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_{ni}表示第n个样本的第i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_{ni}

$$y_{ni} = \sum_{j=1}^{D_{in}} x_{nj} w_{ij} + B_i$$

其中:

- n从1到N
- i从1到D_{out}
- p从1到D_{in}

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

$$\frac{\partial L}{\partial W_{ij}} = \sum_{n=1}^{N} \sum_{k=1}^{D_{out}} \frac{\partial L}{\partial Y_{nk}} \frac{\partial Y_{nk}}{\partial W_{ij}}$$

只有当$k == i$时，Y_{nk}才依赖与W_{ij}，且$\frac{\partial Y_{ni}}{\partial W_{ij}} = X_{nj}$，所以：

$$\frac{\partial L}{\partial W_{ij}} = \sum_{n=1}^{N} \frac{\partial L}{\partial Y_{ni}} X_{nj}$$

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

$$\frac{\partial L}{\partial X_{nj}} = \sum_{k=1}^{D_{out}} \frac{\partial L}{\partial Y_{nk}} \frac{\partial Y_{nk}}{\partial X_{nj}}$$

由于$\frac{\partial Y_{nk}}{\partial X_{nj}} = W_{kj}$,所以:

$$\frac{\partial L}{\partial X_{nj}} = \sum_{k=1}^{D_{out}} \frac{\partial L}{\partial Y_{nk}} W_{kj}$$

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。

$$\frac{\partial L}{\partial B_k} = \sum_{n=1}^{N} \frac{\partial L}{\partial Y_{nk}} \frac{\partial Y_{nk}}{\partial B_k}$$

由于$\frac{\partial Y_{nk}}{\partial B_k} = 1$,所以:

$$\frac{\partial L}{\partial B_k} = \sum_{n=1}^{N} \frac{\partial L}{\partial Y_{nk}}$$

这意味着对偏置的梯度，就是将$\frac{\partial L}{\partial Y}$沿着批次维度(行)求和

$$\frac{\partial L}{\partial B} = \sum_{n=1}^{N} \frac{\partial L}{\partial Y}$$

或者更简洁的表示为:

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