总结

到了现在，大家通过简单的例子以及穿插的代码对神经网络的搭建和训练有了最初步的认识。但是之前的内容仅仅是为了让大家略知一二。当大家对整体过程有了一定认识之后，我们再对每一个地方进行细节的讲解。否则，如果像传统的书籍一样，按顺序对每一个点进行深入而且严谨的讲解，会让读者不清楚这样的目的是什么。

扩展

现在，我们基于之前学习的关于机器学习的基本知识和搭建的初步神经网络对内容进行扩展。让算法和代码的每一部分都变得越来越严谨。

机器学习的任务：分类和回归

分类

回归

激活函数

常见激活函数

非线性性

凹凸性

可学习激活函数

参数初始化

数值优化的初始值对数值优化的结果有巨大影响。而这个问题，我们也可以从不同的角度来看待。

随机初始化

Xavier

对于深度神经网络，会存在这样一个问题，就是当层数过多时，输出的值以及分布会越来越不正常。例如在加权求和中$Z=XW$，如果W的数学期望大于1，那么Z就会比X大一点，那么每一层的输出值都会增加一些，而且值的分布方差也会增加一些，这样经过多层之后会导致许多问题，例如梯度爆炸或者梯度消失。Xavier初始化的目标就是通过选择合适的权重W，使得每一层输出和输入的分布一致，这样的话无论传递多少层，都不会出现极端的情况。

要注意的是，Xavier初始化假设的是激活函数经过零点，且在零点处斜率为1。同时，假设输入X和权重W都是期望为0，方差很小的变量，因此Z也是期望为0，方差很小的变量。这样的话，一个期望值为0，方差较小的随机变量经过激活函数后，分布不变。

我们把输出Z表示为$\mathbb{E}\{Z\}+\delta_Z$，其中$\mathbb{E}\{Z\}=0$，并且δ_Z是一个期望为0方差为$\mathrm{Var}(Z)$的随机变量。那么，激活函数的输出是：

$$\mathrm{act}\left(\mathbb{E}\{Z\}+\delta_z\right)\approx\mathrm{act}\left(\mathbb{E}\{Z\}\right)+\frac{\partial\,\mathrm{act}(\mathbb{E}\{Z\})}{\partial Z}\delta_Z=\delta_Z$$

所以经过激活函数之后的变量的期望和方差是：

$$\begin{aligned}\mathrm{Var}(\mathrm{act}\left(\mathbb{E}\{Z\}+\delta_z\right))&=\mathrm{Var}(\delta_z)=\mathrm{Var}(Z)\\\mathbb{E}\left\{\mathrm{act}(\mathbb{E}\{Z\}\pm\delta_z)\right\}&=\mathbb{E}\left\{\mathbb{E}\{Z\}+\delta_z)\right\}=0=\mathbb{E}\{Z\}\end{aligned}$$

在这种假设之后，我们就可以忽略激活函数对于分布的影响了。这种情况下，我们假设输入是

$$X\in\mathbb{R}^{M\times E}$$

权重矩阵是

$$W\in\mathbb{R}^{N\times M}$$

输出是

$$Z=WX\in\mathbb{R}^{N\times E}$$

它的方差就是

$$\mathrm{Var}(Z)=\frac{1}{E}(WX-\bar{W}\bar{X})(WX-\bar{W}\bar{X})^\top$$

根据假设，W和X的数学期望是0，所以

$$\begin{aligned}\mathrm{Var}(Z)&=\frac{1}{E}(WX)(WX)^\top\\&=\frac{1}{E}WXX^\top W^\top\\&=\frac{1}{E}W\cdot\mathrm{Var}(X)\cdot W^\top\end{aligned}$$

由于各个特征之间相互独立，并且已经假设每个特征的方差都是一样，记为σ^2，我们简化它为$\mathrm{Var}(Z)=\sigma^2I$和$\mathrm{Var}(X)=\sigma^2I$。其中$I$是单位矩阵，维度根据所乘对象而改变。所以：

$$\begin{aligned}\sigma^2I&=\frac{1}{E}W\cdot E\sigma^2I\cdot W^\top\\I&=WW^\top\\&=M\cdot\mathrm{Var}(W)\\\mathrm{Var}(W)&=\frac{I}{M}\end{aligned}$$

其中M是输入特征个数。

在工程上也有用输入和输出的平均值，也就是$\mathrm{Var}(W)=\frac{2I}{M+N}$作为权重初始方差的，这没有什么数学原理，只是工程经验。

随机种子对初始化的影响

优化策略和优化器

分批训练

SGD

动量

自适应长度

学习率

固定学习率

可变学习率

数据集

预处理

数据集分割

Early stop

实验的可重复性

方便自己

方便他人

In [ ]: