Adam算法

1. 产生背景

Adam算法在RMSProp算法的基础上，对小批量随机梯度也做了指数加权移动平均。

1. 算法内容

Adam算法使用动量变量和RMSProp中的小批量随机梯度按元素平方的指数加权移动平均变量，并在时间步0将它们中每个元素初始化为0。给定超参数(文献中作者建议设为0.9)，时间步t的动量变量（即小批量随机梯度的指数加权移动平平均）：

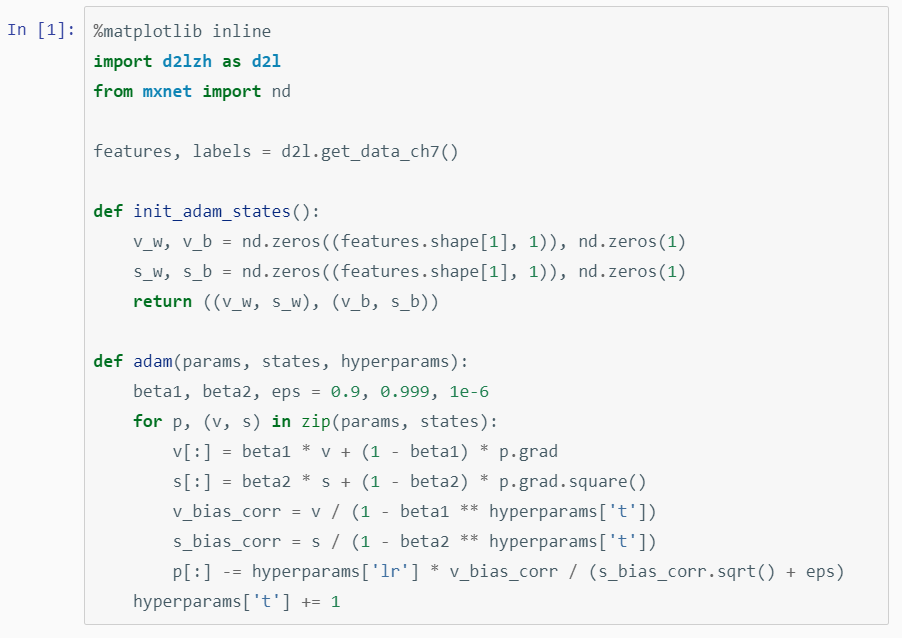
和RMSProp算法中一样，给定超参数(文献中作者建议设为0.999)，将小批量随机梯度按元素平方后的项做指数加权平均得到：

由于我们将和中的元素都初始化为0， 在时间步t我们得到。将过去各时间步小批量随机梯度的权重相加，得到。注意，当 t 较小时，过去各时间步小批量随机梯度权值之和会较小。为了消除这种影响，对于任意时间步t，可以将除以, 从而使过去各时间步小批量随机的梯度权值之和为1，即偏差修正。

然后，Adam算法使用以上偏差修正后的变量与，将模型参数中每个元素的学习率按元素运算重新调整：

其中是学习率，是为了维持数值稳定添加的常数。和AdaGrad算法、RMSProp算法以及AdaDelta算法一样，目标函数自变量中每个元素都分别拥有自己的学习率。最后，使用迭代自变量：

1. 代码实现



1. 相关文献

[1] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.