AdaDelta算法

1. 产生背景

对于解决AdaGrad算法中在迭代后期可能较难找到有用解的问题，除了RMSProp算法外，还可以使用AdaDelta算法实现。特别地，AdaDelta算法并没有学习率这一超参数。

1. 算法内容

AdaDelta算法使用了小批量随机梯度按元素平方的指数加权移动平均变量。在时间步0，他的所有元素被初始化为0。给定超参数(与RMSProp中的对应)，在时间步,与RMSProp一样：

此处，AdaDelta算法还维护一个额外的状态变量,其元素同样在时间步0时被初始化为0，使用来计算自变量的变化量：

其中是维持数值稳定性添加的常数。

然后更新变量：

最后，使用来记录自变量变化量按元素平方的指数加权移动平均：

AdaDelta算法与RMSProp算法的不同之处在于使用来带起超参数

1. 代码实现

AdaDelta算法需要对每个自变量维护两个状态变量，即和。



1. 相关文献

[1] Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.