形状

输入X：形状为(N, D_{in})

- X_{nj}表示第n个样本的第j个输入特征
- N表示批次大小(样本数量)，D_{in}是输入特征的维度

权重W：形状为(D_{out}, D_{in})

- W_{ij}表示连接到第j个输入特征到第i个输出特征的权重
- D_{out}是输出特征的维度

偏置B：形状为(D_{out})

- B_i表示第i个输出特征的偏置

输出Y：形状为(N, D_{out})

- Y_{ij}表示第i个样本的第j个输出特征

损失函数对Y的梯度$\frac{\partial L}{\partial Y}$：形状与$Y$相同

线性层的正向传播公式

$$Y = XW^T + B$$

计算 y_{ij}

$$y_{ij} = \sum_{k=1}^{D_{in}} x_{ik} w_{jk} + b_j$$

其中:

- i从1到N
- j从1到D_{out}
- k从1到D_{in}

这里举个实际的例子

假设：

X:

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

W:

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix}$$

B:

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

W^T:

$$\begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \end{bmatrix}$$

Y 的形状为 $[2,3]$:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{bmatrix}$$

其中：

$$\begin{aligned} y_{11} &= x_{11}w_{11} + x_{12}w_{12} + b_1 \\ y_{12} &= x_{11}w_{21} + x_{12}w_{22} + b_2 \\ y_{13} &= x_{11}w_{31} + x_{12}w_{32} + b_3 \\ y_{21} &= x_{21}w_{11} + x_{22}w_{12} + b_1 \\ y_{22} &= x_{21}w_{21} + x_{22}w_{22} + b_2 \\ y_{23} &= x_{21}w_{31} + x_{22}w_{32} + b_3 \end{aligned}$$

对权重W的梯度($\frac{\partial L}{\partial W}$)

$$\frac{\partial L}{\partial w_{ij}} = \sum_{m=1}^{D_{in}} \sum_{n=1}^{D_{out}} \frac{\partial L}{\partial y_{mn}} \frac{\partial y_{mn}}{\partial w_{ij}}$$

只有当 $n == i$ 时，y_{mn} 才依赖与 w_{ij}, 所以上述公式又可以写成：

$$\frac{\partial L}{\partial w_{ij}} = \sum_{m=1}^{D_{in}} \frac{\partial L}{\partial y_{mi}} \frac{\partial y_{mi}}{\partial w_{ij}}$$

且 $\frac{\partial y_{mi}}{\partial w_{ij}} = x_{mj}$,所以:

$$\frac{\partial L}{\partial w_{ij}} = \sum_{m=1}^{D_{in}} \frac{\partial L}{\partial y_{mi}} x_{mj}$$

将其写成矩阵形式,这意味着 $\frac{\partial L}{\partial W}$ 是 $\left(\frac{\partial L}{\partial Y}\right)^T$ 和 X 的矩阵乘积

$$\frac{\partial L}{\partial W} = (\frac{\partial L}{\partial Y})^T X$$

计算对输入X的梯度($\frac{\partial L}{\partial X}$)

$$\frac{\partial L}{\partial x_{ij}} = \sum_{m=1}^{D_{in}} \sum_{n=1}^{D_{out}} \frac{\partial L}{\partial y_{mn}} \frac{\partial y_{mn}}{\partial x_{ij}}$$

只有当 $m == i$ 时，y_{mn}才依赖于x_{ij}$，所以上述公式又可以写成:

$$\frac{\partial L}{\partial x_{ij}} = \sum_{n=1}^{D_{out}} \frac{\partial L}{\partial y_{in}} \frac{\partial y_{in}}{\partial x_{ij}}$$

且 $\frac{\partial y_{in}}{\partial x_{ij}} = w_{nj}$, 所以:

$$\frac{\partial L}{\partial x_{ij}} = \sum_{n=1}^{D_{out}} \frac{\partial L}{\partial y_{in}} w_{nj}$$

将其写成矩阵形式,这意味着 $\frac{\partial L}{\partial X}$ 是 $\frac{\partial L}{\partial Y}$ 和 W 的矩阵乘积:

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} W$$

对偏置B的梯度$\frac{\partial L}{\partial B}$

偏置 B 是一个向量,其每个元素b_k 会加到输出 Y 的每一行的第 k 列。

$$\frac{\partial L}{\partial b_k} = \sum_{i=1}^{D_{in}} \frac{\partial L}{\partial y_{ik}} \frac{\partial y_{ik}}{\partial b_k}$$

由于 $\frac{\partial y_{ik}}{\partial b_k} = 1$, 所以:

$$\frac{\partial L}{\partial b_k} = \sum_{i=1}^{D_{in}} \frac{\partial L}{\partial y_{ik}} \cdot 1$$

这意味着对偏置的梯度，就是将 $\frac{\partial L}{\partial Y}$ 沿行维度求和

$$\frac{\partial L}{\partial B} = sum\left(\frac{\partial L}{\partial Y}, dim = 0\right)$$

举个例子：

设 $\frac{\partial L}{\partial Y}$ 如下：

$$\begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix}$$

则：

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial y_{11}} + \frac{\partial L}{\partial y_{21}}$$
$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial y_{12}} + \frac{\partial L}{\partial y_{22}}$$
$$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial y_{13}} + \frac{\partial L}{\partial y_{23}}$$