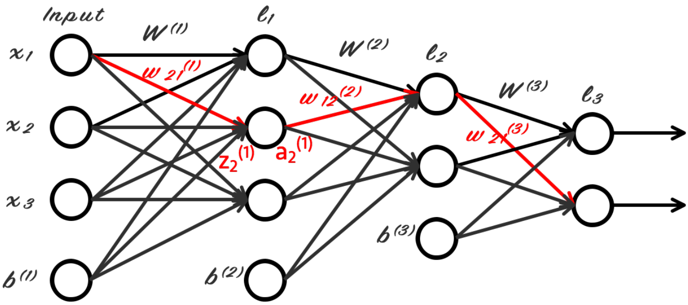
Deep Learning Technology and Application

Ge Li

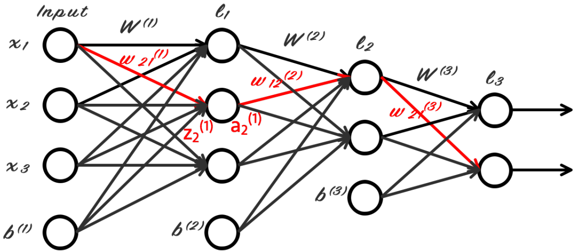
Peking University

Table of contents

前向传播的符号体系



前向传播计算

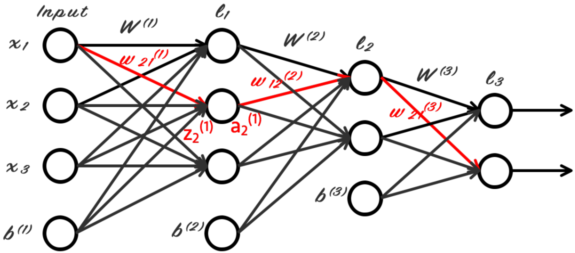


$$z_1^{(1)} = w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3 + b_1^{(1)} \qquad a_1^{(1)} = f(z_1^{(1)})$$

$$z_2^{(1)} = w_{21}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{23}^{(1)} x_3 + b_2^{(1)} \qquad a_2^{(1)} = f(z_2^{(1)})$$

$$z_3^{(1)} = w_{31}^{(1)} x_1 + w_{32}^{(1)} x_2 + w_{33}^{(1)} x_3 + b_3^{(1)} \qquad a_3^{(1)} = f(z_3^{(1)})$$

前向传播计算

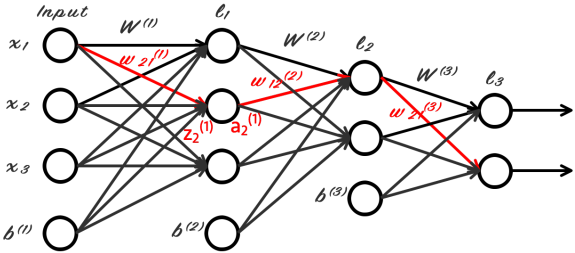


$$z^{(1)} = \begin{bmatrix} z_1^{(1)} \\ z_2^{(1)} \\ z_3^{(1)} \end{bmatrix} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} & w_{13}^{(1)} \\ w_{21}^{(1)} & w_{21}^{(1)} & w_{23}^{(1)} \\ w_{31}^{(1)} & w_{32}^{(1)} & w_{33}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \end{bmatrix} = W^{(1)}X + b^{(1)}$$

$$a^{(1)} = f(z^{(1)}) = f(W^{(1)}X + b^{(1)})$$

$$a^{(2)} = f(z^{(2)}) = f(W^{(2)}a^{(1)} + b^{(2)})$$

前向传播计算

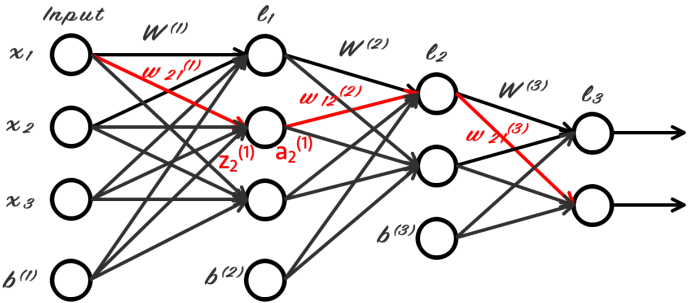


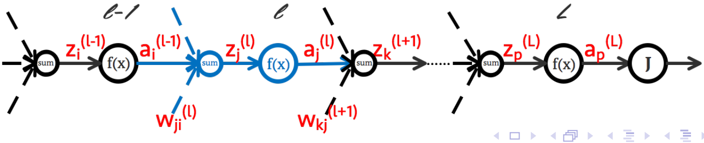
$$z^{(2)} = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \end{bmatrix} = \begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{21}^{(2)} & w_{23}^{(2)} \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \end{bmatrix} + \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \\ b_3^{(2)} \end{bmatrix} = W^{(2)} a^{(1)} + b^{(2)}$$

总之:

$$z^l = W^{(l)} a^{(l-1)} + b^{(l)} \qquad a^{(l)} = f(z^{(l)})$$

反向传播符号体系


Input
$W^{(1)}$
l_1
$W^{(2)}$
l_2
$W^{(3)}$
l_3
x_1
x_2
x_3
$b^{(1)}$
$w_{21}^{(1)}$
$w_{12}^{(2)}$
$w_{21}^{(3)}$
$z_2^{(1)}$
$a_2^{(1)}$
$b^{(2)}$
$b^{(3)}$


$l-1$
l
L
$z_i^{(l-1)}$
$a_i^{(l-1)}$
$z_j^{(l)}$
$a_j^{(l)}$
$z_k^{(l+1)}$
$z_p^{(L)}$
$a_p^{(L)}$
sum
f(x)
J
$w_{ji}^{(l)}$
$w_{kj}^{(l+1)}$

预备知识-多元复合函数求导

由于接下来的计算中要用到多元符合函数的求导，下面我们先来回顾一下“多元复合函数的求导”的方法：

多元复合函数求导-1

设：$z = f(y_1, y_2, ..., y_m)$，其中：$(y_1, y_2, ..., y_m) \in D_f$ 为区域 $D_f \subset R^m$ 上的 m 元函数。又设：

$$\begin{aligned} g : D_g &\to R^m, \\ (x_1, x_2, ..., x_n) &\mapsto (y_1, y_2, ..., y_m) \end{aligned} \tag{1}$$

为区域 $D_g \subset R^n$ 上的 n 元 m 维向量值函数，那么，对于复合函数：

$$z = f \circ g = f[y_1(x_1, x_2, ..., x_n), y_2(x_1, x_2, ..., x_n), ..., y_m(x_1, x_2, ..., x_n)]$$

其中：$(x_1, x_2, ..., x_n) \in D_g$

若 g 在 $x^0 \in D_g$ 点可导，即 $y_1, y_2, ..., y_n$ 在 x^0 点可偏导，且 f 在 $y^0 = g(x^0)$ 点可微，则：

多元复合函数求导-2

$$\frac{\partial z}{\partial x}(x^0) = \left(\frac{\partial z}{\partial x_1}, \frac{\partial z}{\partial x_2}, \cdots \frac{\partial z}{\partial x_i}, \dots, \frac{\partial z}{\partial x_n}\right)_{x=x^0}$$

其中:

$$\frac{\partial z}{\partial x_i}(x^0) = \sum_{j=1}^{m} \frac{\partial z}{\partial y_j}(y^0)\frac{\partial y_j}{\partial x_i}(x^0)$$

即:

$$\frac{\partial z}{\partial x}(x^0) = \left(\frac{\partial z}{\partial y_1}, \frac{\partial z}{\partial y_2}, \dots, \frac{\partial z}{\partial y_m}\right)_{y=y^0} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}_{x=x^0}$$

多元复合函数求导-3

若 g 处处可导，即 $y_1, y_2, \dots, y_n$ 处处可偏导，且 f 处处可微，则：

$$\frac{\partial z}{\partial x} = \left(\frac{\partial z}{\partial x_1}, \frac{\partial z}{\partial x_2}, \cdots \frac{\partial z}{\partial x_i}, \dots, \frac{\partial z}{\partial x_n}\right)$$

其中：

$$\frac{\partial z}{\partial x_i} = \sum_{j=1}^{m} \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i}$$

即：

$$\frac{\partial z}{\partial x} = \left(\frac{\partial z}{\partial y_1}, \frac{\partial z}{\partial y_2}, \dots, \frac{\partial z}{\partial y_m}\right) \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

(此矩阵即 Jacobian 矩阵)

一介全微分的形式不变性

对于多元函数 $z=f(y)$，其中 $y=(y_1,y_2,\dots,y_m)^\top$。当 y 为自变量时，一介全微分形式为：

$$\mathrm{d}z=f'(y)\,\mathrm{d}y$$

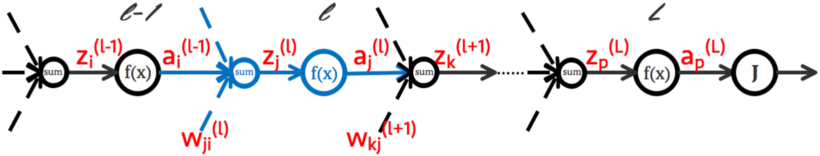
而当 y 为中间变量 $y=g(x)(x=(x_1,x_2,\dots,x_n)^\top)$ 时，$\mathrm{d}y=g'(x)\,\mathrm{d}x$。由链式规则，得：

$$\mathrm{d}z=(f\circ g)'(x)\,\mathrm{d}x=f'(y)g'(x)\,\mathrm{d}x=f'(y)(g'(x)\,\mathrm{d}x)=f'(y)\,\mathrm{d}y$$

注意符号：

$$\frac{\mathrm{d}z}{\mathrm{d}y}=f'(y);\qquad \frac{\mathrm{d}z}{\mathrm{d}x}=(f\circ g)'(x)=f'(y)g'(x)$$

反向传播算法



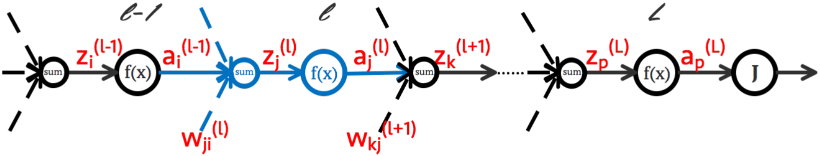
$$z^l = W^{(l)} a^{(l-1)} + b^{(l)} \qquad\qquad a^{(l)} = f(z^{(l)})$$

由梯度下降方法，可知，需要对每个权重权值 $w_{ij}^{(l)}$，求取：

$$w_{ji}^{(l)} = w_{ji}^{(l)} - \alpha \frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} \qquad b_i^{(l)} = b_i^{(l)} - \alpha \frac{\partial J(W,b)}{\partial b_i^{(l)}}$$

其中，关键是如何求取：$\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}}$ 和 $\frac{\partial J(W,b)}{\partial b_i^{(l)}}$

反向传播算法



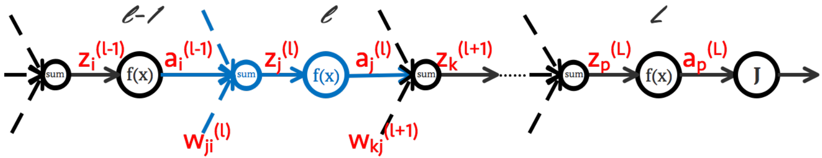
由前向传播过程可知：$z_j^{(l)} = \sum_{i=1}^{n_l} w_{ji}^{(l)} a_i^{(l-1)} + b_i^{(l)}$ 可知：

$$\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} = \frac{\partial J(W,b)}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial w_{ji}^{(l)}} = \frac{\partial J(W,b)}{\partial z_j^{(l)}} a_i^{(l-1)}$$

$$\frac{\partial J(W,b)}{\partial b_i^{(l)}} = \frac{\partial J(W,b)}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial b_i^{(l)}} = \frac{\partial J(W,b)}{\partial z_j^{(l)}}$$

到此为止，关键是如何求取 $\frac{\partial J(W,b)}{\partial z_j^{(l)}}$

反向传播算法

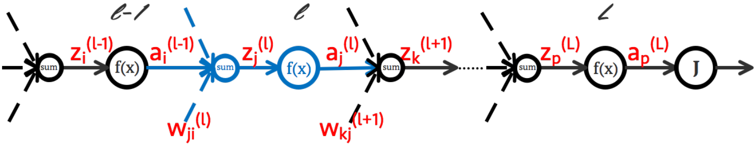


设： $\delta_j^{(l)} = \frac{\partial J(W,b)}{\partial z_j^{(l)}}$

因为： $z_k^{(l+1)} = \sum_{j=1}^{n_{l+1}} w_{kj}^{(l+1)} a_j^{(l)} + b^{(l+1)}$

所以，可以选择从 $z_k^{(l+1)}$ 开始进行对 $z_j^{(l)}$ 进行求导计算：

反向传播算法推导



$$
\begin{aligned}
\delta_j^{(l)} = \frac{\partial J(W,b)}{\partial z_j^{(l)}} &= \sum_{k=1}^{n_{l+1}} \frac{\partial J(W,b)}{\partial z_k^{(l+1)}} \frac{\partial z_k^{(l+1)}}{\partial a_j^{(l)}} \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}} \\
&= \sum_{k=1}^{n_{l+1}} \frac{\partial J(W,b)}{\partial z_k^{(l+1)}} w_{kj}^{(l+1)} f'(z_j^{(l)}) \\
&= \sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{kj}^{(l+1)} f'(z_j^{(l)})
\end{aligned} \tag{2}
$$

反向传播算法推导

对于最后一层:

$$\delta_p^{(L)} = \frac{\partial J(W,b)}{\partial z_p^{(L)}} = \frac{\partial J(W,b)}{\partial a_p^{(L)}} * \frac{\partial a_p^{(L)}}{\partial z_p^{(L)}} = \frac{\partial J(W,b)}{\partial a_p^{(L)}} * f'(z_p^{(L)})$$

并且:

$$\frac{\partial J(W,b)}{\partial w_{pq}^{(L)}} = \frac{\partial J(W,b)}{\partial z_p^{(L)}} a_p^{(L-1)} = \delta_p^{(L)} a_p^{(L-1)}$$

$$\frac{\partial J(W,b)}{\partial b_q^{(L)}} = \frac{\partial J(W,b)}{\partial z_p^{(L)}} = \delta_p^{(L)}$$

反向传播算法推导

小结一下，因为：

$$\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} = \frac{\partial J(W,b)}{\partial z_j^{(l)}} a_i^{(l-1)} \qquad \frac{\partial J(W,b)}{\partial b_i^{(l)}} = \frac{\partial J(W,b)}{\partial z_j^{(l)}}$$

又因为 (上文推导结果)：

$$\frac{\partial J(W,b)}{\partial z_j^{(l)}} = \delta_j^{(l)} = \sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{kj}^{(l+1)} f'(z_j^{(l)})$$

从而得到：

$$\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} a_i^{(l-1)} \qquad \frac{\partial J(W,b)}{\partial b_i^{(l)}} = \delta_j^{(l)}$$

反向传播算法总结

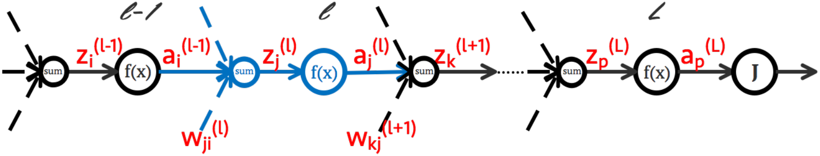
总结一下：

$$\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} a_i^{(l-1)} = \left(\sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{kj}^{(l+1)} f'(z_j^{(l)}) \right) a_i^{(l-1)}$$

$$\frac{\partial J(W,b)}{\partial b_i^{(l)}} = \delta_j^{(l)} = \sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{kj}^{(l+1)} f'(z_j^{(l)})$$

反向传播计算流程

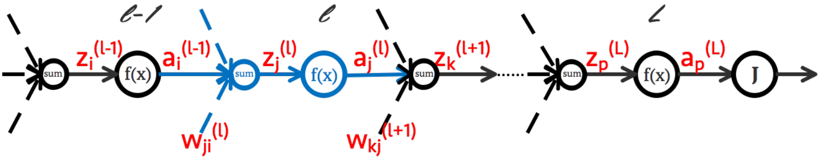
Step-1：依据前向传播算法求解每一层的激活值：



$$z^l = W^{(l)} a^{(l-1)} + b^{(l)} \qquad a^{(l)} = f(z^{(l)})$$

反向传播计算流程

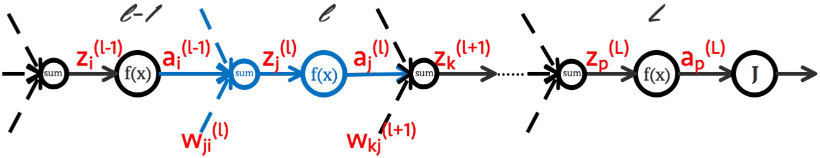
Step-2：计算出最后一层 (L 层) 的每个神经元的 $\delta_p^{(L)}$：



$$\delta_p^{(L)} = \frac{\partial J(W,b)}{\partial a_p^{(L)}} * f'(z_p^{(L)})$$

反向传播计算流程

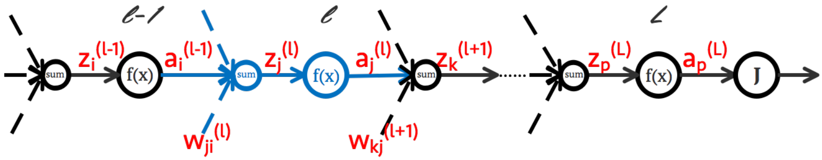
Step-3：由后向前，依次计算出各层（l 层）各个神经元的 $\delta_j^{(l)}$



$$\delta_j^{(l)} = \sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{kj}^{(l+1)} f'(z_j^{(l)})$$

反向传播计算流程

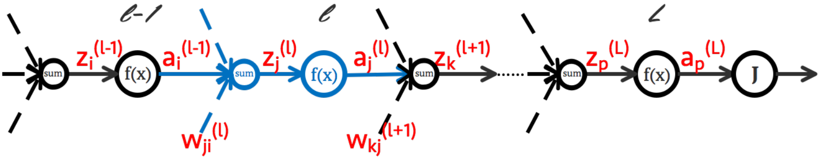
Step-4：计算出各层（l 层）的各个权重（$w_{ji}^{(l)}$）的梯度 $\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}}$ 及各个偏置（$b_i^{(l)}$）的梯度 $\frac{\partial J(W,b)}{\partial b_i^{(l)}}$：



$$\frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} a_i^{(l-1)} \qquad \frac{\partial J(W,b)}{\partial b_i^{(l)}} = \delta_j^{(l)}$$

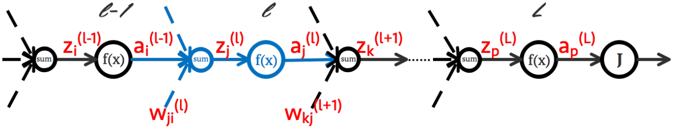
反向传播计算流程

Step-5：对各层（l 层）的各个权重（$w_{ji}^{(l)}$）及各个偏置（$b_i^{(l)}$）进行更新，直到代价函数 $J(W,b)$ 足够小：



$$w_{ji}^{(l)} = w_{ji}^{(l)} - \alpha \frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} \qquad b_i^{(l)} = b_i^{(l)} - \alpha \frac{\partial J(W,b)}{\partial b_i^{(l)}}$$

反向传播核心算式



$$
\begin{aligned}
w_{ji}^{(l)} &= w_{ji}^{(l)} - \alpha \frac{\partial J(W,b)}{\partial w_{ji}^{(l)}} \\
&= w_{ji}^{(l)} - \alpha \frac{\partial J(W,b)}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial w_{ji}^{(l)}} = w_{ji}^{(l)} - \alpha \frac{\partial J(W,b)}{\partial z_j^{(l)}} a_i^{(l-1)} \\
&= w_{ji}^{(l)} - \alpha \delta_j^{(l)} a_i^{(l-1)} \\
&= w_{ji}^{(l)} - \alpha \left(\sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{kj}^{(l+1)} f'(z_j^{(l)}) \right) a_i^{(l-1)}
\end{aligned}
$$

Thanks.