反向传播算法

Yimeng Ren

目录

使用交叉熵损失函数，对于样本 (x, y)，其损失函数为：

$$\mathcal{L}(y, \hat{y}) = -y^{\mathrm{T}} \log \hat{y}$$

对第 l 层中的参数 $W^{(l)}$ 和 $b^{(l)}$ 计算偏导数：

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial w_{ij}^{(l)}} = \frac{\partial z^{(l)}}{\partial w_{ij}^{(l)}} \frac{\partial \mathcal{L}(y, \hat{y})}{\partial z^{(l)}}$$

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial b^{(l)}} = \frac{\partial z^{(l)}}{\partial b^{(l)}} \frac{\partial \mathcal{L}(y, \hat{y})}{\partial z^{(l)}}$$

第一部分偏导 $\frac{\partial z^{(l)}}{\partial w_{ij}^{(l)}}$，由于 $z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$，偏导数

$$\begin{aligned}
\frac{\partial z^{(l)}}{\partial w_{ij}^{(l)}} &= \left[\frac{\partial z_1^{(l)}}{\partial w_{ij}^{(l)}}, \cdots, \frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}}, \cdots, \frac{\partial z_{m^{(l)}}^{(l)}}{\partial w_{ij}^{(l)}} \right] \\
&= \left[0, \cdots, \frac{\partial \left(w_{i:}^{(l)} a^{(l-1)} + b_i^{(l)} \right)}{\partial w_{ij}^{(l)}}, \cdots, 0 \right] \\
&= \left[0, \cdots, a_j^{(l-1)}, \cdots, 0 \right] \\
&\triangleq \mathbb{I}_i \left(a_j^{(l-1)} \right) \quad \in \mathbb{R}^{m^{(l)}}
\end{aligned} \tag{1}$$

第二部分偏导 $\frac{\partial z^{(l)}}{\partial b^{(l)}}$，因为 $z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$，偏导数

$$\frac{\partial z^{(l)}}{\partial b^{(l)}} = I_{m^{(l)}} \in \mathbb{R}^{m^{(l)} \times m^{(l)}}$$

第三部分偏导记为 $\frac{\partial \mathcal{L}(y,\hat{y})}{\partial z^{(l)}} \triangleq \delta^{(l)}$

这个值表示第 l 层神经元对最终损失的影响（贡献程度），也反映了最终损失对第 l 层神经元的敏感程度。

根据 $z^{(l+1)} = W^{(l+1)}a^{(l)} + b^{(l+1)}$，有 $\frac{\partial z^{(l+1)}}{\partial a^{(l)}} = \left(W^{(l+1)}\right)^{\mathrm{T}}$；

根据 $a^{(l)} = f_l\left(z^{(l)}\right)$，有

$$\begin{aligned}
\frac{\partial a^{(l)}}{\partial z^{(l)}} &= \frac{\partial f_l\left(z^{(l)}\right)}{\partial z^{(l)}} \\
&= \operatorname{diag}\left(f_l'\left(z^{(l)}\right)\right)
\end{aligned}$$

根据链式法则，有

$$\begin{aligned}
\delta^{(l)} &\triangleq \frac{\partial \mathcal{L}(y,\hat{y})}{\partial z^{(l)}} \\
&= \frac{\partial a^{(l)}}{\partial z^{(l)}} \cdot \frac{\partial z^{(l+1)}}{\partial a^{(l)}} \cdot \frac{\partial \mathcal{L}(y,\hat{y})}{\partial z^{(l+1)}} \\
&= \operatorname{diag}\left(f_l'\left(z^{(l)}\right)\right) \cdot \left(W^{(l+1)}\right)^{\mathrm{T}} \cdot \delta^{(l+1)} \\
&= f_l'\left(z^{(l)}\right) \odot \left(\left(W^{(l+1)}\right)^{\mathrm{T}} \delta^{(l+1)}\right)
\end{aligned} \tag{2}$$

第 l 层的一个神经元的敏感性是所有与该神经元相连的第 l+1 层的神经元的敏感性加权求和之后再乘该神经元的激活函数梯度。

计算出上面的三个偏导数之后，损失函数对于权重 w 的梯度就可以写成：

$$\begin{aligned}
\frac{\partial \mathcal{L}(y,\hat{y})}{\partial w_{ij}^{(l)}} &= \mathbb{I}_i\left(a_j^{(l-1)}\right)\delta^{(l)} \\
&= \left[0, \cdots, a_j^{(l-1)}, \cdots, 0\right]\left[\delta_1^{(l)}, \cdots, \delta_i^{(l)}, \cdots, \delta_{m^{(l)}}^{(l)}\right]^{\mathrm{T}} \\
&= \delta_i^{(l)} a_j^{(l-1)}
\end{aligned} \tag{3}$$

即：

$$\left[\frac{\partial \mathcal{L}(y,\hat{y})}{\partial W^{(l)}}\right]_{ij} = \left[\delta^{(l)}\left(a^{(l-1)}\right)^{\mathrm{T}}\right]_{ij}$$

$\mathcal{L}(y, \hat{y})$ 关于第 l 层权重 $W^{(l)}$ 的梯度为:

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial W^{(l)}} = \delta^{(l)} \left(a^{(l-1)}\right)^{\mathrm{T}}$$

$\mathcal{L}(y, \hat{y})$ 关于第 l 层偏移项 $b^{(l)}$ 的梯度为:

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial b^{(l)}} = \delta^{(l)}$$