- Tanh 也有消失梯度问题。

5.3.7 Softmax

Softmax 函数计算事件在 n 个不同事件上的概率分布。一般来说，这个函数将计算每个目标类在所有可能目标类上的概率。随后计算的概率将有助于确定给定输入的目标类。

5.4 架构设计

神经网络设计的另一个关键点是确定它的架构。架构（architecture）一词是指网络的整体结构：它应该具有多少单元，以及这些单元应该如何连接。

大多数神经网络被组织成称为层的单元组。大多数神经网络架构将这些层布置成链式结构，其中每一层都是前一层的函数。

在这些链式架构中，主要的架构考虑是选择网络的深度和每一层的宽度。我们将会看到，即使只有一个隐藏层的网络也足够适应训练集。更深层的网络通常能够对每一层使用更少的单元数和更少的参数，并且经常容易泛化到测试集，但是通常也更难以优化。对于一个具体的任务，理想的网络架构必须通过实验，观测在验证集上的误差来找到。

5.5 前向与反向传播

5.5.1 前向传播

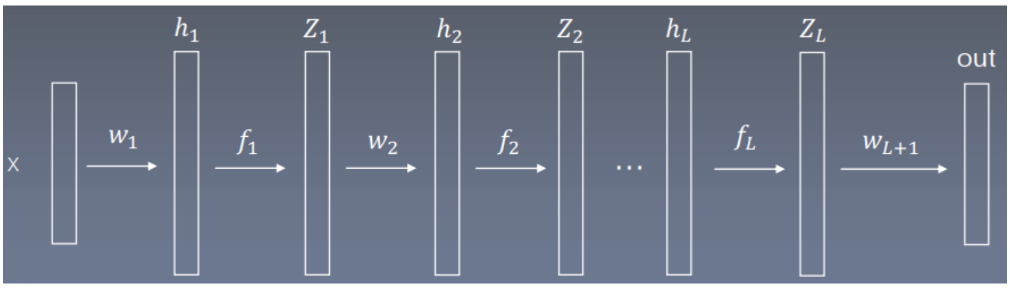


图 46 前向传播示意图

假设 X 为 N*m 的矩阵（其中 N 为样本数，m 为特征维数）

h_1 与 Z_1 的维数为 $m_1 \rightarrow W_1$ 为 $m*m_1$ 的矩阵，$b_1 \in R^{m1}$，

h_2 与 Z_2 的维数为 $m_2 \rightarrow W_2$ 为 m_1*m_2 的矩阵，$b_2 \in R^{m2}$，

…

h_L 与 Z_L 的维数为 $m_L \rightarrow W_L$ 为 $m_{L-1}*m_L$ 的矩阵，$b_L \in R^{m_L}$。

前向算法：

$h_1=XW_1+b_1^{\wedge}$，$Z_1=f_1(h_1)$，其中 $b_1^{\wedge}$为 b_1^T 后沿着行方向扩展 N 行，即

$$\mathrm{b1}^{\wedge} = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{Nm} \end{pmatrix}$$

…

$h_2=Z_1W_2+b_2^{\wedge}$，$Z_2=f_2(h_2)$，

$h_L=Z_{L-1}W_L+b_L^{\wedge}$，$Z_L=f_L(h_L)$，

$out=Z_LW_{L+1}+b_{L+1}^{\wedge}$

假设输出为 n 维，则 out 为 N*n 的矩阵。

$\frac{\partial L}{\partial out}$可以根据 mse 或者交叉熵 ce 准则求出（均是对 out 求导，可以看出是网络输出矩阵与标签矩阵相减）。

$$J=\frac{1}{2N}\sum_{i=1}^{N}\|y^i-\tilde{y}^i\|^2$$

$$\frac{\partial J}{\partial y^i}=\frac{1}{2N}\sum_{i=1}^{N}(y^i-\tilde{y}^i)*2$$

$$=\frac{1}{N}\sum_{i=1}^{N}(y^i-\tilde{y}^i)$$

$$\frac{\partial J}{\partial y_m}=\frac{\partial J}{\partial y_m}(\log\sum_{i=0}^{n-1}e^{y_i}-y_k)$$

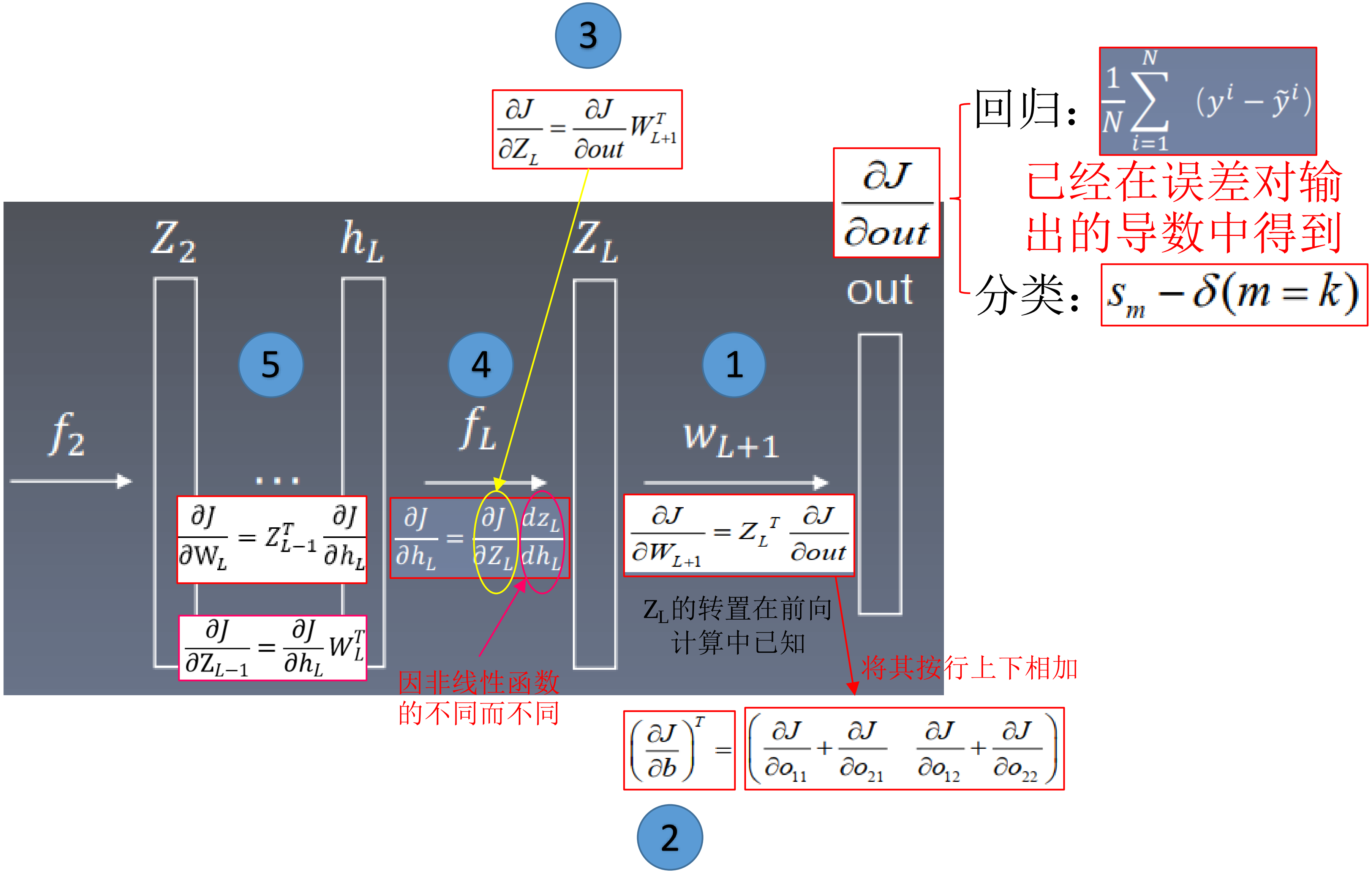
$$=\frac{e^{y_m}}{\sum_{i=0}^{n-1}e^{y_i}}-\delta(m=k)$$

$$=s_m-\delta(m=k)$$

图 47 回归问题和分类问题的损失函数

5.5.2 反向传播

通过特例推导一般式，可以假设输入 2 个样本，输出 2 个标签，如下图所示。


3
$\frac{\partial J}{\partial Z_L} = \frac{\partial J}{\partial out} W_{L+1}^T$
回归：$\frac{1}{N}\sum_{i=1}^{N}(y^i - \tilde{y}^i)$
$\frac{\partial J}{\partial out}$
已经在误差对输出的导数中得到
分类：$s_m - \delta(m=k)$
Z_2
h_L
Z_L
out
5
4
1
f_2
f_L
w_{L+1}
...
$\frac{\partial J}{\partial W_L} = Z_{L-1}^T \frac{\partial J}{\partial h_L}$
$\frac{\partial J}{\partial h_L} = \frac{\partial J}{\partial Z_L} \frac{dz_L}{dh_L}$
$\frac{\partial J}{\partial W_{L+1}} = Z_L{}^T \frac{\partial J}{\partial out}$
$\frac{\partial J}{\partial Z_{L-1}} = \frac{\partial J}{\partial h_L} W_L^T$
Z_L的转置在前向计算中已知
将其按行上下相加
因非线性函数的不同而不同
$\left(\frac{\partial J}{\partial b}\right)^T = \left(\frac{\partial J}{\partial o_{11}} + \frac{\partial J}{\partial o_{21}} \quad \frac{\partial J}{\partial o_{12}} + \frac{\partial J}{\partial o_{22}}\right)$
2

①

$$Z_L=\begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \end{pmatrix}_{2\times 3}, W_{L+1}=\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}_{3\times 2}, \tilde{b}_{L+1}=\begin{pmatrix} b_1 & b_2 \\ b_1 & b_2 \end{pmatrix}_{2\times 2}, out=\begin{pmatrix} o_{11} & o_{12} \\ o_{21} & o_{22} \end{pmatrix}$$

$$Z_L \cdot W_{L+1}=\begin{pmatrix} z_{11}w_{11}+z_{12}w_{21}+z_{13}w_{31} & , z_{11}w_{12}+z_{12}w_{22}+z_{13}w_{32} \\ z_{21}w_{11}+z_{22}w_{21}+z_{23}w_{31} & , z_{21}w_{12}+z_{22}w_{22}+z_{23}w_{32} \end{pmatrix}$$

$$o_{11}=z_{11}w_{11}+z_{12}w_{21}+z_{13}w_{31}+b_1,$$
$$o_{12}=z_{11}w_{12}+z_{12}w_{22}+z_{13}w_{32}+b_2,$$
$$o_{21}=z_{21}w_{11}+z_{22}w_{21}+z_{23}w_{31}+b_1,$$
$$o_{22}=z_{21}w_{12}+z_{22}w_{22}+z_{23}w_{32}+b_2.$$

$$\frac{\partial J}{\partial w_{11}}=\frac{\partial J}{\partial o_{11}}z_{11}+\frac{\partial J}{\partial o_{21}}z_{21}, \frac{\partial J}{\partial w_{12}}=\frac{\partial J}{\partial o_{12}}z_{11}+\frac{\partial J}{\partial o_{22}}z_{21}$$
$$\frac{\partial J}{\partial w_{21}}=\frac{\partial J}{\partial o_{11}}z_{12}+\frac{\partial J}{\partial o_{21}}z_{22}, \frac{\partial J}{\partial w_{22}}=\frac{\partial J}{\partial o_{12}}z_{12}+\frac{\partial J}{\partial o_{22}}z_{22}$$
$$\frac{\partial J}{\partial w_{31}}=\frac{\partial J}{\partial o_{11}}z_{13}+\frac{\partial J}{\partial o_{21}}z_{23}, \frac{\partial J}{\partial w_{32}}=\frac{\partial J}{\partial o_{12}}z_{13}+\frac{\partial J}{\partial o_{22}}z_{23}$$

$$\begin{pmatrix} \frac{\partial J}{\partial w_{11}} & \frac{\partial J}{\partial w_{12}} \\ \frac{\partial J}{\partial w_{21}} & \frac{\partial J}{\partial w_{22}} \\ \frac{\partial J}{\partial w_{31}} & \frac{\partial J}{\partial w_{32}} \end{pmatrix}=\begin{pmatrix} z_{11} & z_{21} \\ z_{12} & z_{22} \\ z_{13} & z_{23} \end{pmatrix}\begin{pmatrix} \frac{\partial J}{\partial o_{11}} & \frac{\partial J}{\partial o_{12}} \\ \frac{\partial J}{\partial o_{21}} & \frac{\partial J}{\partial o_{22}} \end{pmatrix}$$

$$\boxed{\frac{\partial J}{\partial W_{L+1}}=Z_L^T\frac{\partial J}{\partial out}}$$

②

$$\begin{cases} \frac{\partial J}{\partial b_1}=\frac{\partial J}{\partial o_{11}}+\frac{\partial J}{\partial o_{21}} \\ \frac{\partial J}{\partial b_2}=\frac{\partial J}{\partial o_{12}}+\frac{\partial J}{\partial o_{22}} \end{cases}, \Rightarrow \boxed{\left(\frac{\partial J}{\partial b}\right)^T=\begin{pmatrix} \frac{\partial J}{\partial b_1} & \frac{\partial J}{\partial b_2} \end{pmatrix}=\begin{pmatrix} \frac{\partial J}{\partial o_{11}}+\frac{\partial J}{\partial o_{21}} & \frac{\partial J}{\partial o_{12}}+\frac{\partial J}{\partial o_{22}} \end{pmatrix}}=\text{将}\frac{\partial J}{\partial out}\text{的每一行加起来}$$

③

$$\frac{\partial J}{\partial z_{11}}=\frac{\partial J}{\partial o_{11}}w_{11}+\frac{\partial J}{\partial o_{12}}w_{12};\frac{\partial J}{\partial z_{12}}=\frac{\partial J}{\partial o_{11}}w_{21}+\frac{\partial J}{\partial o_{12}}w_{22};\frac{\partial J}{\partial z_{13}}=\frac{\partial J}{\partial o_{11}}w_{31}+\frac{\partial J}{\partial o_{12}}w_{32}$$

$$\frac{\partial J}{\partial z_{21}}=\frac{\partial J}{\partial o_{21}}w_{11}+\frac{\partial J}{\partial o_{22}}w_{12};\frac{\partial J}{\partial z_{22}}=\frac{\partial J}{\partial o_{21}}w_{21}+\frac{\partial J}{\partial o_{12}}w_{22};\frac{\partial J}{\partial z_{23}}=\frac{\partial J}{\partial o_{21}}w_{31}+\frac{\partial J}{\partial o_{22}}w_{32}$$

$$\begin{pmatrix}\frac{\partial J}{\partial z_{11}} & \frac{\partial J}{\partial z_{12}} & \frac{\partial J}{\partial z_{13}}\\ \frac{\partial J}{\partial z_{21}} & \frac{\partial J}{\partial z_{22}} & \frac{\partial J}{\partial z_{23}}\end{pmatrix}=\begin{pmatrix}\frac{\partial J}{\partial o_{11}} & \frac{\partial J}{\partial o_{12}}\\ \frac{\partial J}{\partial o_{21}} & \frac{\partial J}{\partial o_{22}}\end{pmatrix}\begin{pmatrix}w_{11} & w_{21} & w_{31}\\ w_{12} & w_{22} & w_{32}\end{pmatrix}$$

$$\boxed{\frac{\partial J}{\partial Z_L}=\frac{\partial J}{\partial out}W_{L+1}^T}$$

④

1) 非线性函数 f_L 为 sigmoid

$$Z_L=\frac{1}{1+e^{-h_L}}$$

$$\frac{\partial J}{\partial h_L}=\frac{\partial J}{\partial Z_L}\frac{dz_L}{dh_L}=\frac{\partial J}{\partial Z_L}\frac{e^{-hL}}{(1+e^{-h_L})^2}=\frac{\partial J}{\partial Z_L}\frac{1}{1+e^{-h_L}}\frac{e^{-h_L}}{1+e^{-h_L}}$$

$$=\frac{\partial J}{\partial Z_L}Z_L(1-Z_L)$$

2) 非线性函数 f_L 为 Tanh

$$Z_L=\frac{e^{h_L}-e^{-h_L}}{e^{h_L}+e^{-h_L}}$$

$$\frac{\partial J}{\partial h_L}=\frac{\partial J}{\partial Z_L}\frac{dZ_L}{dh_L}=\frac{\partial J}{\partial Z_L}\frac{4}{(e^{h_L}+e^{-h_L})^2}=\frac{\partial J}{\partial Z_L}[1-(\frac{e^{h_L}-e^{-h_L}}{e^{h_L}+e^{-h_L}})^2]$$

$$=\frac{\partial J}{\partial z_L}[1-{z_L}^2]$$

3) 非线性函数 f_L 为 ReLU

$$Z_L = relu(h_L) = \begin{cases} 0, h_L \leq 0 \\ h_L, h_L > 0 \end{cases}$$

$$\frac{\partial J}{\partial h_L} = \frac{\partial J}{\partial Z_L}\frac{dZ_L}{dh_L} = \begin{cases} 0, h_L \leq 0 \\ \frac{\partial J}{\partial Z_L}, h_L > 0 \end{cases}$$

⑤

$$\frac{\partial J}{\partial \mathrm{W}_L} = Z_{L-1}^T \frac{\partial J}{\partial h_L}$$

$$\frac{\partial J}{\partial \mathrm{Z}_{L-1}} = \frac{\partial J}{\partial h_L} W_L^T$$

因此通过这种逆向的计算，就有

不同算法：

$$\frac{\partial J}{\partial out} \rightarrow \boxed{\begin{cases} \frac{\partial J}{\partial w_{L+1}} = z_L^T \frac{\partial J}{\partial out} \\ \frac{\partial J}{\partial z_L} = \frac{\partial J}{\partial out} w_{L+1}^T \\ (\frac{\partial J}{\partial b_{L+1}})^T = SumRow(\frac{\partial J}{\partial out}) \\ w_{L+1}^{t+1} = w_{L+1}^t - \eta \frac{\partial J}{\partial w_{L+1}} \\ b_{L+1}^{t+1} = b_{L+1}^t - \eta \frac{\partial J}{\partial b_{L+1}} \end{cases}} \rightarrow \boxed{\frac{\partial J}{\partial h_L}} \rightarrow \boxed{\begin{cases} \frac{\partial J}{\partial w_L} = z_{L-1}^T \frac{\partial J}{\partial h_L} \\ \frac{\partial J}{\partial z_{L-1}} = \frac{\partial J}{\partial h_L} w_L^T \\ \vdots \\ \vdots \end{cases}} \cdots$$

5.5.3 案例 1

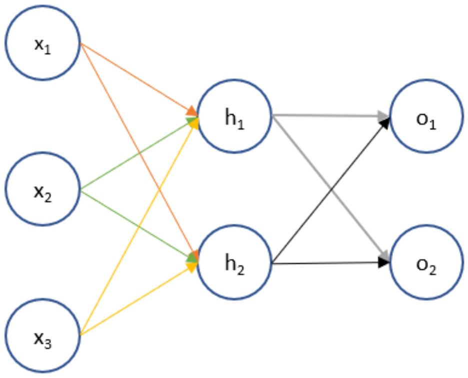


图 48 3 个输入，一个隐藏层（2 个神经元）和一个输出层（2 个神经元）

参数更新将按照以下 5 步进行：

A. 初始化待训练参数的权重；
B. 沿着网络前向传播以得到输出值；
C. 定义误差或损失函数并计算其导数；
D. 沿着网络反向传播以确定误差导数；
E. 使用误差导数和当前值更新参数估计；

Step1：

此问题的输入和目标值为 x1=1，x2=4，x3=5 以及 t1=0.1 及 t2=0.05。下图中已经进行了权重的初始化。为了后续方便表示，定义$\mathrm{W1}=\begin{pmatrix} w1 & w2 \\ w3 & w4 \\ w5 & w6 \end{pmatrix}$，$\mathrm{W2}=\begin{pmatrix} w7 & w8 \\ w9 & w10 \end{pmatrix}$

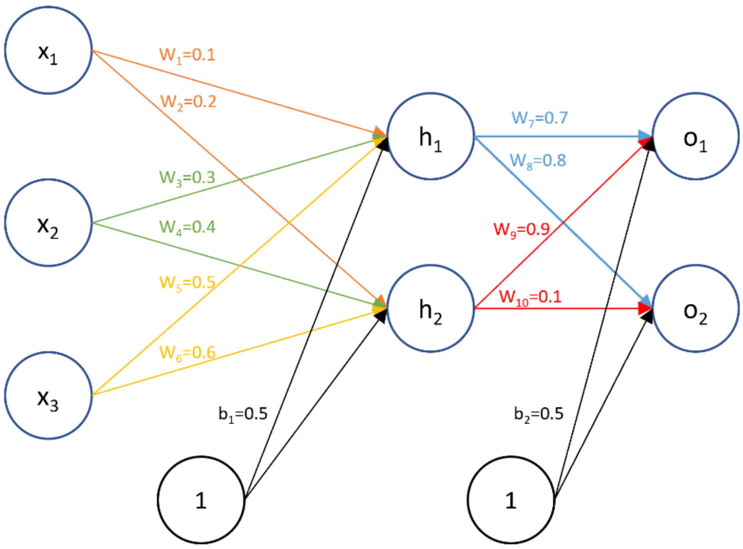


图 49 初始化权重

Step2：

对于输入输出层，使用 z_{h1}，z_{h2}，z_{o1} 和 z_{o2} 表示应用激活函数前的值，h_1，h_2，o_1 和 o_2 表示应用激活函数后的值。

输入到隐藏层

$$(x1 \quad x2 \quad x3)\begin{pmatrix} w1 & w2 \\ w3 & w4 \\ w5 & w6 \end{pmatrix} + \begin{pmatrix} b1 \\ b1 \end{pmatrix}^T = \begin{pmatrix} w1x1 + w3x2 + w5x3 + b1 \\ w2x1 + w4x2 + w6x3 + b1 \end{pmatrix}^T = \begin{pmatrix} z_{h1} \\ z_{h2} \end{pmatrix}^T$$

$$\sigma\left(\begin{pmatrix} z_{h1} \\ z_{h2} \end{pmatrix}^T\right) = \begin{pmatrix} h1 \\ h2 \end{pmatrix}^T$$

隐藏层到输出层

$$\begin{pmatrix} h1 \\ h2 \end{pmatrix}^T \begin{pmatrix} w7 & w8 \\ w9 & w10 \end{pmatrix} + \begin{pmatrix} b2 \\ b2 \end{pmatrix}^T = \begin{pmatrix} w7h1 + w9h2 + b2 \\ w8h1 + w10h2 + b2 \end{pmatrix}^T = \begin{pmatrix} z_{o1} \\ z_{o2} \end{pmatrix}^T$$

$$\sigma\left(\begin{pmatrix} z_{o1} \\ z_{o2} \end{pmatrix}^T\right) = \begin{pmatrix} o1 \\ o2 \end{pmatrix}^T$$

可以通过上式沿着网络前向传播。

$$\begin{pmatrix} z_{h1} \\ z_{h2} \end{pmatrix}^T = \begin{pmatrix} 0.1 \cdot 1 + 0.3 \cdot 4 + 0.5 \cdot 5 + 0.5 \\ 0.2 \cdot 1 + 0.4 \cdot 4 + 0.6 \cdot 5 + 0.5 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 4.3 \\ 5.3 \end{pmatrix}^T$$

$$\begin{pmatrix} h1 \\ h2 \end{pmatrix}^T = \begin{pmatrix} \sigma(4.3) \\ \sigma(5.3) \end{pmatrix}^T = \begin{pmatrix} 0.9866 \\ 0.9950 \end{pmatrix}^T$$

$$\begin{pmatrix} z_{o1} \\ z_{o2} \end{pmatrix}^T = \begin{pmatrix} 0.7 \cdot 0.9866 + 0.9 \cdot 0.9950 + 0.5 \\ 0.8 \cdot 0.9866 + 0.1 \cdot 0.9950 + 0.5 \end{pmatrix}^T = \begin{pmatrix} 2.0862 \\ 1.3888 \end{pmatrix}^T$$

$$\begin{pmatrix} o1 \\ o2 \end{pmatrix}^T = \begin{pmatrix} \sigma(2.0862) \\ \sigma(1.3888) \end{pmatrix}^T = \begin{pmatrix} 0.8896 \\ 0.8004 \end{pmatrix}^T$$

Step3：

根据目标值和前向传播中最后一层的结果计算误差平方和。

$$\mathrm{E} = \frac{1}{2}\begin{pmatrix} (o1 - t1)^2 \\ (o2 - t2)^2 \end{pmatrix}^{\mathrm{T}}$$

$$\frac{\mathrm{dE}}{\mathrm{do}} = \begin{pmatrix} o1 - t1 \\ o2 - t2 \end{pmatrix}^{\mathrm{T}}$$

Step4：

现在将按照链式法则，沿着网络反向传播以计算误差关于参数的导数。

根据

$$w_7 h_1 + w_9 h_2 + b_2 = z_{o_1}$$

$$w_8 h_1 + w_{10} h_2 + b_2 = z_{o_2}$$

得

$$\frac{dz_{o_1}}{dw_7} = h_1, \frac{dz_{o_2}}{dw_8} = h_1, \frac{dz_{o_1}}{dw_9} = h_2, \frac{dz_{o_2}}{dw_{10}} = h_2$$

$$\frac{dz_{o_1}}{db_2} = 1, \text{and} \frac{dz_{o_2}}{db_2} = 1$$

$$\frac{\mathrm{dE}}{\mathrm{dW2}} = \begin{pmatrix} \frac{\mathrm{dE}}{dw7} & \frac{dE}{dw8} \\ \frac{\mathrm{dE}}{dw9} & \frac{dE}{dw10} \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{dE}}{do1}\frac{do1}{dzo1}\frac{dzo1}{dw7} & \frac{\mathrm{dE}}{do2}\frac{do2}{dzo2}\frac{dzo2}{dw8} \\ \frac{\mathrm{dE}}{do1}\frac{do1}{dzo1}\frac{dzo1}{dw9} & \frac{dE}{do2}\frac{do2}{dzo2}\frac{dzo2}{dw10} \end{pmatrix}$$

$$= \begin{pmatrix} (o1 - t1)\big(o1(1 - o1)\big)h1 & (o2 - t2)\big(o2(1 - o2)\big)h1 \\ (o1 - t1)\big(o1(1 - o1)\big)h2 & (o2 - t2)\big(o2(1 - o2)\big)h2 \end{pmatrix}$$

$$= \begin{pmatrix} (0.8896 - 0.1)\big(0.8896(1 - 0.8896)\big)0.9866 & (0.8004 - 0.05)\big(0.8004(1 - 0.8004)\big)0.9866 \\ (0.8896 - 0.1)\big(0.8896(1 - 0.8896)\big)0.9950 & (0.8004 - 0.05)\big(0.8004(1 - 0.8004)\big)0.9950 \end{pmatrix}$$

$$= \begin{pmatrix} 0.0765 & 0.1183 \\ 0.0772 & 0.1193 \end{pmatrix}$$

参数 b2 的误差导数稍微复杂一些，因为 b2 的改变会通过 o1 和 o2 影响误差。

$$\frac{dE}{db_2} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{db_2} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{db_2}$$

$$\frac{dE}{db_2} = (0.7896)(0.0983)(1) + (0.7504)(0.1598)(1)$$

$$\frac{dE}{db_2} = 0.1975$$

至此已经计算出误差关于 w7,w8,w9,w10 和 b2 的导数，接下来反向传播下一层来计

算输入层到隐藏层之间的参数的误差导数。

$$\frac{dE}{dw_1} = \frac{dE}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{dw_1}$$

$$\frac{dE}{dh_1} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dh_1} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{dh_1}$$

$$\frac{dE}{dh_1} = (0.7896)(0.0983)(0.7) + (0.7504)(0.1598)(0.8) = 0.1502$$

$$\frac{dE}{dw_1} = (0.1502)(0.0132)(1) = 0.0020$$

$$\frac{dE}{dw_3} = \frac{dE}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{dw_3}$$

$$\frac{dE}{dw_3} = (0.1502)(0.0132)(4) = 0.0079$$

$$\frac{dE}{dw_5} = \frac{dE}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{dw_5}$$

$$\frac{dE}{dw_5} = (0.1502)(0.0132)(5) = 0.0099$$

$$\frac{dE}{dw_2} = \frac{dE}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{dw_2}$$

$$\frac{dE}{dh_2} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dh_2} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{dh_2}$$

$$\frac{dE}{dh_2} = (0.7896)(0.0983)(0.9) + (0.7504)(0.1598)(0.1) = 0.0818$$

$$\frac{dE}{dw_2} = (0.0818)(0.0049)(1) = 0.0004$$

$$\frac{dE}{dw_4} = \frac{dE}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{dw_4}$$

$$\frac{dE}{dw_4} = (0.0818)(0.0049)(4) = 0.0016$$

$$\frac{dE}{dw_6} = \frac{dE}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{dw_6}$$

$$\frac{dE}{dw_6} = (0.0818)(0.0049)(5) = 0.0020$$

$\frac{dE}{db_1} = \frac{dE}{do_1}\frac{do_1}{dz_{o_1}}\frac{dz_{o_1}}{dh_1}\frac{dh_1}{dz_{h_1}}\frac{dz_{h_1}}{db_1} + \frac{dE}{do_2}\frac{do_2}{dz_{o_2}}\frac{dz_{o_2}}{dh_2}\frac{dh_2}{dz_{h_2}}\frac{dz_{h_2}}{db_1}$

$\frac{dE}{db_1} = (0.7896)(0.0983)(0.7)(0.0132)(1)+(0.7504)(0.1598)(0.1)(0.0049)(1) = 0.0008$

计算得到所有的误差导数，在第一次跌打反向传播后进行参数更新，设置学习率为0.01：

$$w_1 := w_1 - \alpha\frac{dE}{dw_1} = 0.1 - (0.01)(0.0020) = 0.1000$$

$$w_2 := w_2 - \alpha\frac{dE}{dw_2} = 0.2 - (0.01)(0.0004) = 0.2000$$

$$w_3 := w_3 - \alpha\frac{dE}{dw_3} = 0.3 - (0.01)(0.0079) = 0.2999$$

$$w_4 := w_4 - \alpha\frac{dE}{dw_4} = 0.4 - (0.01)(0.0016) = 0.4000$$

$$w_5 := w_5 - \alpha\frac{dE}{dw_5} = 0.5 - (0.01)(0.0099) = 0.4999$$

$$w_6 := w_6 - \alpha\frac{dE}{dw_6} = 0.6 - (0.01)(0.0020) = 0.6000$$

$$w_7 := w_7 - \alpha\frac{dE}{dw_7} = 0.7 - (0.01)(0.0765) = 0.6992$$

$$w_8 := w_8 - \alpha\frac{dE}{dw_8} = 0.8 - (0.01)(0.1183) = 0.7988$$

$$w_9 := w_9 - \alpha\frac{dE}{dw_9} = 0.9 - (0.01)(0.0772) = 0.8992$$

$$w_{10} := w_{10} - \alpha\frac{dE}{dw_{10}} = 0.1 - (0.01)(0.1193) = 0.0988$$

$$b_1 := b_1 - \alpha\frac{dE}{db_1} = 0.5 - (0.01)(0.0008) = 0.5000$$

$$b_2 := b_2 - \alpha\frac{dE}{db_2} = 0.5 - (0.01)(0.1975) = 0.4980$$

重复此过程直至误差小于一定值或参数估计收敛。

5.5.4 案例 2

考虑如下网络：

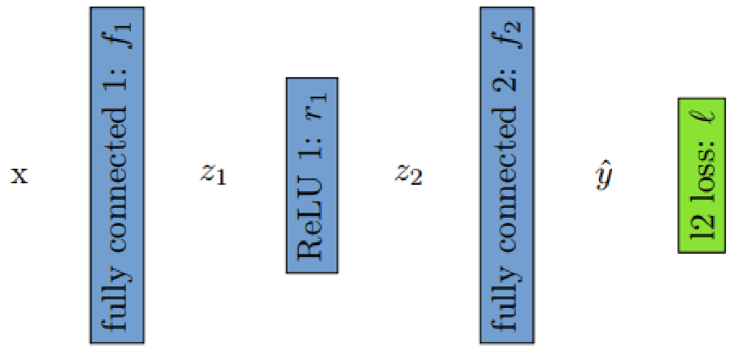


图 50 两个全链接层和一个 ReLU 层

对于输入 $x \in R^2$ 和连续标签 $y \in R$，网络定义如下：

$$z_1 = f_1(x) = \begin{bmatrix} 1 & -2 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$z_2 = r_1(z_1) = \max(z_1, 0)$$

$$\hat{y} = f_2(x) = \begin{bmatrix} 1 & -1 \end{bmatrix} z_2$$

$$\ell(z_2) = \frac{1}{2}\|\hat{y} - y\|^2$$

对于以下输入，计算前向传播和网络损失：

a) $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, y = 2$

$$z1 = \begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$z2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\hat{y} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1$$

$$l = \frac{1}{2}$$

b) $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, y = 0$

$$z1 = \begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

$$z2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\hat{y} = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -1$$

$$l = \frac{1}{2}$$

c) $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, y = -2$

$$z1 = \begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$z2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\hat{y} = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -1$$

$$l = \frac{1}{2}$$

对于以上输入，使用反向传播计算 $\frac{d}{dx}\ell(\hat{y})$

a)

$$\frac{\mathrm{d}l(\hat{y})}{\mathrm{dx}} = \frac{\mathrm{d}l(\hat{y})}{\mathrm{d}\hat{y}}\frac{d\hat{y}}{dz2}\frac{dz2}{dz1}\frac{dz1}{dx}$$

$$= -1 \cdot \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} -1 & 2 \end{pmatrix}$$

b)

$$\frac{\mathrm{d}l(\hat{y})}{\mathrm{dx}} = \frac{\mathrm{d}l(\hat{y})}{\mathrm{d}\hat{y}}\frac{d\hat{y}}{dz2}\frac{dz2}{dz1}\frac{dz1}{dx}$$

$$= -1 \cdot \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \end{pmatrix}$$

c)

$$\frac{\mathrm{d}l(\hat{\mathrm{y}})}{\mathrm{dx}} = \frac{\mathrm{d}l(\hat{\mathrm{y}})}{\mathrm{d}\hat{\mathrm{y}}}\frac{d\hat{\mathrm{y}}}{dz2}\frac{dz2}{dz1}\frac{dz1}{dx}$$

$$= 1 \cdot (1 \quad -1)\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix} = (0 \quad 2)$$