_1^{\{3\}}}x_2^{\{3\}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}x_2^{\{m\}}$$

$$\frac{\partial E}{\partial w_{13}} = \frac{\partial E}{\partial y_1^{\{1\}}}\frac{\partial y_1^{\{1\}}}{\partial w_{13}} + \frac{\partial E}{\partial y_1^{\{2\}}}\frac{\partial y_1^{\{2\}}}{\partial w_{13}} + \frac{\partial E}{\partial y_1^{\{3\}}}\frac{\partial y_1^{\{3\}}}{\partial w_{13}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}\frac{\partial y_1^{\{m\}}}{\partial w_{13}}$$
$$= \frac{\partial E}{\partial y_1^{\{1\}}}x_3^{\{1\}} + \frac{\partial E}{\partial y_1^{\{2\}}}x_3^{\{2\}} + \frac{\partial E}{\partial y_1^{\{3\}}}x_3^{\{3\}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}x_3^{\{m\}}$$

$$\frac{\partial E}{\partial w_{21}} = \frac{\partial E}{\partial y_2^{\{1\}}}\frac{\partial y_2^{\{1\}}}{\partial w_{21}} + \frac{\partial E}{\partial y_2^{\{2\}}}\frac{\partial y_2^{\{2\}}}{\partial w_{21}} + \frac{\partial E}{\partial y_2^{\{3\}}}\frac{\partial y_2^{\{3\}}}{\partial w_{21}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}\frac{\partial y_2^{\{m\}}}{\partial w_{21}}$$
$$= \frac{\partial E}{\partial y_2^{\{1\}}}x_1^{\{1\}} + \frac{\partial E}{\partial y_2^{\{2\}}}x_1^{\{2\}} + \frac{\partial E}{\partial y_2^{\{3\}}}x_1^{\{3\}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}x_1^{\{m\}}$$

$$\frac{\partial E}{\partial w_{22}} = \frac{\partial E}{\partial y_2^{\{1\}}}\frac{\partial y_2^{\{1\}}}{\partial w_{22}} + \frac{\partial E}{\partial y_2^{\{2\}}}\frac{\partial y_2^{\{2\}}}{\partial w_{22}} + \frac{\partial E}{\partial y_2^{\{3\}}}\frac{\partial y_2^{\{3\}}}{\partial w_{22}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}\frac{\partial y_2^{\{m\}}}{\partial w_{22}}$$
$$= \frac{\partial E}{\partial y_2^{\{1\}}}x_2^{\{1\}} + \frac{\partial E}{\partial y_2^{\{2\}}}x_2^{\{2\}} + \frac{\partial E}{\partial y_2^{\{3\}}}x_2^{\{3\}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}x_2^{\{m\}}$$

$$\frac{\partial E}{\partial w_{23}} = \frac{\partial E}{\partial y_2^{\{1\}}}\frac{\partial y_2^{\{1\}}}{\partial w_{23}} + \frac{\partial E}{\partial y_2^{\{2\}}}\frac{\partial y_2^{\{2\}}}{\partial w_{23}} + \frac{\partial E}{\partial y_2^{\{3\}}}\frac{\partial y_2^{\{3\}}}{\partial w_{23}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}\frac{\partial y_2^{\{m\}}}{\partial w_{23}}$$
$$= \frac{\partial E}{\partial y_2^{\{1\}}}x_3^{\{1\}} + \frac{\partial E}{\partial y_2^{\{2\}}}x_3^{\{2\}} + \frac{\partial E}{\partial y_2^{\{3\}}}x_3^{\{3\}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}x_3^{\{m\}}$$

好了，现在我们得到了$\frac{\partial E}{\partial W}$的求导结果了。我们用矩阵的形式给出$\frac{\partial E}{\partial W}$:

$$\frac{\partial E}{\partial W_{2\times 3}} = \begin{pmatrix} \left[\frac{\partial E}{\partial y_1^{\{1\}}}x_1^{\{1\}} + \frac{\partial E}{\partial y_1^{\{2\}}}x_1^{\{2\}} + \frac{\partial E}{\partial y_1^{\{3\}}}x_1^{\{3\}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}x_1^{\{m\}}, \right. \\ \frac{\partial E}{\partial y_1^{\{1\}}}x_2^{\{1\}} + \frac{\partial E}{\partial y_1^{\{2\}}}x_2^{\{2\}} + \frac{\partial E}{\partial y_1^{\{3\}}}x_2^{\{3\}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}x_2^{\{m\}}, \\ \left.\frac{\partial E}{\partial y_1^{\{1\}}}x_3^{\{1\}} + \frac{\partial E}{\partial y_1^{\{2\}}}x_3^{\{2\}} + \frac{\partial E}{\partial y_1^{\{3\}}}x_3^{\{3\}} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}x_3^{\{m\}}\right], \\ \left[\frac{\partial E}{\partial y_2^{\{1\}}}x_1^{\{1\}} + \frac{\partial E}{\partial y_2^{\{2\}}}x_1^{\{2\}} + \frac{\partial E}{\partial y_2^{\{3\}}}x_1^{\{3\}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}x_1^{\{m\}}, \right. \\ \frac{\partial E}{\partial y_2^{\{1\}}}x_2^{\{1\}} + \frac{\partial E}{\partial y_2^{\{2\}}}x_2^{\{2\}} + \frac{\partial E}{\partial y_2^{\{3\}}}x_2^{\{3\}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}x_2^{\{m\}}, \\ \left.\frac{\partial E}{\partial y_2^{\{1\}}}x_3^{\{1\}} + \frac{\partial E}{\partial y_2^{\{2\}}}x_3^{\{2\}} + \frac{\partial E}{\partial y_2^{\{3\}}}x_3^{\{3\}} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}x_3^{\{m\}}\right] \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial E}{\partial y_1^{\{1\}}} & \frac{\partial E}{\partial y_1^{\{2\}}} & \frac{\partial E}{\partial y_2^{\{3\}}} & \cdots & \frac{\partial E}{\partial y_1^{\{m\}}} \\ \frac{\partial E}{\partial y_2^{\{1\}}} & \frac{\partial E}{\partial y_2^{\{2\}}} & \frac{\partial E}{\partial y_2^{\{3\}}} & \cdots & \frac{\partial E}{\partial y_2^{\{m\}}} \end{pmatrix} \cdot \begin{pmatrix} x_1^{\{1\}} & x_2^{\{1\}} & x_3^{\{1\}} \\ x_1^{\{2\}} & x_2^{\{2\}} & x_3^{\{2\}} \\ \vdots & \vdots & \vdots \\ x_1^{\{m\}} & x_2^{\{m\}} & x_3^{\{m\}} \end{pmatrix}$$

$$= \frac{\partial E}{\partial y} \cdot X^T$$

上面的计算公式得到关于*权重*W的每一个元素的求导结果，是m个样本累计相加得到的结果。所以我们要除以样本数m，等到平均值。所以我们的公式中有个除以m的操作。所以最终结果就是:

$$\frac{\partial E}{\partial W_{2\times 3}} = \frac{\partial E}{\partial y} \cdot X^T / m$$

那么我们得到了关于权重的导数，就可以更新权重了，更新权重公式如下：

$$W = W - \eta * \frac{\partial E}{\partial y} \cdot \frac{X^T}{m} \ (\eta是学习率)$$

最后我们来证明最后一个公式，$\frac{\partial E}{\partial b} = sum\left(\frac{\partial E}{\partial y}, axis = 1\right)/m$.

$$\frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial y_1^{\{1\}}}\frac{\partial y_1^{\{1\}}}{\partial b_1} + \frac{\partial E}{\partial y_1^{\{2\}}}\frac{\partial y_1^{\{2\}}}{\partial b_1} + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}}\frac{\partial y_1^{\{m\}}}{\partial b_1}$$
$$= \frac{\partial E}{\partial y_1^{\{1\}}} \cdot 1 + \frac{\partial E}{\partial y_1^{\{2\}}} \cdot 1 + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}} \cdot 1$$

$$\frac{\partial E}{\partial b_2} = \frac{\partial E}{\partial y_2^{\{1\}}}\frac{\partial y_2^{\{1\}}}{\partial b_2} + \frac{\partial E}{\partial y_2^{\{2\}}}\frac{\partial y_2^{\{2\}}}{\partial b_2} + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}}\frac{\partial y_2^{\{m\}}}{\partial b_2}$$
$$= \frac{\partial E}{\partial y_2^{\{1\}}} \cdot 1 + \frac{\partial E}{\partial y_2^{\{2\}}} \cdot 1 + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot 1$$

那么我们也就得到了$\frac{\partial E}{\partial b}$的求导结果了，我们以矩阵形式给出:

$$\frac{\partial E}{\partial b} = \begin{pmatrix} \frac{\partial E}{\partial y_1^{\{1\}}} \cdot 1 + \frac{\partial E}{\partial y_1^{\{2\}}} \cdot 1 + \cdots + \frac{\partial E}{\partial y_1^{\{m\}}} \cdot 1 \\ \frac{\partial E}{\partial y_2^{\{1\}}} \cdot 1 + \frac{\partial E}{\partial y_2^{\{2\}}} \cdot 1 + \cdots + \frac{\partial E}{\partial y_2^{\{m\}}} \cdot 1 \end{pmatrix}$$
$$= sum(\frac{\partial E}{\partial y}, axis = 1)$$

$\frac{\partial E}{\partial b}$的结果相当于将其元素按照列方向相加。同样地我们从$\frac{\partial E}{\partial b}$的表达式可以看到，这是$m$个样本累加的结果，所以我们要除以$m$得到平均值。所以最终公式为:

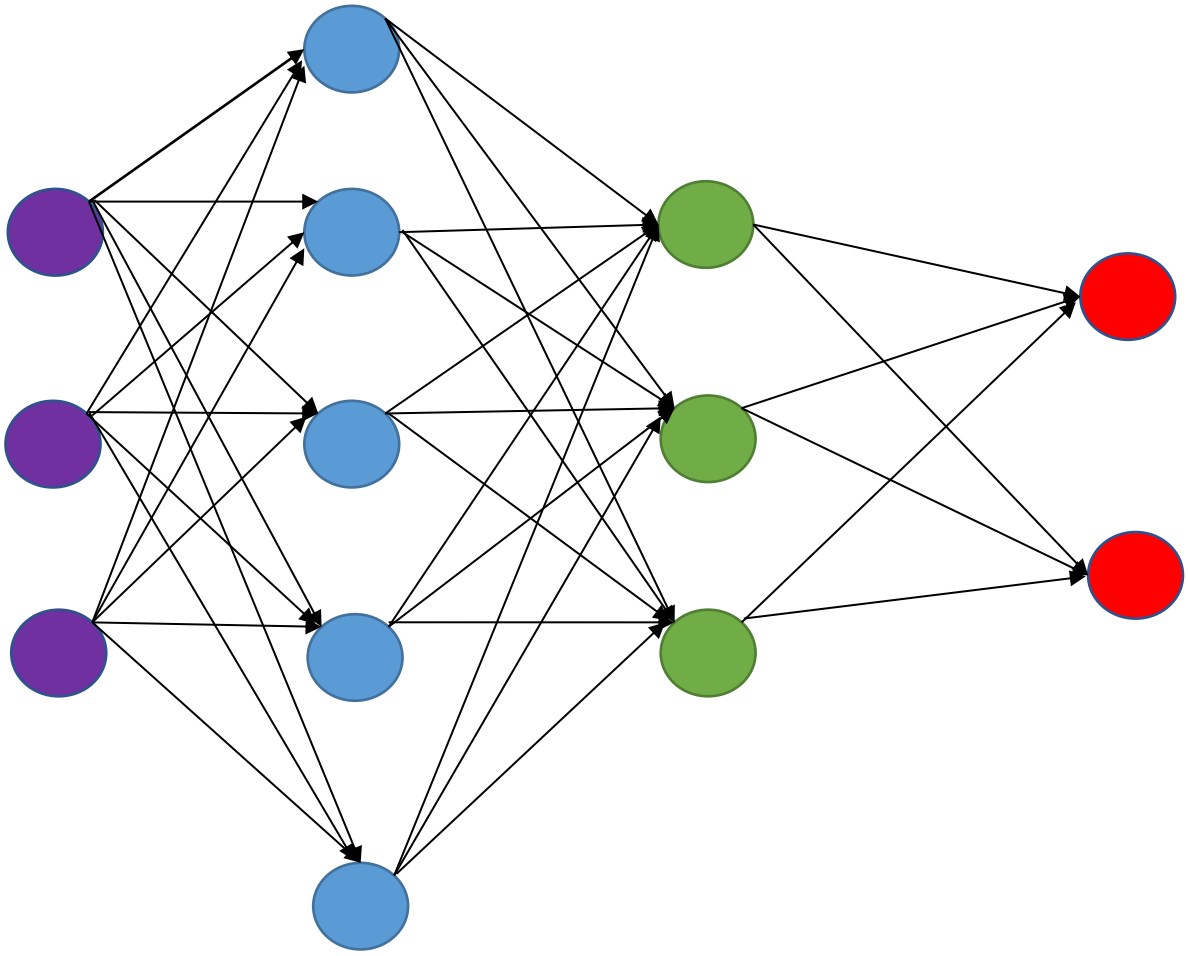
$$\frac{\partial E}{\partial b} = sum\left(\frac{\partial E}{\partial y}, axis = 1\right)/m$$

我们最后也就得到更新b的计算公式了:

$$b = b - \eta * sum\left(\frac{\partial E}{\partial y}, axis = 1\right)/m$$

好了关于反向传播的求导公式，我们给出了矩阵的计算形式。再懂得了这些计算公式后。我们间接可可对任何层次的神经网络进行关于梯度的计算了，从而可以实现反向传播。其实通篇文章下来，就是一个关于复合函数的求导运算，使用了链式法则。再就没有什么其它数学知识了。好了到这里为止大家应该是懂得了如何进行计算了，我们来做一个小练习吧。请看下图:

。



我这里给出了一个 3 层的神经网络的模型，其中两个隐藏层，蓝色和绿色节点的。一个输出层，红色节点的。不包含输入层，紫色节点的，所以是 3 层。我设置输入数据是$X_{3\times m}$。第一个隐藏层的输出，蓝色节点的，用$y_{5\times m}^{[1]}$来表示。中括号里的数字表示是神经网络的第几层。连接输入层和第一层隐藏层的权重，用$W_{5\times 3}^{[1]}$表示。第二层隐藏层的输出用$y_{3\times m}^{[2]}$表示。连接第一层隐藏层和第二层隐藏层之间的权重用$W_{3\times 5}^{[2]}$。最后的输出层的输出用$y_{2\times m}^{[3]}$，连接第二层隐藏层和输出层之间的权重矩阵用$W_{2\times 3}^{[3]}$。执行损失计算前，用激活函数σ对输出做了一次非线性激活。也就是$\sigma\left(y_{2\times m}^{[3]}\right)=a_{2\times m}^{[3]}$。$a$表示激活值。同时除了输出层的激活用 σ 激活函数，其余各层得到的输出使用了g函数来激活,$g\left(y_{5\times m}^{[1]}\right)=a_{5\times m}^{[1]}$, $g\left(y_{3\times m}^{[2]}\right)=a_{3\times m}^{[2]}$。最后使用的损失函数为 E 。让我们把所有的运算关系罗列下来，方便我们分析。如下所示：

前向传播关系:

$$W_{5\times3}^{[1]} \cdot X_{3\times m} + b_{5\times1}^{[1]} = y_{5\times m}^{[1]} \quad (1)$$

$$\mathcal{g}\left(y_{5\times m}^{[1]}\right) = a_{5\times m}^{[1]} \quad (2)$$

$$W_{3\times5}^{[2]} \cdot a_{5\times m}^{[1]} + b_{3\times1}^{[2]} = y_{3\times m}^{[2]} \ (3)$$

$$\mathcal{g}\left(y_{3\times m}^{[2]}\right) = a_{3\times m}^{[2]} \quad (4)$$

$$W_{2\times3}^{[3]} \cdot a_{3\times m}^{[2]} + b_{2\times1}^{[3]} = y_{2\times m}^{[3]} \ (5)$$

$$\sigma\left(y_{2\times m}^{[3]}\right) = a_{2\times m}^{[3]} \ (6)$$

$$\mathbb{E}\left(a_{2\times m}^{[3]}\right) = \text{Loss}$$

我们一步一步来分解这个过程，因为是反向传播，所以我们先从最后面的一层开始求起。′∗′这个符号表示矩阵是按位乘，而不是点乘。

$$\frac{\partial \mathbb{E}}{\partial y_{2\times m}^{[3]}} = \frac{\partial \mathbb{E}}{\partial a_{2\times m}^{[3]}} * \frac{\partial a_{2\times m}^{[3]}}{\partial y_{2\times m}^{[3]}}$$

这里之所以是按位乘而不是点乘，其实很明显，$y_{2\times m}^{[3]}$和$a_{2\times m}^{[3]}$是一对一关系，每个元素都对应着一条链式法则求导路径。损失函数$\mathbb{E}$对y中每个元素求导，根据链式求导规则，需要先对a中每个元素进行求导，然后a再对y的每个元素进行求导，才符合链式法则。所以当然是按位乘了。我们再给出之前推导的关于多个样本反向传播计算公式，便于我们阐述问题求解思路。

$$\frac{\partial E}{\partial x} = w^T \cdot \frac{\partial E}{\partial y} \quad (1)$$

$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial y} \cdot X^T / m \quad (2)$$

$$\frac{\partial E}{\partial b} = sum\left(\frac{\partial E}{\partial y}, axis = 1\right) / m \quad (3)$$

$$\frac{\partial \mathrm{E}}{\partial W_{2\times3}^{[3]}}=\frac{\partial \mathrm{E}}{\partial y_{2\times m}^{[3]}}\cdot\left(a_{3\times m}^{[2]}\right)^{T}/m$$

根据上面给出的前向传播关系(5),需要知道$\frac{\partial y_{2\times m}^{[3]}}{\partial W_{2\times3}^{[3]}}$，根据我们之前证明的多个样本反向传播计算公式(2),$\frac{\partial E}{\partial W}=\frac{\partial E}{\partial y}\cdot X^{T}/m$，这个公式里的$X^{T}$，就对应的是$W_{2\times3}^{[3]}\cdot a_{3\times m}^{[2]}+b_{2\times1}^{[3]}=y_{2\times m}^{[3]}$里的$a_{3\times m}^{[2]}$的转置。最后我们除以样本数$m$得到均值。

$$\frac{\partial \mathrm{E}}{\partial a_{3\times m}^{[2]}}=\left(W_{2\times3}^{[3]}\right)^{T}\cdot\frac{\partial \mathrm{E}}{\partial y_{2\times m}^{[3]}}$$

这里我们使用了$\frac{\partial E}{\partial x}=w^{T}\cdot\frac{\partial E}{\partial y}$这个公式，得到了上面关于$\frac{\partial \mathrm{E}}{\partial a_{3\times m}^{[2]}}$的计算关系。很方便啊。我们用同样的方法来处理其余的。

$$\frac{\partial \mathrm{E}}{\partial b_{2\times1}^{[3]}}=sum\left(\frac{\partial \mathrm{E}}{\partial y_{2\times m}^{[3]}},axis=1\right)/m$$

直接调用$\frac{\partial E}{\partial b}=sum\left(\frac{\partial E}{\partial y},axis=1\right)/m$，是不是很爽快啊，理解了原理，其余就很好办了。、

$$\frac{\partial \mathrm{E}}{\partial y_{3\times m}^{[2]}}=\frac{\partial \mathrm{E}}{\partial a_{3\times m}^{[2]}}*\frac{\partial a_{3\times m}^{[2]}}{\partial y_{3\times m}^{[2]}}$$

这从$\mathrm{g}\left(y_{3\times m}^{[2]}\right)=a_{3\times m}^{[2]}$这个关系能看出来。

$$\frac{\partial \mathrm{E}}{\partial a_{5\times m}^{[1]}}=\left(W_{3\times5}^{[2]}\right)^{T}\cdot\frac{\partial \mathrm{E}}{\partial y_{3\times m}^{[2]}}$$

从这个$W_{3\times5}^{[2]}\cdot a_{5\times m}^{[1]}+b_{3\times1}^{[2]}=y_{3\times m}^{[2]}$的等式就可以看出来参数彼此之间的关系了，然后再根据公式$\frac{\partial E}{\partial x}=w^{T}\cdot\frac{\partial E}{\partial y}$自然就得到上面的式子了。大家应该越发熟练这个过程了吧。

$$\frac{\partial \mathrm{E}}{\partial W_{3\times5}^{[2]}}=\frac{\partial \mathrm{E}}{\partial y_{3\times m}^{[2]}}\cdot\left(a_{5\times m}^{[1]}\right)^{T}/m$$

这个根据$W_{3\times5}^{[2]}\cdot a_{5\times m}^{[1]}+b_{3\times1}^{[2]}=y_{3\times m}^{[2]}$和$\frac{\partial E}{\partial W}=\frac{\partial E}{\partial y}\cdot X^T/m$，就可以得到上面的式子了。

$$\frac{\partial \mathrm{E}}{\partial b_{3\times1}^{[2]}}=\ sum\left(\frac{\partial \mathrm{E}}{\partial y_{3\times m}^{[2]}},axis=1\right)/m$$

根据$W_{3\times5}^{[2]}\cdot a_{5\times m}^{[1]}+b_{3\times1}^{[2]}=y_{3\times m}^{[2]}$，然后调用$\frac{\partial E}{\partial b}=sum\left(\frac{\partial E}{\partial y},axis=1\right)/m$轻松搞定啊。

$$\frac{\partial \mathrm{E}}{\partial y_{5\times m}^{[1]}}=\frac{\partial \mathrm{E}}{\partial a_{5\times m}^{[1]}}*\frac{\partial a_{5\times m}^{[1]}}{\partial y_{5\times m}^{[1]}}$$

根据$\mathrm{g}\left(y_{5\times m}^{[1]}\right)=a_{5\times m}^{[1]}$可得到上式。

$$\frac{\partial \mathrm{E}}{\partial X_{3\times m}}=\left(W_{3\times5}^{[2]}\right)^T\cdot\frac{\partial \mathrm{E}}{\partial y_{5\times m}^{[1]}}$$

根据$W_{5\times3}^{[1]}\cdot X_{3\times m}+b_{5\times1}^{[1]}=\ y_{5\times m}^{[1]}$的等式关系，以及公式$\frac{\partial E}{\partial x}=\ w^T\cdot\frac{\partial E}{\partial y}$可得上式。

$$\frac{\partial \mathrm{E}}{\partial W_{5\times3}^{[1]}}=\frac{\partial \mathrm{E}}{\partial y_{5\times m}^{[1]}}\cdot(X_{3\times m})^T/m$$

根据$W_{5\times3}^{[1]}\cdot X_{3\times m}+b_{5\times1}^{[1]}=\ y_{5\times m}^{[1]}$的等式关系,以及公式$\frac{\partial E}{\partial W}=\frac{\partial E}{\partial y}\cdot X^T/m$得到上式。

$$\frac{\partial \mathrm{E}}{\partial b_{5\times1}^{[1]}}=\ sum\left(\frac{\partial \mathrm{E}}{\partial y_{5\times m}^{[1]}},axis=1\right)/m$$

根据$W_{5\times3}^{[1]}\cdot X_{3\times m}+b_{5\times1}^{[1]}=\ y_{5\times m}^{[1]}$的等式关系,以及公式$\frac{\partial E}{\partial b}=sum\left(\frac{\partial E}{\partial y},axis=1\right)/m$得到上式。

好了我们将这个问题已经全部解决。套用我们之前给出的多个样本反向传播计算公式，然后搞清楚前向传播之间的计算关系，利用链式求导法则，就可以很容易计算反向传播的梯度。其实这些公式我们也已经清楚就是为了我们编写代码方便，通过矩阵的各种运算就能给出反向传播的梯度来进对权重等参数进行更新。矩阵的操作编写代码比较容易并且容易理解。好了这篇文章就结束了，希望对大家能有所帮助。