指数加权移动平均

（Exponential Weighted Moving Average）

1. EMA简介：

EMA：是以指数式递减加权的移动平均，各数值的胶圈影响力随时间呈指数式递减，时间越靠近当前时刻的数据加权影响力（权重）越大。

发展过程：

算术平均（权重相等）——>加权平均（权重不等）——>移动平均（指定时间段，对时间序列数据进行移动计算平均值）——>批量归一化（BN）及各种优化算法的基础

1. 推导过程及理解

给定超参数，当前时间步t的变量是上一时间步t-1的变量和当前时间步另一变量的线性组合：

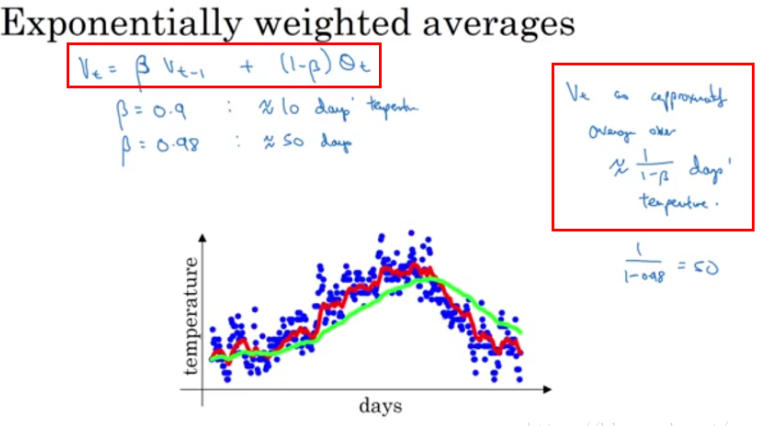
对展开：

令，则：

若把exp(-1)当作一个比较小的数，则在近似中忽略所有含和比高阶的系数的项。

所以在实际过程中常常将看作是对最近个时间步的值的加权平均，并且当前时间步t越接近的值获得的权重越大（越接近1）。

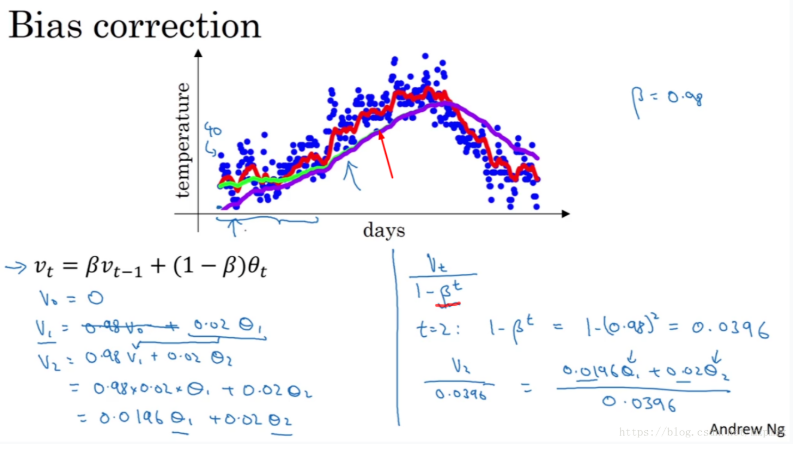
举例：（Andrew Ng deep learning coursea）



例举对于温度的的预测：

1. ，公式中的为t时刻的实际温度，系数表示加权下降的快慢，值越小权重下降的越快，为t时刻的EMA值。
2. 当时，可得：，从公式中可以看出：每天温度（）的权重以指数等比形式缩小，时间越靠近当前时刻的数据加权权重越大。
3. 在优化算法中，一般取，而，所以当t足够大时，，此时便是严格意义上的指数加权移动平均。
4. 关于移动平均的解释：取，此时有，天后，曲线高度下降到了约原来的三分之一，由于时间越往前推移，权重越来越小，所以只考虑最近的天的数据计算当前时刻的EMA，也就是移动平均的来源。

EMA偏差修正：



在时，理想情况应为绿色曲线，而实际是紫色曲线，起点比真实值低，不能很好地预测起始位置的温度，此问题称为：冷启动问题。是由于造成的。

解决方案：将所有时刻的EMA除以后作为修正后的EMA。当t很小时，可以在起始阶段的预测更加准确，当t很大时，偏差修正几乎不起作用，对原来试自几乎没有影响。注意：计算t时刻修正后的EMA时，使用t-1时刻修正前的EMA。

1. EMA优点
2. 占用内存较少：计算指数加权平均数只占用单行数字的存储和内存，然后把最新数据带入公式，不断覆盖即可。
3. 移动平均线能较好地反应时间序列的变化趋势，权重的大小不同起到的作用不同，时间交久远的变量值权重较低，其影响力也相对较低，时间较近的变量值权重较高，影响力也相对较高。
4. 动量法优化

在梯度下降中，给定学习率，梯度下降迭代自变量会使自变量在不同方向下降速度不同，此时需要较小的学习率避免在某方向上超过目标函数的最优解，而同时，会导致自变量在其他方向上朝最优解移动速度减慢。

为解决上述问题，使用动量法，同小批量随机梯度下降（mini batch）中时间步t的小批量随机梯度的定义，设时间步t的自变量为，学习率为。 在时间步为0，动量法创建速度变量，并将其元素初始化为0。在时间步t>0，动量法对每步迭代做以下操作：

其中动量超参数满足：，当时，动量法等价于小批量随机梯度下降。

结合EMA：

对动量法中的速度变量进行变形：

由指数加权移动平均的形式可知：速度变量实际上是对序列做了指数加权移动平均，相当于小批量随机梯度下降。所以，在动量法中，自变量在各个方向上的移动幅度不仅取决当前梯度，还取决于过去的各个梯度在各个方向上是否一致。

在迭代后期，由于随机噪声问题，经常会在收敛值附近震荡，动量法会起到减速作用，增加稳定性。