

如图所示三层神经网络：输入层  $Layer_1$ 、隐含层  $Layer_2$ 、输出层  $Layer_3$ 。输入层包括三个神经元，即输入为  $\mathbf{x} = \{x_1, x_2, x_3\}$ ，样本数据集为  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ ，目标输出为  $o_1^T$  和  $o_2^T$ 。  $\omega_{ij}^{L_k}$  为第  $k-1$  层的第  $i$  个神经元到第  $k$  层的第  $j$  个神经元之间的权重。  $\mathbf{b}^{L_k} = \{b_i^{L_k}\}$  为第  $k$  层神经元偏移量集合，  $b_j^{L_k}$  为第  $k$  层第  $j$  个神经元的偏移量。  $net_i^{L_k}$  表示第  $k$  层网络的第  $i$  个神经元的输入，  $o_i^{L_k}$  表示第  $k$  层网络的第  $i$  个神经元的输出，  $f_i^{L_k}(\cdot)$  表示第  $k$  层网络的第  $i$  个神经元的激活函数。

假设激活函数为 sigmoid 函数：

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

其导数为：

$$f'(x) = f(x)(1 - f(x)) \quad (2)$$

对于  $Layer_2$  层第  $j$  个神经元的输出：

$$net_j^{L_2} = \sum_{i=1}^3 \omega_{ij}^{L_1} \times x_i + b_j^{L_2} \quad (3)$$

对于  $Layer_3$  层的第  $j$  个神经元的输出：

$$net_j^{L_3} = \sum_{i=1}^2 \omega_{ij}^{L_2} \times o_i^{L_2} + b_j^{L_3} \quad (4)$$

第  $L_i$  层的第  $j$  个神经元的输出：

$$o_j^{L_i} = f_j^{L_i}(net_j^{L_i}) = \frac{1}{1 + e^{-net_j^{L_i}}} \quad (5)$$

输出的总误差为：

$$E_{\text{Total}} = \frac{1}{2} \sum_{j=1}^2 (o_j^T - o_j^{L_3})^2 = \frac{1}{2} ((o_1^T - o_1^{L_3})^2 + (o_2^T - o_2^{L_3})^2) \quad (6)$$

是  $o_1^{L_3}$  和  $o_2^{L_3}$  的函数

## 1 $Layer_2$ 与 $Layer_3$ 间参数调整

调整  $Layer_2$  与  $Layer_3$  间的权重  $\omega_{ij}^{L_2}$

$$\frac{\partial E_{\text{Total}}}{\partial \omega_{ij}^{L_2}} = \frac{\partial E_{\text{Total}}}{\partial o_j^{L_3}} \cdot \frac{\partial o_j^{L_3}}{\partial net_j^{L_3}} \cdot \frac{\partial net_j^{L_3}}{\partial \omega_{ij}^{L_2}} \quad (7)$$

$$= -(o_j^T - o_j^{L_3}) \cdot f_j^{L_3}(net_j^{L_3})(1 - f_j^{L_3}(net_j^{L_3})) \cdot o_i^{L_2} \quad (8)$$

$$= -(o_j^T - o_j^{L_3}) \cdot o_j^{L_3}(1 - o_j^{L_3}) \cdot o_i^{L_2} \quad (9)$$

$$(10)$$

令  $\delta_{ij}^{L_2}$  表示  $Layer_2$  层第  $i$  个神经元与  $Layer_3$  层第  $j$  个神经元间权重的梯度项，则

$$\delta_{ij}^{L_2} = \frac{\partial E_{\text{Total}}}{\partial o_j^{L_3}} \cdot \frac{\partial o_j^{L_3}}{\partial \text{net}_j^{L_3}} \quad (11)$$

$$= -(o_j^T - o_j^{L_3}) \cdot o_j^{L_3} (1 - o_j^{L_3}) \quad (12)$$

可以看出， $Layer_2$  层第  $i$  个神经元与  $Layer_3$  层第  $j$  个神经元间权重的梯度项是与  $i$  无关的，故连接至  $Layer_3$  层第  $j$  个神经元相对应的权重的梯度向均为  $\delta_{ij}^{L_2}$ ，用  $\delta_{\cdot j}^{L_2}$  表示，

$$\delta_{\cdot j}^{L_2} = \delta_{ij}^{L_2} = -(o_j^T - o_j^{L_3}) \cdot o_j^{L_3} (1 - o_j^{L_3}) \quad (13)$$

则整体误差  $E_{\text{Total}}$  对  $\omega_{ij}^{L_2}$  的偏导数公式写作：

$$\frac{\partial E_{\text{Total}}}{\partial \omega_{ij}^{L_2}} = \delta_{\cdot j}^{L_2} \cdot o_i^{L_2} \quad (14)$$

则权重  $\omega_{ij}^{L_2}$  的学习公式为：

$$\hat{\omega}_{ij}^{L_2} = \omega_{ij}^{L_2} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial \omega_{ij}^{L_2}} = \omega_{ij}^{L_2} - \eta \cdot \delta_{\cdot j}^{L_2} \cdot o_i^{L_2} \quad (15)$$

$\eta$  为学习速率

$Layer_3$  的第  $j$  个神经元的偏移量的梯度为

$$\frac{\partial E_{\text{Total}}}{\partial b_j^{L_3}} = \frac{\partial E_{\text{Total}}}{\partial o_j^{L_3}} \cdot \frac{\partial o_j^{L_3}}{\partial \text{net}_j^{L_3}} \cdot \frac{\partial \text{net}_j^{L_3}}{\partial b_j^{L_3}} \quad (16)$$

$$= -(o_j^T - o_j^{L_3}) \cdot f_j^{L_3}(\text{net}_j^{L_3}) (1 - f_j^{L_3}(\text{net}_j^{L_3})) \cdot 1 \quad (17)$$

$$= -(o_j^T - o_j^{L_3}) \cdot o_j^{L_3} (1 - o_j^{L_3}) \quad (18)$$

$$= \delta_{\cdot j}^{L_2} \quad (19)$$

偏移  $b_j^{L_3}$  的学习公式为：

$$\hat{b}_j^{L_3} = b_j^{L_3} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial b_j^{L_3}} = b_j^{L_3} - \eta \cdot \delta_{\cdot j}^{L_2} \quad (20)$$

例如，对于权重  $\omega_{11}^{L_2}$  的更新学习公式为：

$$\hat{\omega}_{11}^{L_2} = \omega_{11}^{L_2} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial \omega_{11}^{L_2}} \quad (21)$$

$$= \omega_{11}^{L_2} - \eta \cdot \delta_{11}^{L_2} \cdot o_1^{L_2} \quad (22)$$

$$= \omega_{11}^{L_2} - \eta \cdot (-(o_1^T - o_1^{L_3}) \cdot o_1^{L_3} (1 - o_1^{L_3})) \cdot o_1^{L_2} \quad (23)$$

$$= \omega_{11}^{L_2} + \eta \cdot (o_1^T - o_1^{L_3}) \cdot o_1^{L_3} (1 - o_1^{L_3}) \cdot o_1^{L_2} \quad (24)$$

例如，对于偏移  $b_1^{L_3}$  的更新学习公式为：

$$\hat{b}_1^{L_3} = b_1^{L_3} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial b_1^{L_3}} \quad (25)$$

$$= b_1^{L_3} + \eta \cdot (o_1^T - o_1^{L_3}) \cdot o_1^{L_3} (1 - o_1^{L_3}) \quad (26)$$

## 2 $Layer_1$ 与 $Layer_2$ 间参数调整

调整  $Layer_1$  与  $Layer_2$  间的权重  $\omega_{ij}^{L_1}$

$$\frac{\partial E_{\text{Total}}}{\partial \omega_{ij}^{L_1}} = \left( \sum_{k=1}^2 \frac{\partial E_{\text{Total}}}{\partial o_k^{L_3}} \frac{\partial o_k^{L_3}}{\partial net_k^{L_3}} \frac{\partial net_k^{L_3}}{\partial o_j^{L_2}} \right) \cdot \frac{\partial o_j^{L_2}}{\partial net_j^{L_2}} \cdot \frac{\partial net_j^{L_2}}{\partial \omega_{ij}^{L_1}} \quad (27)$$

$$= \left( \sum_{k=1}^2 \delta_{\cdot k}^{\text{Layer}_2} \omega_{3k}^{L_2} \right) \cdot f_j^{L_2}(net_j^{L_2})(1 - f_j^{L_2}(net_j^{L_2})) \cdot x_i \quad (28)$$

$$= \left( \sum_{k=1}^2 \delta_{\cdot k}^{\text{Layer}_2} \omega_{3k}^{L_2} \right) \cdot o_j^{L_2}(1 - o_j^{L_2}) \cdot x_i \quad (29)$$

$$(30)$$

令  $\delta_{ij}^{L_1}$  表示  $Layer_1$  层第  $i$  个神经元与  $Layer_2$  层第  $j$  个神经元间权重的梯度项，则

$$\delta_{ij}^{L_1} = \left( \sum_{k=1}^2 \frac{\partial E_{\text{Total}}}{\partial o_k^{L_3}} \frac{\partial o_k^{L_3}}{\partial net_k^{L_3}} \frac{\partial net_k^{L_3}}{\partial o_j^{L_2}} \right) \cdot \frac{\partial o_j^{L_2}}{\partial net_j^{L_2}} \quad (31)$$

$$= \left( \sum_{k=1}^2 \delta_{\cdot k}^{\text{Layer}_2} \omega_{3k}^{L_2} \right) \cdot o_j^{L_2}(1 - o_j^{L_2}) \quad (32)$$

可以看出， $Layer_1$  层第  $i$  个神经元与  $Layer_2$  层第  $j$  个神经元间权重的梯度项是与  $i$  无关的，故连接至  $Layer_2$  层第  $j$  个神经元相对应的权重的梯度向均为  $\delta_{ij}^{L_1}$ ，用  $\delta_{\cdot j}^{L_1}$  表示，

$$\delta_{\cdot j}^{L_1} = \delta_{ij}^{L_1} = \left( \sum_{k=1}^2 \delta_{\cdot k}^{\text{Layer}_2} \omega_{3k}^{L_2} \right) \cdot o_j^{L_2}(1 - o_j^{L_2}) \quad (33)$$

则整体误差  $E_{\text{Total}}$  对  $\omega_{ij}^{L_1}$  的偏导数公式写作：

$$\frac{\partial E_{\text{Total}}}{\partial \omega_{ij}^{L_1}} = \delta_{\cdot j}^{L_1} \cdot x_i \quad (34)$$

则权重  $\omega_{ij}^{L_1}$  的学习公式为：

$$\hat{\omega}_{ij}^{L_1} = \omega_{ij}^{L_1} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial \omega_{ij}^{L_1}} = \omega_{ij}^{L_1} - \eta \cdot \delta_{\cdot j}^{L_1} \cdot x_i \quad (35)$$

$$\frac{\partial E_{\text{Total}}}{\partial b_j^{L_2}} = \left( \sum_{k=1}^2 \frac{\partial E_{\text{Total}}}{\partial o_k^{L_3}} \frac{\partial o_k^{L_3}}{\partial net_k^{L_3}} \frac{\partial net_k^{L_3}}{\partial o_j^{L_2}} \right) \cdot \frac{\partial o_j^{L_2}}{\partial net_j^{L_2}} \cdot \frac{\partial net_j^{L_2}}{\partial b_j^{L_2}} \quad (36)$$

$$= \left( \sum_{k=1}^2 \delta_{\cdot k}^{\text{Layer}_2} \omega_{3k}^{L_2} \right) \cdot f_j^{L_2}(net_j^{L_2})(1 - f_j^{L_2}(net_j^{L_2})) \cdot 1 \quad (37)$$

$$= \left( \sum_{k=1}^2 \delta_{\cdot k}^{\text{Layer}_2} \omega_{3k}^{L_2} \right) \cdot o_j^{L_2}(1 - o_j^{L_2}) \quad (38)$$

$$= \delta_{\cdot j}^{L_1} \quad (39)$$

偏移  $b_j^{L_2}$  的学习公式为：

$$\hat{b}_j^{L_2} = b_j^{L_2} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial b_j^{L_2}} = b_j^{L_2} - \eta \cdot \delta_{\cdot j}^{L_1} \quad (40)$$

例如，对于权重  $\omega_{23}^{L_1}$  的更新学习公式为：

$$\hat{\omega}_{23}^{L_1} = \omega_{23}^{L_2} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial \omega_{23}^{L_1}} \quad (41)$$

$$= \omega_{23}^{L_2} - \eta \cdot \left( \left( \sum_{j=1}^2 \frac{\partial E_{\text{Total}}}{\partial o_j^{L_3}} \frac{\partial o_j^{L_3}}{\partial \text{net}_j^{L_3}} \frac{\partial \text{net}_j^{L_3}}{\partial o_3^{L_2}} \right) \cdot \frac{\partial o_3^{L_2}}{\partial \text{net}_3^{L_2}} \cdot \frac{\partial \text{net}_3^{L_2}}{\partial \omega_{23}^{L_1}} \right) \quad (42)$$

$$= \omega_{23}^{L_2} - \eta \cdot \left( \sum_{j=1}^2 \delta_{\cdot j}^{\text{Layer}_2} \omega_{3j}^{L_2} \right) \cdot f_3^{L_2}(\text{net}_3^{L_2}) (1 - f_3^{L_2}(\text{net}_3^{L_2})) \cdot x_2 \quad (43)$$

$$= \omega_{23}^{L_2} - \eta \cdot \left( \sum_{j=1}^2 \delta_{\cdot j}^{\text{Layer}_2} \omega_{3j}^{L_2} \right) \cdot o_3^{L_2} (1 - o_3^{L_2}) \cdot x_2 \quad (44)$$

例如，对于偏移  $b_3^{L_2}$  的更新学习公式为：

$$\hat{b}_3^{L_2} = b_3^{L_2} - \eta \cdot \frac{\partial E_{\text{Total}}}{\partial b_3^{L_2}} \quad (45)$$

$$= b_3^{L_2} - \eta \cdot \left( \left( \sum_{j=1}^2 \frac{\partial E_{\text{Total}}}{\partial o_j^{L_3}} \frac{\partial o_j^{L_3}}{\partial \text{net}_j^{L_3}} \frac{\partial \text{net}_j^{L_3}}{\partial o_3^{L_2}} \right) \cdot \frac{\partial o_3^{L_2}}{\partial \text{net}_3^{L_2}} \cdot \frac{\partial \text{net}_3^{L_2}}{\partial b_3^{L_2}} \right) \quad (46)$$

$$= b_3^{L_2} - \eta \cdot \left( \sum_{j=1}^2 \delta_{\cdot j}^{\text{Layer}_2} \omega_{3j}^{L_2} \right) \cdot f_3^{L_2}(\text{net}_3^{L_2}) (1 - f_3^{L_2}(\text{net}_3^{L_2})) \cdot 1 \quad (47)$$

$$= \omega_{23}^{L_2} - \eta \cdot \left( \sum_{j=1}^2 \delta_{\cdot j}^{\text{Layer}_2} \omega_{3j}^{L_2} \right) \cdot o_3^{L_2} (1 - o_3^{L_2}) \quad (48)$$