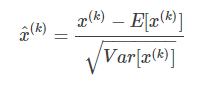
## BN

1. 优点
   1. 快速训练收敛。可以选择比较大的初始学习率。对于学习率、参数初始化、权重衰减系数、Drop out比例等，不需要那么刻意的慢慢调整参数。
   2. 不用处理过拟合中drop out、L2正则项参数的选择问题，采用BN算法后，可以移除这两项了参数，或者可以选择更小的L2正则约束参数了，因为BN具有提高网络泛化能力的特性；
   3. 不需要使用使用局部响应归一化层。BN中也会对数据进行归一化。
   4. 在训练的时候可以把训练数据彻底打乱（防止每批训练的时候，某一个样本都经常被挑选到），文献说这个可以提高1%的精度。
2. 数据归一化的原因
   1. 神经网络学习过程本质就是为了学习数据分布，一旦训练数据与测试数据的分布不同，那么网络的泛化能力也大大降低；
   2. 另外一方面，一旦每批训练数据的分布各不相同(batch 梯度下降)，那么网络就要在每次迭代都去学习适应不同的分布，这样将会大大降低网络的训练速度.
3. BN概述
   1. 实一层。BN操作层，它位于X=WU+B激活值获得之后，非线性函数变换质。在网络的每一层输入的时候，又插入了一个归一化层，也就是先做一个归一化处理，然后再进入网络的下之前
   2. 数据预处理之白话预处理

数据满足条件：a、特征之间的相关性降低，这个就相当于pca；b、数据均值、标准差归一化，也就是使得每一维特征均值为0，标准差为1。

但是白话处理要满足上述两个条件的话，计算量特别大。

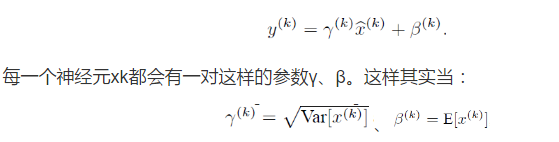
1. BN算法核心思想
   1. 归一化公式（伪白化）

可计算，该公式所有x(k)的估计值的均值为0，方差为1，此时均为标准正太分布~（0,1）。

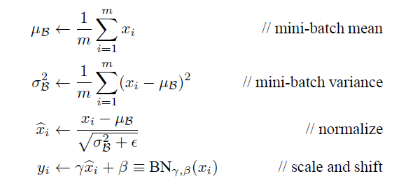
推到过程：

* 1. 数据分布恢复

因为上公式强制将网络中间某一层学习到特征数据给我归一化处理、标准差也限制在了1，把数据变换成分布于s函数的中间部分，损害了该层网络所学到的特征。



可以推到的y(k)=x(k).

则BN网络前向传导公式为：

此时困惑：上面讲过x的估计值服从标准正太分布，则x经过线性变换后的y，仍然服从正太分布，可计算y的均值为(beita谐音)，方差为（gama谐音）。