梯度相关

July 1, 2020

1 泰勒公式和梯度下降法的关系

见《1-预备知识》

2 浅谈梯度消失的问题 [1]

见《1-预备知识》中的" MSE 均方误差 +Sigmoid 激活函数 "和" 交叉熵损失 +Sigmoid 激活函数 "。

可以看出，损失函数包含两个部分：1. 计算方法（均方差、交叉熵等）；2. 激活函数。

而之前我们遇到的是均方差损失 +sigmoid 激活函数造成了输出层神经元学习率缓慢，解决思路是交叉熵损失 +Sigmoid 激活函数；其实我们破坏任意一个条件都有可能解决这个问题：

1. 均方误差损失 → 交叉熵损失；

2. Sigmoid 函数 → 不会造成梯度消失的函数，例如 ReLU 函数，不仅能解决输出层学习率缓慢，还能解决隐藏层学习率缓慢问题。

2.1 ReLU 相对于 tanh、sigmoid 的好处

第一，采用 sigmoid 等函数，算激活函数是（指数运算），计算量大；反向传播求误差梯度时，求导涉及除法，计算量相对大。而采用 Relu 激活函数，整个过程的计算量节省很多。

第二，对于深层网络，sigmoid 函数反向传播时，很容易就会出现梯度消失的情况（在 sigmoid 接近饱和区时，变换太缓慢，导数趋于 0），这种情况会造成信息丢失，梯度消失在网络层数多的时候尤其明显，从而无法完成深层网络的训练。

第三，ReLU 会使一部分神经元的输出为 0，这样就造成了网络的稀疏性，并且减少了参数的相互依存关系，缓解了过拟合问题的发生。

3 ASGD 收敛性慢的原因

泰勒展开是在原点附近展开时，收敛性较好；否则收敛性较差，甚至不收敛（见《数学相关/泰勒公式在不同点展开的收敛性讨论》）

4 梯度验证 [2]

为了求解一个优化问题，最重要的操作是计算目标函数的梯度。在一些机器学习的应用中，例如深度神经网络，目标函数的梯度公式非常复杂，需要验证自己写出的实现代码是否正确。

4.1 问题描述

假设你需要求解优化问题

$$\min_{w\in\mathbb{R}^2} L(w)$$

并且用代码实现了求目标函数值和求目标函数梯度的功能。问如何利用求目标函数值的功能来验证求目标函数梯度的功能是否正确?

4.2 解答和分析

根据梯度的定义，目标函数的梯度为:

$$\nabla L(w) = \big[\frac{\partial L(w)}{\partial w_1}, \cdots, \frac{\partial L(w)}{\partial w_n}\big]^T$$

其中，对于任意的 $i = 1, 2, \cdots, n$，有

$$\frac{\partial L(w)}{\partial w_i} = \lim_{h\to 0} \frac{L(w + he_i) - L(w - he_i)}{2h}$$

其中，向量 e_i 的长度与 w 的长度相同（二者维度相同），仅在第 i 个位置取值为 1，其余位置取值为 0。因此，我们可以取 h 为一个较小的数（例如 10^{-7}），则有:

$$\frac{\partial L(w)}{\partial w_i} \approx \frac{L(w + he_i) - L(w - he_i)}{2h}$$

该近似式的左边为目标函数梯度的第 i 个分量，右边仅和目标函数值有关。

下面我们根据泰勒展开及拉格朗日余项公式，有:

$$L(w + he_i) = L(w) + L'(w)(he_i) + \frac{1}{2}L''(w)(he_i)^2 + \frac{1}{6}L^{(3)}(w)(p_i)(he_i)^3$$

$$L(w-he_i) = L(w) - L'(w)(he_i) + \frac{1}{2}L''(w)(he_i)^2 - \frac{1}{6}L^{(3)}(w)(q_i)(he_i)^3$$

其中，$p_i \in (0,h), q_i \in (-h,0)$。

两个式子相减，等号两边同时除以 $2h$，并且因为有 $e_i = 1$，可得：

$$\frac{L(w+he_i) - L(w-he_i)}{2h} = \frac{\partial L(w)}{\partial w_i} + \frac{1}{12}(\tilde{L}^{(3)}(w)(p_i) + \tilde{L}^{(3)}(w)(q_i))h^2$$

当 h 较小时，可以近似认为 h^2 项前面的系数为常数 M，则近似误差为：

$$\left|\frac{L(w+he_i) - L(w-he_i)}{2h} - \frac{\partial L(w)}{\partial w_i}\right| \approx Mh^2$$

当 h 较小时，h 每减小为原来的 10^{-1}，近似误差减小为原来的 10^{-2}，也就是近似误差是 h 的高阶无穷小。

实际中，我们随机初始化 w，取 h 为较小的数（例如 10^{-7}），并对 $t = 1, 2, \cdots, n$，验证：

$$\left|\frac{L(w+he_i) - L(w-he_i)}{2h} - \frac{\partial L(w)}{\partial w_i}\right| <= h$$

是否成立。如果对于某个下标 i，该不等式不成立，那么有两种情况：该下标对应的 M 过大，或者该梯度分量计算不正确。此时可以固定 w，减小 h 为原来的 10^{-1}，并再次计算下标 i 对应的近似误差。若该近似误差约减小为原来的 10^{-2}，则对应于第一种情况，我们应该采用更小的 h 重新做一次验证；否则对应于第二种情况，应该检查求梯度的代码是否有错误。

References

[1] “常见损失函数小结.” [Online]. Available: https://zhuanlan.zhihu.com/p/37217242

[2] “Hulu 机器学习问题与解答系列 | 第三弹: 优化简介.” [Online]. Available: https://mp.weixin.qq.com/s?__biz=MzA5NzQyNTcxMA==&mid=2656430247&idx=1&sn=fd2ce2a046ac740449db342968d3cfd5&scene=19#wechat_redirect