

5. 梯度下降法的改进过程 (SGD->Adam)

1. 梯度下降法

1.1 梯度下降

2.1 SGD 梯度下降法

2.1.1 梯度下降 (Gradient Descent)

梯度是指函数在某处偏导数，指函数上升方向。

因此梯度下降法是指用梯度的负数 $-g$ 更新参数，从而使下一次的计算沿着函数下降方向逼近，从而得到最小值。

更新时参数学习率：

$$W^* = W - \alpha \nabla W = W - \alpha \frac{\partial}{\partial W} L$$

1.2 批次梯度下降

2.1.2 批次梯度下降 (Batch Gradient Descent)

以所有 m 个数据作为一次批次，

每次计算 loss 值与梯度 g 时，都是计算所有数据累加和。

更新时，也以所有数据的梯度累加和进行计算更新。

缺点：计算所有数据的梯度非常耗时。

1.3 随机梯度下降

2.1.3 随机梯度下降 (Stochastic Gradient Descent, SGD)

虽然 m 个数据为一个批次，但是更新参数时仅使用随机一个数据的梯度进行更新。



扫描全能王 创建

Date: _____

Page: _____

缺点：随机性强，噪声影响严重，不一定向整体最优点下降。

1.4 小批次梯度下降

2.1.4 小批次梯度下降 (Mini-batch GD (MBGD))

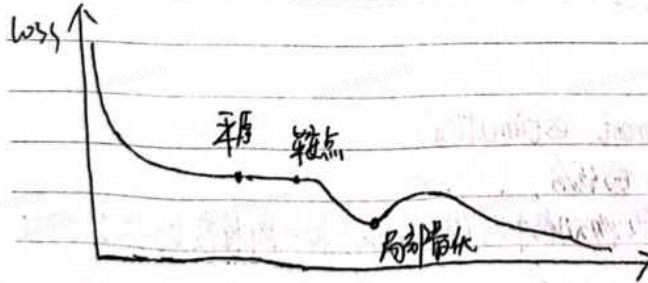
把所有样本分为 m 个 batch (一般是随机的), 每次计算损失和梯度都用一个 batch 的数据进行计算, 并更新参数, 从而避免了唯一随机性和全局计算的耗时性。

优点: 得到的梯度下降方向是局部最优的, 整体速度快。

(一般说的 SGD 就是指 Mini-batch GD)

2. 动量梯度下降

梯度下降法可能会停滞在平原、鞍点、局部最优点 (这三处梯度均为 0), 因此带动量的梯度下降法能依靠之前的梯度值冲过平原、鞍点和局部最优, 提高泛化性。



$$W = W - \alpha \nabla W$$

α : 历史参数的更新值大小, 相当于现在与过去的梯度的一个加权。

可以视为当前一个动量。

3. 自适应梯度下降

2.3 自适应梯度算法 Adagrad (Adaptive gradient)

针对不同的变量提供不同的学习率。

解决方法:

为每一参数建立历史累计梯度值, 利用历史累计梯度作为分母, 从而使各个参数在训练后期被给予不同的除数, 得到自适应参数值。

为什么要不同学习率?

当一些变量被优化到最优点时, 另外一些并没有, 使用统一的学习率就会影响优化过程, 太大会导致震荡, 太小会导致收敛慢。

4. RMSprop 自适应学习率算法

2.4 RMSprop 自适应学习率算法 (root mean square propagation)
不再直接暴力累加平方梯度, 而是使用一个衰减系数来控制历史信息的影响。

5. Adam (Adaptive moment estimation)

是 RMSprop 与 Momentum 的结合,
使用了一阶梯度的指数移动平均 (Momentum) 和二阶梯度的指数移动平均 (RMSprop)。

优点: 每一次迭代学习率都有一个明确的范围, 使得参数变化很平稳。

$$m_w^{t+1} = \beta_1 m_w^t + (1 - \beta_1) \Delta L^t, \quad m \text{ 为一阶矩估计}$$

$$v_w^{t+1} = \beta_2 v_w^t + (1 - \beta_2) (\Delta L^t)^2, \quad v \text{ 为二阶矩估计}$$

$$w^{t+1} = w^t - \eta \frac{m_w}{\sqrt{v_w} + \epsilon}$$