基本公式

$$y = \frac{x - \mu}{\sqrt{\sigma + \epsilon}} \odot \gamma + \beta$$

其中:

均值

$$\mu = \frac{1}{D} \sum_{k=1}^{D} x_i$$

方差

$$\sigma^2 = rac{1}{D} \sum_{k=1}^D \left(x_i - \mu
ight)^2$$

• 计算中间值,用于反向传播

$$rstd = rac{1}{\sqrt{\sigma^2 + \epsilon}}$$

前向传播公式

对于每个输入张量 x,层归一化对每个样本 $x_i = [x_{i1}, x_{i2}, ..., x_{iD}]$ 进行操作。

1. 计算均值 (Mean)

对每个样本 i,计算其所在特征的均值 μ_i :

$$\mu_i = rac{1}{D} \sum_{j=1}^D x_{ij}$$

2. 计算方差 (Variance)

对于每个样本 i, 计算其所有特征的方差 σ_i^2 :

$$\sigma_i^2 = rac{1}{D}\sum_{i=1}^D \left(x_{ij} - \mu_i
ight)^2$$

3. 归一化(Normalization)

使用均值和方差对每个特征进行归一化,得到 \hat{x}_{ij} :

$$\hat{x}_{ij} = rac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

其中 ϵ 是一个很小的常数 (例如 $\frac{1}{10^5}$),用于防止除以零。

4. 缩放和平移 (Scale and Shift)

引入可学习参数 $\gamma \in \mathbb{R}^D$ (缩放因子) 和 $\beta \in \mathbb{R}^D$ (平移因子), 对归一化后的特征进行缩放和平移:

$$y_{ij} = \gamma_j \hat{x}_{ij} + eta_j$$

这里, γ_j 和 β_j 是对应于第 j 个特征的参数。

反向传播公式推导

反向传播的目标是计算损失函数 L 对输入 x 以及可学习参数 γ 和 β 的梯度。

假设我们已经得到了损失函数 L 对输出 y 的梯度 $\frac{\partial L}{\partial y_{ii}}$.

1. 对 γ 和 β 的梯度:

根据 $y_{ij} = \gamma_j \hat{x}_{ij} + \beta_j$,我们有:

$$rac{\partial L}{\partial \gamma_j} = \sum_{i=1}^N rac{\partial L}{\partial y_{ij}} rac{\partial y_{ij}}{\partial \gamma_j} = \sum_{i=1}^N rac{\partial L}{\partial y_{ij}} \hat{x}_{ij}$$

$$rac{\partial L}{\partial eta_j} = \sum_{i=1}^N rac{\partial L}{\partial y_{ij}} rac{\partial y_{ij}}{\partial eta_j} = \sum_{i=1}^N rac{\partial L}{\partial y_{ij}} \cdot 1$$

2. 对 \hat{x}_{ij} 的梯度:

$$rac{\partial L}{\partial \hat{x}_{ij}} = rac{\partial L}{\partial y_{ij}} rac{\partial y_{ij}}{\partial \hat{x}_{ij}} = rac{\partial L}{\partial y_{ij}} \gamma_j$$

3. 对 μ_i 和 σ_i^2 的梯度:

这个部分比较复杂,因为 μ_i 和 σ_i^2 都是 x_{ij} 的函数,并且 \hat{x}_{ij} 也依赖他们。

我们有:

$$\hat{x}_{ij} = rac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

因此:

$$egin{aligned} rac{\partial \hat{x}_{ij}}{\partial \mu_i} &= -rac{1}{\sqrt{\sigma_i^2 + \epsilon}} \ rac{\partial \hat{x}_{ij}}{\partial \sigma_i^2} &= -rac{1}{2} \left(x_{ij} - \mu_i
ight) \left(\sigma_i^2 + \epsilon
ight)^{-rac{3}{2}} \end{aligned}$$

因此,根据链式法则, $\frac{\partial L}{\partial \mu_i}$ 和 $\frac{\partial L}{\partial \sigma_i^2}$ 的计算需要对所有 j 求和:

$$\frac{\partial L}{\partial \mu_i} = \sum_{i=1}^{D} \frac{\partial L}{\partial \hat{x}_{ij}} \frac{\partial \hat{x}_{ij}}{\partial \mu_i}$$

$$rac{\partial L}{\partial \sigma_i^2} = \sum_{i=1}^D rac{\partial L}{\partial \hat{x}_{ij}} rac{\partial \hat{x}_{ij}}{\partial \sigma_i^2}$$

代入 $\frac{\partial \hat{x}_{ij}}{\partial \mu_i}$ 和 $\frac{\partial \hat{x}_{ij}}{\partial \sigma_i^2}$ 的表达式:

$$egin{aligned} rac{\partial L}{\partial \mu_i} &= \sum_{j=1}^D rac{\partial L}{\partial \hat{x}_{ij}} \left(-rac{1}{\sqrt{\sigma_i^2 + \epsilon}}
ight) = \left(-rac{1}{\sqrt{\sigma_i^2 + \epsilon}}
ight) \sum_{j=1}^D rac{\partial L}{\partial \hat{x}_{ij}} \ rac{\partial L}{\partial \sigma_i^2} &= \sum_{j=1}^D rac{\partial L}{\partial \hat{x}_{ij}} \left(-rac{1}{2} \left(x_{ij} - \mu_i
ight) \left(\sigma_i^2 + \epsilon
ight)^{-rac{3}{2}}
ight) = -rac{1}{2 \left(\sigma_i^2 + \epsilon
ight)^{rac{3}{2}}} \sum_{j=1}^D rac{\partial L}{\partial \hat{x}_{ij}} \left(x_{ij} - \mu_i
ight) \end{aligned}$$

4. 对 x_{ij} 的梯度

这里比较复杂,因为 x_{ij} 不仅直接影响 \hat{x}_{ij} ,还通过 μ 和 σ_i^2 间接影响 \hat{x}_{ij}

$$rac{\partial L}{\partial x_{ij}} = rac{\partial L}{\partial \hat{x}_{ij}} rac{\partial \hat{x}_{ij}}{\partial x_{ij}} + rac{\partial L}{\partial \mu_i} rac{\partial \mu_i}{\partial x_{ij}} + rac{\partial L}{\partial \sigma_i^2} rac{\partial \sigma_i^2}{\partial x_{ij}}$$

我们分别计算这些偏导数:

$$ullet$$
 $\frac{\partial \hat{x}_{ij}}{\partial x_{ij}} = \frac{1}{\sqrt{\sigma_i^2 + \epsilon}}$

•
$$\frac{\partial \mu_i}{\partial x_{ij}} = \frac{\partial \left(\frac{1}{D} \sum_{j=1}^D x_{ij}\right)}{\partial x_{ij}} = \frac{1}{D}$$

• $\frac{\partial \sigma_i^2}{\partial x_{ij}}$, 注意: σ_i^2 是 x_{ij} 的函数,也是 μ_i 的函数,而 μ_i 是 x_{ij} 的函数,因此:

$$\frac{\partial \sigma_i^2}{\partial x_{ij}} = \frac{\partial \left(\frac{1}{D} \sum_{k=1}^D (x_{ik} - \mu_i)^2\right)}{\partial x_{ij}}$$

$$= \frac{1}{D} \left(\frac{\partial \left(\sum_{k=1}^D \left(x_{ik}^2 - 2x_{ik}\mu_i + \mu_i^2\right)\right)}{\partial x_{ij}}\right)$$

$$= \frac{1}{D} \left(\frac{\partial \sum_{k=1}^D x_{ik}^2}{\partial x_{ij}} - \frac{2\partial \sum_{k=1}^D x_{ik}\mu_i}{\partial x_{ij}} + \frac{\partial \sum_{k=1}^D \mu_i^2}{\partial x_{ij}}\right)$$

$$= \frac{1}{D} \left[2x_{ij} - 2\left(\frac{\partial \sum_{k=1}^D x_{ik}}{\partial x_{ij}}\mu_i + \frac{\partial \mu_i}{\partial x_{ij}}\sum_{k=1}^D x_{ik}\right) + \left(2\sum_{k=1}^D \mu_i \frac{\partial \mu_i}{\partial x_{ij}}\right)\right]$$

$$= \frac{1}{D} \left[2x_{ij} - 2\left(\mu_i + \frac{1}{D}D\mu_i\right) + 2\sum_{k=1}^D \mu_i \frac{1}{D}\right]$$

$$= \frac{1}{D} \left[2x_{ij} - 2\left(\mu_i + \mu_i\right) + 2\mu_i\right]$$

$$= \frac{2}{D} \left(x_{ij} - \mu_i\right)$$

或者, 更加简洁的推导:

$$rac{\partial \sigma_i^2}{\partial x_{ij}} = rac{\partial \left(rac{1}{D} \sum_{k=1}^D \left(x_{ik} - \mu_i
ight)^2
ight)}{\partial x_{ij}}$$

$$egin{aligned} &=rac{1}{D}\sum_{k=1}^{D}rac{\partial\left(x_{ik}-\mu_{i}
ight)^{2}}{\partial x_{ij}} \ &=rac{1}{D}\sum_{k=1}^{D}2\left(x_{ik}-\mu_{i}
ight)\left(rac{\partial x_{ik}}{\partial x_{ij}}-rac{\partial\mu_{i}}{\partial x_{ij}}
ight) \ &=rac{1}{D}\sum_{k=1}^{D}2\left(x_{ik}-\mu_{i}
ight)\left(\delta_{jk}-rac{\partial\mu_{i}}{\partial x_{ij}}
ight) \ &=rac{2}{D}\left[\sum_{k=1}^{D}\left(x_{ik}-\mu_{i}
ight)\delta_{jk}-\sum_{k=1}^{D}\left(x_{ik}-\mu_{i}
ight)rac{\partial\mu_{i}}{\partial x_{ij}}
ight] \ &=rac{2}{D}\left[\left(x_{ij}-\mu_{i}
ight)-rac{1}{D}\sum_{k=1}^{D}\left(x_{ik}-\mu_{i}
ight)
ight] \end{aligned}$$

计算 $\sum_{k=1}^{D} (x_{ik} - \mu_i)$:

$$\sum_{k=1}^{D} \left(x_{ik} - \mu_i
ight) = \sum_{k=1}^{D} x_{ik} - \sum_{k=1}^{D} \mu_i = D \mu_i - D \mu_i = 0$$

因此:

$$rac{\partial \sigma_i^2}{\partial x_{ij}} = rac{2}{D} \left[(x_{ij} - \mu_i) - rac{1}{D} \sum_{k=1}^D \left(x_{ik} - \mu_i
ight)
ight] = rac{2}{D} \left(x_{ij} - \mu_i
ight)$$

现在,将所有项代入 $\frac{\partial L}{\partial x_{ij}}$ 的公式:

$$egin{aligned} rac{\partial L}{\partial x_{ij}} &= rac{\partial L}{\partial \hat{x}_{ij}} rac{1}{\sqrt{\sigma_i^2 + \epsilon}} + \left(-rac{1}{\sqrt{\sigma_i^2 + \epsilon}} \sum_{j=1}^D rac{\partial L}{\partial \hat{x}_{ij}}
ight) rac{1}{D} + \left(-rac{1}{2\left(\sigma_i^2 + \epsilon
ight)^{rac{3}{2}}} \sum_{k=1}^D rac{\partial L}{\partial \hat{x}_{ik}} \left(x_{ik} - \mu_i
ight)
ight) rac{2}{D} \left(x_{ij}
ight) \ &= rac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left[rac{\partial L}{\partial \hat{x}_{ij}} - rac{1}{D} \sum_{k=1}^D rac{\partial L}{\partial \hat{x}_{ik}} - rac{x_{ij} - \mu_i}{D\left(\sigma_i^2 + \epsilon
ight)} \sum_{k=1}^D rac{\partial L}{\partial \hat{x}_{ik}} \left(x_{ik} - \mu_i
ight)
ight] \end{aligned}$$

其中:

$$\hat{x}_{ij} = rac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \ \hat{x}_{ik} = rac{x_{ik} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

所以:

$$rac{\partial L}{\partial x_{ij}} = rac{1}{D} rac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left(D rac{\partial L}{\partial \hat{x}_{ij}} - \sum_{k=1}^D rac{\partial L}{\partial \hat{x}_{ik}} - \hat{x}_{ij} \sum_{k=1}^D \hat{x}_{ik} rac{\partial L}{\partial \hat{x}_{ik}}
ight).$$

最终, 所以的梯度公式如下:

$$\begin{split} \frac{\partial L}{\partial \gamma_j} &= \sum_{i=1}^N \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \gamma_j} = \sum_{i=1}^N \frac{\partial L}{\partial y_{ij}} \hat{x}_{ij} \\ \frac{\partial L}{\partial \beta_j} &= \sum_{i=1}^N \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial \beta_j} = \sum_{i=1}^N \frac{\partial L}{\partial y_{ij}} \cdot 1 \\ \frac{\partial L}{\partial x_{ij}} &= \frac{1}{D} \frac{1}{\sqrt{\sigma_i^2 + \epsilon}} \left(D \frac{\partial L}{\partial \hat{x}_{ij}} - \sum_{k=1}^D \frac{\partial L}{\partial \hat{x}_{ik}} - \hat{x}_{ij} \sum_{k=1}^D \hat{x}_{ik} \frac{\partial L}{\partial \hat{x}_{ik}} \right) \end{split}$$

为了提高计算效率,我们在前向传播时,需要缓存中间计算结果;

$$mean_i = rac{1}{D} \sum_{k=1}^{D} x_{ik} \ rstd_i = rac{1}{\sqrt{\sigma_i^2 + \epsilon}}$$

从另外的视角看反向传播公式

• 根据链式法则:

$$rac{\partial L}{\partial x_{ij}} = rac{\partial L}{\partial \hat{x}_{ij}} rac{\partial \hat{x}_{ij}}{\partial x_{ij}} + rac{\partial L}{\partial \mu_i} rac{\partial \mu_i}{\partial x_{ij}} + rac{\partial L}{\partial \sigma_i^2} rac{\partial \sigma_i^2}{\partial x_{ij}} + rac{\partial L}{\partial \sigma_i^2} rac{\partial \sigma_i^2}{\partial \mu_i} rac{\partial \mu_i}{\partial x_{ij}}$$

因为:

$$rac{\partial \sigma_i^2}{\partial \mu_i} = rac{2}{D} \sum_{k=1}^D \left(\mu_i - x_{ik}
ight) = 2 \left(rac{1}{D} D \mu_i - rac{1}{D} \sum_{k=1}^D x_{ik}
ight) = 2 \left(\mu_i - \mu_i
ight) = 0$$

有一般性的结论:由于方差的数学性质,方差对均值的导数恰好为零。 可以从最小二乘的角度看,方差实际上是:

$$\sigma^2 = \min_c rac{1}{D} \sum_{i=1}^D \left(x_i - c
ight)^2$$

在最优点处,目标函数对参数的导数必须为零:

$$\frac{\partial \sigma^2}{\partial c} = 0$$

解得:

$$c = rac{1}{D} \sum_{i=1}^D x_i.$$

这正是均值,因而,方差的最优化解是在导数为零的时候取得。