手写字符识别（下）

赵耀

回顾

▶ 既然希望神经网络的输出尽可能的接近真正想要预测的值，那么就可以通过比较当前网络的**预测值**与**目标值**，根据两者的差异情况来更新每一层的权重矩阵。

➤ 如何比较差异情况？ 损失函数(loss function)

➤ 如何更新权重矩阵？梯度下降(Gradient descend)

回顾

- 损失函数，又称为误差函数（error function），也有叫cost function。是用于衡量预测值和目标值差异的函数。

- loss function的输出值越高表示差异性越大，那么神经网络的训练目标就是尽可能的缩小差异

- 本次实验中采用的loss function为：$E(w) = \sum_{n=1}^{N} \sum_{k=1}^{k} -t_k^n ln y_k(x_n, w)$（推导过程详见准备知识.pdf文档）。训练目标是调整w，使得$E(w)$最小

回顾

- 梯度下降：通过使loss值向当前点对应梯度的反方向不断移动，来降低loss。一次移动多少是由学习速率（learning rate）来控制的。
- 难点1：有可能梯度下降只能找到局部最小值，找不到全局最小值
- 难点2：如何快速的计算梯度？如何更新隐藏层的权重？本次实验用到的是反向传播算法。（反向传播算法是一种计算梯度的方法，其贡献如同FFT）

背景知识

- 链式法则是微积分中的求导法则，用以求一个复合函数的导数。所谓的复合函数，是指以一个函数作为另一个函数的自变量。

- 链式法则描述：若h(x)=f(g(x))，则h' (x)=f'(g(x))g'(x)

- 例子：

例如f(x)=2x+2，g(x)=3x+3，g(f(x))就是一个复合函数，并且g'(f(x))=6

误差反向传播-整体代价函数

- 训练目标是调整W，使得损失函数$E(w,b) = \sum_{n=1}^{N}\sum_{k=1}^{10} -t_k^n \ln y_k(x_n,w)$越来越小

- 在批量梯度下降法中，给定一个包含n个样例的数据集，我们可以定义整体代价函数为：$E_{total}(w,b) = \frac{1}{N}\sum_{n=1}^{N} E(W^{(t)}, B^{(t)}) + \frac{\lambda}{2}|W^{(t)}|^2$
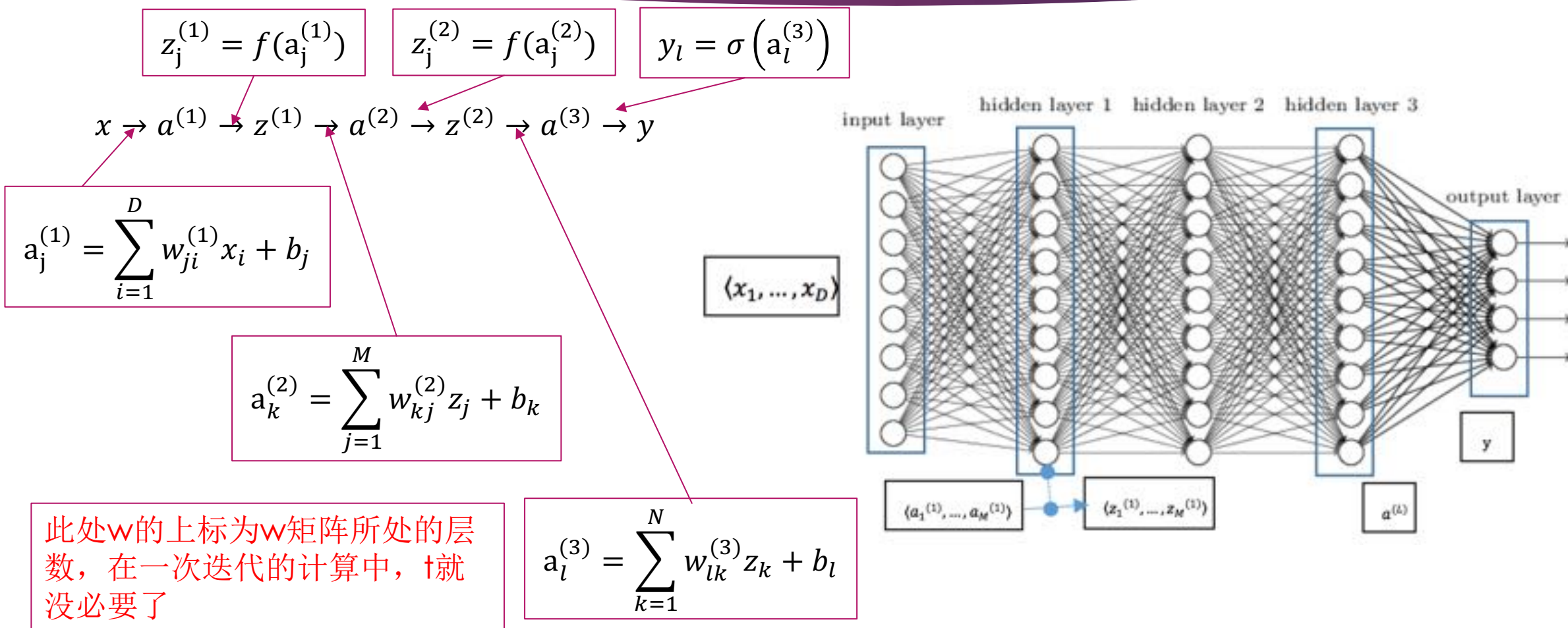
- 第二项是一个正则化项（也叫**权重衰减项**）：L_2正则，其目的是减小权重的幅度，防止过度拟合。

梯度下降

▶ 梯度下降：通过使loss值向当前点对应梯度的反方向不断移动，来降低loss。t+1次迭代的W是上一次的W减去$\eta * E_{total}(W^{(t)}, B^{(t)})$对$W^{(t)}$的偏导值，同样的方式处理B偏置向量。一次移动多少是由学习速率η来控制的。在本实验中，前50的大周期推荐设置η=0.1；后50次的大周期推荐设置η=0.01。

▶ $W^{(t+1)} = W^{(t)} - \eta \frac{\partial E_{total}(W^{(t)}, B^{(t)})}{\partial W^{(t)}}$

▶ $B^{(t+1)} = B^{(t)} - \eta \frac{\partial E_{total}(W^{(t)}, B^{(t)})}{\partial B^{(t)}}$

难点在于偏导怎么计算？

- $\dfrac{\partial E_{total}(W^{(t)}, B^{(t)})}{\partial W^{(t)}}$

$= \dfrac{\partial \left(\frac{1}{N} \sum_{n=1}^{N} E_n(W^{(t)}, B^{(t)}) + \frac{\lambda}{2} |W^{(t)}|^2 \right)}{\partial W^{(t)}}$

$= \dfrac{1}{N} \sum_{n=1}^{N} \dfrac{\partial E_n(W^{(t)}, B^{(t)})}{\partial W^{(t)}} + \lambda W^{(t)}$

- $\dfrac{\partial E_{total}(W^{(t)}, B^{(t)})}{\partial B^{(t)}} = \dfrac{1}{N} \sum_{n=1}^{N} \dfrac{\partial E_n(W^{(t)}, B^{(t)})}{\partial B^{(t)}}$

回顾一次迭代前向计算的过程

$$z_j^{(1)} = f(a_j^{(1)})$$

$$z_j^{(2)} = f(a_j^{(2)})$$

$$y_l = \sigma\left(a_l^{(3)}\right)$$

$$x \rightarrow a^{(1)} \rightarrow z^{(1)} \rightarrow a^{(2)} \rightarrow z^{(2)} \rightarrow a^{(3)} \rightarrow y$$

$$a_j^{(1)} = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + b_j$$

$$a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + b_k$$

$$a_l^{(3)} = \sum_{k=1}^{N} w_{lk}^{(3)} z_k + b_l$$

此处w的上标为w矩阵所处的层数，在一次迭代的计算中，t就没必要了

E对 $w^{(3)}$, $b^{(3)}$ 求导

$\blacktriangleright$ $\dfrac{\partial E}{\partial W^{(3)}} = \dfrac{\partial E\,(y)}{\partial y} \dfrac{\partial y(a^{(3)})}{\partial a^{(3)}} \dfrac{\partial a^{(3)})}{\partial w^{(3)}}$

$\blacktriangleright$ $\dfrac{\partial E}{\partial b^{(3)}} = \dfrac{\partial E\,(y)}{\partial y} \dfrac{\partial y(a^{(3)})}{\partial a^{(3)}} \dfrac{\partial a^{(3)})}{\partial b^{(3)}}$

记为: $\delta^{(3)} = \dfrac{\partial E}{\partial a^{(3)}} = \dfrac{\partial E\,(y)}{\partial y} \dfrac{\partial y(a^{(3)})}{\partial a^{(3)}}$

$z_j^{(1)} = f(a_j^{(1)})$

$z_j^{(2)} = f(a_j^{(2)})$

$y_l = \sigma\left(a_l^{(3)}\right)$

$x \rightarrow a^{(1)} \rightarrow z^{(1)} \rightarrow a^{(2)} \rightarrow z^{(2)} \rightarrow a^{(3)} \rightarrow y$

$a_j^{(1)} = \displaystyle\sum_{i=1}^{D} w_{ji}^{(1)} x_i + b_j^{(1)}$

$a_k^{(2)} = \displaystyle\sum_{j=1}^{M} w_{kj}^{(2)} z_j^{(1)} + b_k^{(2)}$

$a_l^{(3)} = \displaystyle\sum_{k=1}^{N} w_{lk}^{(3)} z_k^{(2)} + b_l^{(3)}$

求导 $\delta^{(3)}$-1

- $\delta^{(3)} = \dfrac{\partial E\ (y)}{\partial y}\dfrac{\partial y(a^{(3)})}{\partial a^{(3)}} = \dfrac{\partial \sum_{k=1}^{10} -t_k^n ln y_k}{\partial y}\dfrac{\partial(\boxed{\dfrac{e^{a_k^{(3)}}}{\sum_{k=1}^{10} e^{a_k^{(3)}}}})}{\partial a^{(3)}}$ $\longrightarrow$ y_k

- 假设 t_k^{10} 中 k=j 的时候 t_j^n 位为1，那么当 k=j 时：

- $\delta_j^{(3)} = \dfrac{\partial(-ln y_j)}{\partial y_j}\dfrac{\partial(\dfrac{e^{a_j^{(3)}}}{\sum_{k=1}^{10} e^{a_k^{(3)}}})}{\partial a_j^{(3)}}$

$= (-\dfrac{1}{y_j})(\dfrac{e^{a_j^{(3)}} \sum_{k=1}^{10} e^{a_k^{(3)}} - e^{a_j^{(3)}} e^{a_j^{(3)}}}{(\sum_{k=1}^{10} e^{a_k^{(3)}})^2})$

$= (-\dfrac{1}{y_j})(\dfrac{e^{a_j^{(3)}}}{\sum_{k=1}^{10} e^{a_k^{(3)}}} - \dfrac{e^{a_j^{(3)}} e^{a_j^{(3)}}}{(\sum_{k=1}^{10} e^{a_k^{(3)}})^2})$

$= (-\dfrac{1}{y_j})(y_j - y_j^2) = y_j - 1$

相关求导公式：

$y = ln x \quad y' = \dfrac{1}{x}$

$y = e^x \quad y' = e^x$

$y = x^n \quad y' = nx^{n-1}$

$\left(\dfrac{v(x)}{u(x)}\right)' = \dfrac{u(x)v'(x) - u'(x)v(x)}{(u(x))^2}$

求导$\delta^{(3)}$-2

▶ 假设t_k^n中k=j的时候t_j^n位为1，那么当$k \neq j$时：

▶ $\delta_i^{(3)} = \dfrac{\partial(-ln y_j)}{\partial y_j} \dfrac{\partial(\dfrac{e^{a_j^{(3)}}}{\Sigma_{k=1}^{10} e^{a_k^{(3)}}})}{\partial a_i^{(3)}}$

$= (-\dfrac{1}{y_j})(\dfrac{0 * \Sigma_{k=1}^{10} e^{a_k^{(3)}} - e^{a_j^{(3)}} e^{a_i^{(3)}}}{(\Sigma_{k=1}^{10} e^{a_k^{(3)}})^2})$

$= (-\dfrac{1}{y_j})(- \dfrac{e^{a_j^{(3)}} e^{a_i^{(3)}}}{(\Sigma_{k=1}^{10} e^{a_k^{(3)}})^2})$

$= (-\dfrac{1}{y_j})(-y_j y_i) = y_j$

求导$\delta^{(3)}$-3

- 由于综合起来t_k^n中k=j的时候t_j^n位为1，$k \neq j$时为0，所以
- $\delta^{(3)} = y - t$

对 $\dfrac{\partial E}{\partial w_{ij}^{(3)}}$，$\dfrac{\partial E}{\partial b_i^{(3)}}$ 求导

▶ $\dfrac{\partial E}{\partial w_{ij}^{(3)}} = \dfrac{\partial E}{\partial a_i^{(3)}}\dfrac{\partial(a_i^{(3)})}{\partial w_{ij}^{(3)}} = \dfrac{\partial E}{\partial a_i^{(3)}}\dfrac{\partial(\sum_{k=1}^{N} w_{ik}^{(3)} z_k^{(2)}+b_i^{(2)})}{\partial w_{ij}^{(3)}}=\delta_i^{(3)} z_j^{(2)}$

▶ 综合起来，$\dfrac{\partial E\,(W^{(3)},b^{(3)})}{\partial W^{(3)}}=\delta^{(3)}(z^{(2)})^T$

▶ $\dfrac{\partial E}{\partial b_i^{(3)}} = \dfrac{\partial E}{\partial a_i^{(3)}}\dfrac{\partial(a_i^{(3)})}{\partial b_i^{(3)}} = \dfrac{\partial E}{\partial a_i^{(3)}}\dfrac{\partial(\sum_{k=1}^{N} w_{ik}^{(3)} z_k^{(2)}+b_i^{(2)})}{\partial b_i^{(3)}}=\delta_i^{(3)} * 1=\delta_i^{(3)}$

E对$w^{(2)}$, $b^{(2)}$求导分解

$\frac{\partial E}{\partial W^{(2)}} = \boxed{\frac{\partial E\ (y)}{\partial y} \frac{\partial y(a^{(3)})}{\partial a^{(3)}} \frac{\partial a^{(3)})}{\partial z^{(2)}} \frac{\partial z^{(2)})}{\partial a^{(2)}}} \frac{\partial a^{(2)})}{\partial W^{(2)}}$

$\frac{\partial E}{\partial b^{(2)}} = \boxed{\frac{\partial E\ (y)}{\partial y} \frac{\partial y(a^{(3)})}{\partial a^{(3)}} \frac{\partial a^{(3)})}{\partial z^{(2)}} \frac{\partial z^{(2)})}{\partial a^{(2)}}} \frac{\partial a^{(2)})}{\partial b^{(2)}}$

记为： $\delta^{(2)} = \frac{\partial E\ (y)}{\partial y} \frac{\partial y(a^{(3)})}{\partial a^{(3)}} \frac{\partial a^{(3)})}{\partial z^{(2)}} \frac{\partial z^{(2)})}{\partial a^{(2)}} = \frac{\partial E}{\partial a^{(2)}}$

$z_j^{(1)} = f(a_j^{(1)})$

$z_j^{(2)} = f(a_j^{(2)})$

$y_l = \sigma\left(a_l^{(3)}\right)$

$x \to a^{(1)} \to z^{(1)} \to a^{(2)} \to z^{(2)} \to a^{(3)} \to y$

$a_j^{(1)} = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + b_j$

$a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(2)} z_j^{(1)} + b_k^{(2)}$

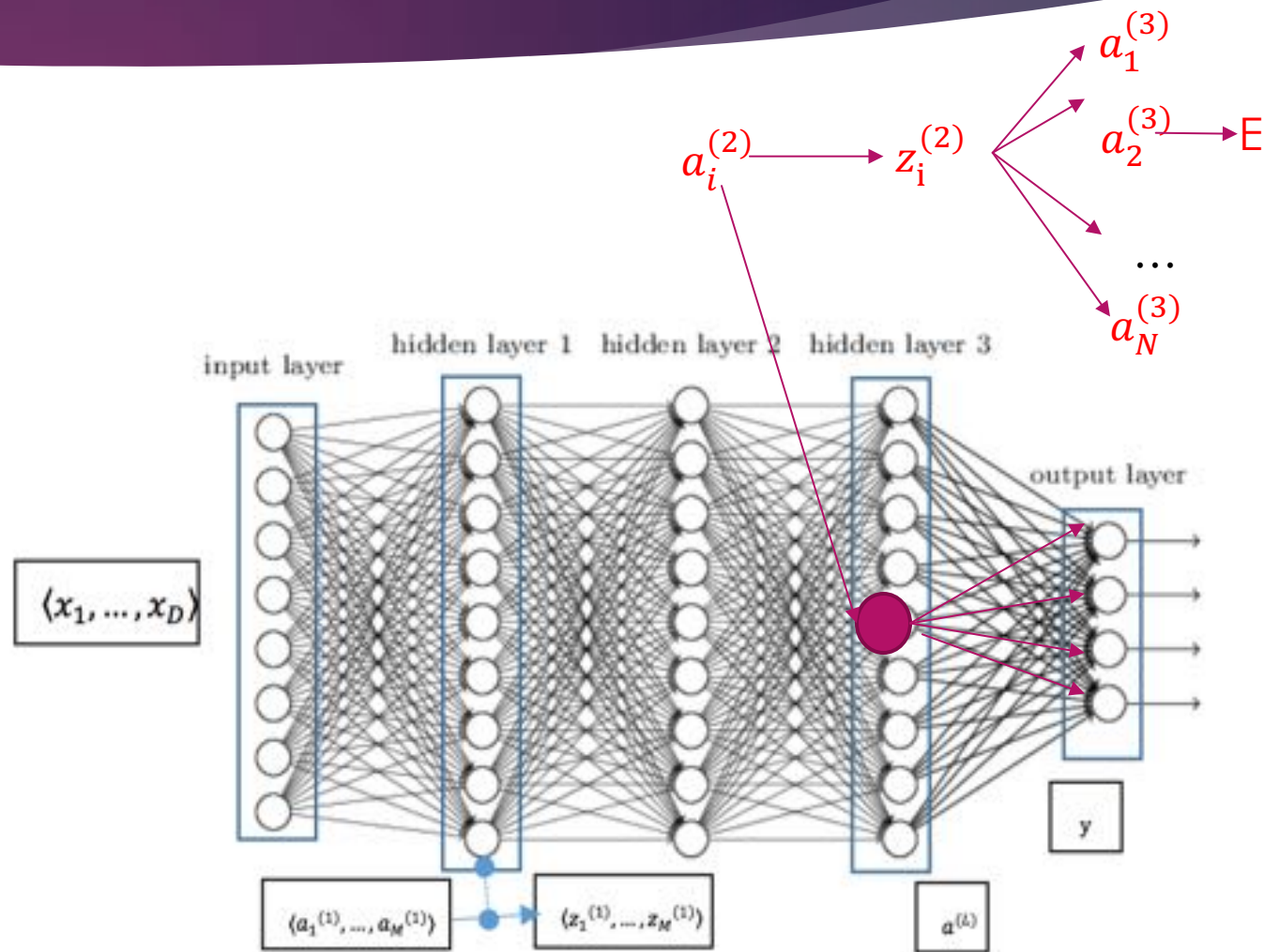$a_l^{(3)} = \sum_{k=1}^{N} w_{jk}^{(3)} z_k^{(2)} + b_k^{(3)}$

求 $\delta^{(2)}$

▶ $\delta_i^{(2)} = \sum_{j=1}^{L} \delta_j^{(3)} \frac{\partial a_j^{(3)}}{\partial z_i^{(2)}} \frac{\partial z_i^{(2)}}{\partial a_i^{(2)}}$

$= \sum_{j=1}^{L} \delta_j^{(3)} \frac{\partial (\sum_{k=1}^{N} w_{jk}^{(3)} z_k + b_k^{(3)})}{\partial z_i^{(2)}} \frac{\partial z_i^{(2)}}{\partial a_i^{(2)}}$

$= \sum_{j=1}^{L} \delta_j^{(3)} w_{ji}^{(3)} f'(z_i^{(2)})$

表示成矩阵的形式为：

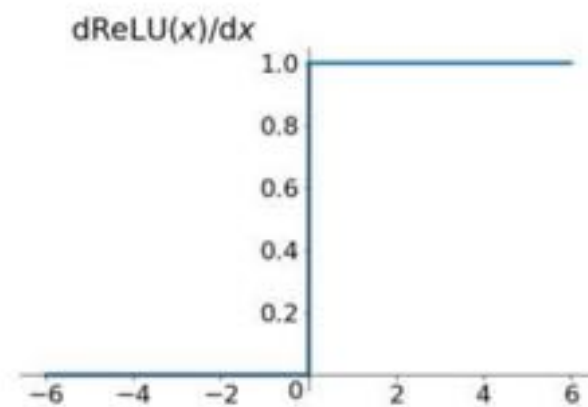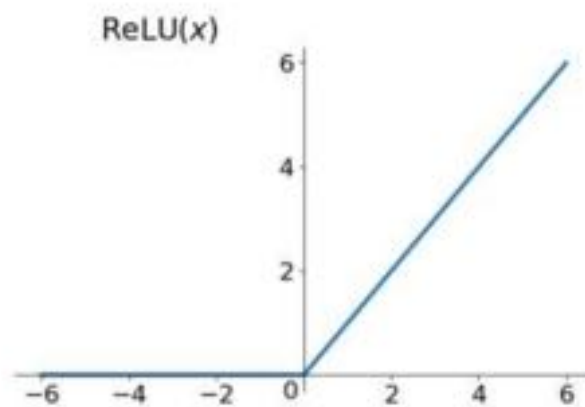$\delta^{(2)} = \delta^{(3)} (w^{(3)})^T \odot f'(z^{(2)})$

此处，$z_j^{(2)} = f(a_j^{(2)})$，f为relu函数。

$a_i^{(2)} \longrightarrow z_i^{(2)}$

$a_1^{(3)}$

$a_2^{(3)} \rightarrow E$

$\ldots$

$a_N^{(3)}$

input layer   hidden layer 1   hidden layer 2   hidden layer 3

$\langle x_1, \ldots, x_D \rangle$

output layer

y

$\langle a_1^{(1)}, \ldots, a_M^{(1)} \rangle$ $\langle z_1^{(1)}, \ldots, z_M^{(1)} \rangle$ $a^{(L)}$

relu函数及其求导

▶ relu函数的定义：$f(x) = \max(0, x)$

▶ relu函数的导数：$f'^{(x)} =$
$$\begin{cases} if\ x > 0, f'(x) = 1 \\ \quad else\ f'(x) = 0 \end{cases}$$

对 $\dfrac{\partial E}{\partial w_{ij}^{(2)}}$, $\dfrac{\partial E}{\partial b_i^{(2)}}$ 求导

- $\dfrac{\partial E}{\partial w_{ij}^{(2)}} = \dfrac{\partial E}{\partial a_i^{(2)}} \dfrac{\partial(a_i^{(2)})}{\partial w_{ij}^{(2)}} = \delta_i^{(2)} \dfrac{\partial(\sum_{j=1}^{M} w_{kj}^{(2)} z_j^{(1)} + b_k^{(2)})}{\partial w_{ij}^{(2)}} = \delta_i^{(2)} z_j^{(1)}$

- 综合起来，$\dfrac{\partial E}{\partial W^{(2)}} = \delta^{(2)} (z^{(1)})^T$

- $\dfrac{\partial E}{\partial b_i^{(2)}} = \dfrac{\partial E}{\partial a_i^{(2)}} \dfrac{\partial(a_i^{(2)})}{\partial b_i^{(2)}} = \dfrac{\partial E}{\partial a_i^{(2)}} \dfrac{\partial(\sum_{j=1}^{M} w_{kj}^{(2)} z_j^{(1)} + b_k^{(2)})}{\partial b_i^{(3)}} = \delta_i^{(2)} * 1 = \delta_i^{(2)}$

对 $\dfrac{\partial E}{\partial w_{ij}^{(1)}}$, $\dfrac{\partial E}{\partial b_i^{(1)}}$ 求导

- 同 $\delta^{(2)}$ 的过程，求出：$\delta^{(1)} = \delta^{(2)}(w^{(2)})^T \odot f'(z^{(1)})$

- $\dfrac{\partial E}{\partial W^{(1)}} = \delta^{(1)}(x)^T$

- $\dfrac{\partial E}{\partial b^{(1)}} = \delta^{(1)}$

综上，误差反向传播求导及调整步骤

▶ 求出输出层的$\delta^{(3)}$：本例中经过推导 $\delta^{(3)} = y - t$

▶ 求出$\frac{\partial E}{\partial w_{ij}^{(3)}}$，$\frac{\partial E}{\partial w_{ij}^{(3)}}$：本例中经过推导 $\frac{\partial E}{\partial w_{ij}^{(3)}} = \delta_i^{(3)} z_j^{(2)}$，$\frac{\partial E}{\partial b_i^{(3)}} = \delta_i^{(3)}$，推导时从后往前，i表示的是L+1层维度，j表示的是L层维度，然而前向计算时矩阵W的i表示的是L层维度，j表示L+1层维度，这里要特别注意。此处以前置矩阵定义为主。所以以下做调整时是有将推导的公式做了转置处理。

▶ 调整$W^{(3)(t+1)} = W^{(3)(t)} - \eta \frac{1}{N} \sum_{n=1}^{N} \frac{\partial E}{\partial W^{(3)(t)}} - \eta \lambda W^{(3)(t)} = W^{(3)(t)} - \eta z^{(2)(T)} \delta^{(3)} - \eta \lambda W^{(3)(t)}$，此处的加入的$t$表示的是迭代的次数，此处的$z$和$\delta$均为加了样本点维度的矩阵，跟推导时用的单个样本点不同。

▶ 调整$b^{(3)(t+1)} = b^{(3)(t)} - \eta \frac{1}{N} \sum_{n=1}^{N} \delta_{ni}^{(3)}$，此处的$\delta$为加了样本点维度的矩阵

▶ 求出倒数第二层的$\delta^{(2)}$：本例经过推导$\delta^{(2)} = \delta^{(3)} (w^{(3)(t)})^T \odot f'(z^{(2)})$

▶ 调整$W^{(2)(t+1)} = W^{(2)(t)} - \eta z^{(1)(T)} \delta^{(2)} - \eta \lambda W^{(2)(t)}$，此处的加入的$t$表示的是迭代的次数，此处的$z$和$\delta$均为加了样本点维度的矩阵，跟推导时用的单个样本点不同。

▶ 调整$b^{(2)(t+1)} = b^{(2)(t)} - \eta \frac{1}{N} \sum_{n=1}^{N} \delta_{ni}^{(2)}$，此处的$\delta$为加了样本点维度的矩阵

▶ 求出第一层的$\delta^{(1)}$：本例经过推导$\delta^{(1)} = \delta^{(2)} (w^{(2)(t)})^T \odot f'(z^{(1)})$

▶ 调整$W^{(1)(t+1)} = W^{(1)(t)} - \eta x^{(T)} \delta^{(1)} - \eta \lambda W^{(1)(t)}$，此处的加入的$t$表示的是迭代的次数，此处的$x$和$\delta$均为加了样本点维度的矩阵，跟推导时用的单个样本点不同。

▶ 调整$b^{(1)(t+1)} = b^{(1)(t)} - \eta \frac{1}{N} \sum_{n=1}^{N} \delta_{ni}^{(1)}$，此处的$\delta$为加了样本点维度的矩阵