# Batch Normalization

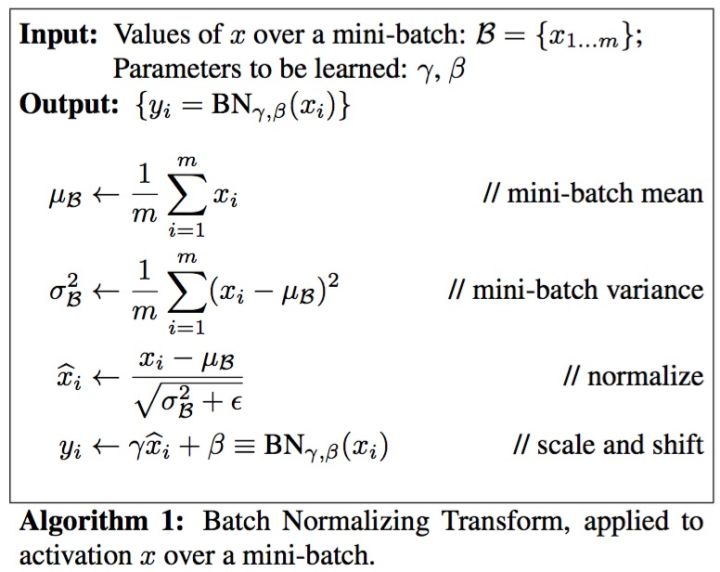
## 解释

Batch Normalization顾名思义即是批规范化（标准化）。

在每一次SGD中，通过mini-batch来对相应的激活值做规范化操作，使得输出结果符合均值为0、方差为1的分布

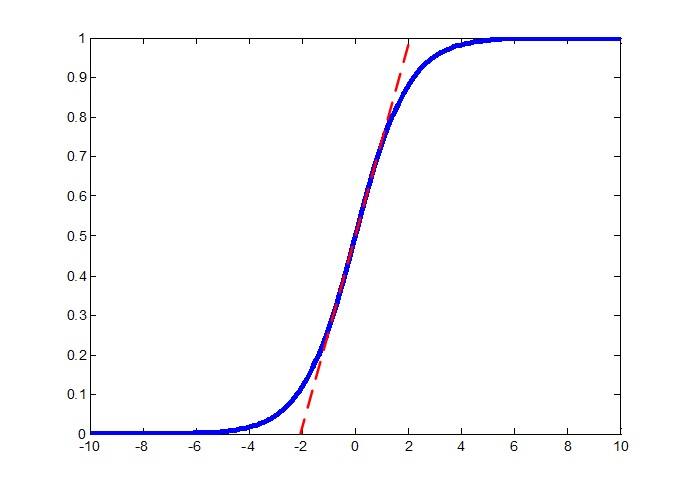
而随后的scale and shift则是为了让规范化后的内容有能力还原最初的输入（不然都规范化，则会丢失原先数据的诸多特性），那么如何在规范化后的结果和原输入中找到一个平衡，依靠的便是可学习的参数和下方的scale and shift操作。

这样既达到了BN的效果，又维护了网络的宽容性（capacity）。



#### 参数的解释

如果我们让等于之前求得的标准差，让等于均值的话，可以发现这个变换相当于Normalization的反操作，还原了原先数据的分布。假设我们使用了sigmoid激活函数，那么Normalization后，数据可能只分布在中心区域，使得学习到的函数几乎是线性的。



对数据做彻底的Normalization是不合适的，但如果还原原数据，相当于就没有使用BN算法，所以被配置为可学习的参数，以找到两者之间的平衡。

## 正向传播

在一个mini-batch中，求解出本层输入（注意不是样本的输入）的均值和方差，两者均是针对一个batch：

其中表示第个输入，表示本层输入的大小.

BN层：

## 反向传播

反向传播即使用链式求导法则即可，注意为了利用链式求导权重，我们需要对求偏导，除了公式中显式存在的，在均值和方差中也存在，方差中也存在均值，所以要仔细整合到一起。

当时取，可转化为下式

上式为对的完全求解，事实上，因为下式我们预先求解了，所以不需要再考虑，当然巧合的是即使考虑了，计算出来的结果也是一样的。