深度学习基础

一、 训练误差和泛化误差

训练误差主要是指的型在训练数据集上表现出的误差；泛化误差主要是指的模型在任意一个测试数据样本上表现出的误差的期望。

一般情况下，由训练数据集学到的模型参数会使模型在训练数据集上的表现优于或等于在测试数据集上的表现。

1.1 模型选择

选择模型（线性回归，逻辑回归等），也可以是选择有着不同超参数的同类模型（参数不同，调参）。

1.1.1 验证数据集

可以预留一部分在训练数据集和测试数据集以外的数据来进行模型选择。这部分数据被称为验证数据集，简称验证集（validation set）。例如，我们可以从给定的训练集中随机选取一小部分作为验证集，而将剩余部分作为真正的训练集。

1.1.2 k 折交叉验证

原始训练数据被分成k个不重叠的子集。 然后执行次模型训练和验证，每次在$k-1$个子集上进行训练， 并在剩余的一个子集（在该轮中没有用于训练的子集）上进行验证。 最后，通过对k次实验的结果取平均来估计训练和验证误差。

1.2 模型复杂度

时间复杂度和空间复杂度是衡量一个算法的两个重要指标,用于表示算法的最差状态所需的时间增长量和所需辅助空间.

在深度学习神经网络模型中我们也通过：

计算量/FLOPS（时间复杂度）即模型的运算次数

访存量/Bytes（空间复杂度）即模型的参数数量

1.3 欠拟合与过拟合

给定训练数据集，如果模型的复杂度过低，很容易出现欠拟合；如果模型复杂度过高，很容易出现过拟合。应对欠拟合和过拟合的一个办法是针对数据集选择合适复杂度的模型。

1.3.1 欠拟合

模型无法得到较低的训练误差，我们将这一现象称作欠拟合（underfitting）

1.3.2 过拟合

模型的训练误差远小于它在测试数据集上的误差，我们称该现象为过拟合

二、 权重衰减

权重衰减是一个正则化技术，作用是抑制模型的过拟合，以此来提高模型的泛化性。正则化是减少数据扰动对预测结果的影响。训练数据点距离真实模型的偏离程度就是数据扰动。

模型权重数值越小，模型的复杂度越低。通过增加惩罚项可以限制参数大小，抑制过拟合。可以用公式表示为：

$$L = L_0 + \frac{\lambda}{2} \|W\|^2 \tag{2 - 1}$$

式中 λ ———— 超参数

$\|\mathbb{W}\|^2$是模型参数的 2 范数的平方

L_0 是原本的损失函数

假设模型有n个参数，$W = [w_1 \ \ w_2 \ \ w_3 \ \ ..... \ \ w_n]$，L 可以表示为：

$$\begin{aligned} L &= L_0 + \frac{\lambda}{2}\left(\sqrt{w_1^2 + w_2^2 + w_3^2 + ... + w_n^2}\right)^2 \\ &= L_0 + \frac{\lambda}{2}(w_1^2 + w_2^2 + w_3^2 + ... + w_n^2) \end{aligned} \tag{2 - 2}$$

这样在 SGD 中的参数更新由$w_i \leftarrow w_i - \gamma\frac{\partial(L)}{\partial(w_i)}$ 变为

$$\begin{aligned} w_i &\leftarrow w_i - \gamma\left(\frac{\partial(L_0)}{\partial(w_i)} + \lambda w_i\right) \\ &= w_i - \gamma\lambda w_i - \gamma\frac{\partial(L_0)}{\partial(w_i)} \\ &= w_i(1 - \gamma\lambda) - \gamma\frac{\partial(L_o)}{\partial(w_i)} \end{aligned} \tag{2 - 3}$$

L_2范数正则化令权重w_1和w_2, 先自乘小于 1 的数，再减去不含惩罚项的梯度。

三、 丢弃法(倒置丢弃法)

使用丢弃法也可以应对过拟合的问题。随机丢弃一部分神经元（同时丢弃其对应的连接边）来避免过拟合。

在多层感知机中单个隐藏层单元的计算为：

$$h_i = \varphi(x_1 w_1 i + x_2 w_2 i + x_3 w_3 i + x_4 w_4 i + b_i) \tag{3 - 1}$$

当对该隐藏层使用丢弃法时，该层的隐藏单元将有一定概率被丢弃掉。设丢弃概率为p， 那么有p的概率h_i会被清零，有$1-p$的概率h_i会除以$1-p$做拉伸。丢弃概率是丢弃法的超参数。可以表示为：

$$h_i' = \begin{cases} 0 & \text{if } p \\ \frac{\zeta_i}{1-p} h_i & \text{else } 1-p \end{cases} \tag{3 - 2}$$

式中 p ———— 超参数

四、 正向传播，反向传播，计算图

4.1 正向传播

正向传播（forward propagation）是指对神经网络沿着从输入层到输出层的顺序，依次计算并存储模型的中间变量（包括输出）。

4.2 反向传播

反向传播（back-propagation）指的是计算神经网络参数梯度的方法。总的来说，反向传播依据微积分中的链式法则，沿着从输出层到输入层的顺序，依次计算并存储目标函数有关神经网络各层的中间变量以及参数的梯度。

4.3 计算图

通过绘制计算图（computational graph）来可视化运算符和变量在计算中的依赖关系。

五、 数值稳定性和模型初始化

层数较多时，梯度的计算也更容易出现衰减或爆炸。每层的参数值会变的特别大或特别小。

5.1 随机初始化模型参数

如果将每个隐藏单元的参数都初始化为相等的值，那么在正向传播时每个隐藏单元将根据相同的输入计算出相同的值，并传递至输出层。在反向传播中，每个隐藏单元的参数梯度值相等。因此，这些参数在使用基于梯度的优化算法迭代后值依然相等。之后的迭代也是如此。在这种情况下，无论隐藏单元有多少，隐藏层本质上只有 1 个隐藏单元在发挥作用。因此，正如在前面的实验中所做的那样，我们通常对神经网络的模型参数，特别是权重参数，进行随机初始化。

六、 引用

[1] 深度学习模型数值稳定性——梯度衰减和梯度爆炸的说明-CSDN 博客 [EB].https://blog.csdn.net/m0_49963403/article/details/132394707.

[2] 3.11. 模型选择、欠拟合和过拟合 — 《动手学深度学习》 文档[EB] https://zh-v1.d2l.ai/chapter_deep-learning-basics/underfit-overfit.html.

[3] 4.4. 模型选择、欠拟合和过拟合 — 动手学深度学习 2.0.0 documentation[EB]. https://zh.d2l.ai/chapter_multilayer-perceptrons/underfit-overfit.html.

[4] 机器学习 *K* 折交叉验证知识详解（深刻理解版）（全网最详细）五折交叉验证得到五个模型-CSDN 博客[EB].https://blog.csdn.net/Rocky6688/article/details/107296546.

[5] 理解深度学习模型复杂度评估 全连接层的计算复杂度-CSDN 博客[EB] https://blog.csdn.net/coco_12345/article/details/105742205.

[6] 权重衰减 weight_decay 参数从入门到精通 weight decay-CSDN 博客[EB] https://blog.csdn.net/zhaohongfei_358/article/details/129625803.

[7] 深度学习入门笔记-13 正则化-丢弃法 Dropout[EB]-知乎专栏. https://zhuanlan.zhihu.com/p/608914928.