Batch Nomalization

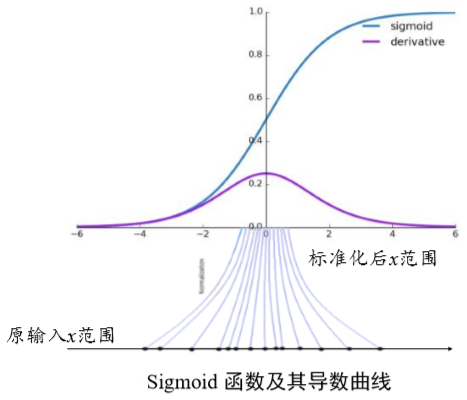
Batch Nomalization是一种参数标准化手段。

# Batch Nomalization的好处

采用两个示例说明。

## 示例一

考虑 Sigmoid 激活函数和它的梯度分布，如下图所示，Sigmoid 函数在𝑥∈[−2 2]区间的导数值在[0.1,0.25]区间分布；当𝑥 > 2或𝑥 < −2时，Sigmoid函数的导数变得很小，逼近于0，从而容易出现梯度弥散现象。为了避免因为输入较大或者较小而导致Sigmoid函数出现梯度弥散现象，将函数输入𝑥标准化映射到0附近的一段较小区间将变得非常重要，可以从下图看到，通过标准化重映射后，值被映射在0附近，此处的导数值不至于过小，从而不容易出现梯度弥散现象。这是使用标准化手段受益的一个例子。



## 示例二

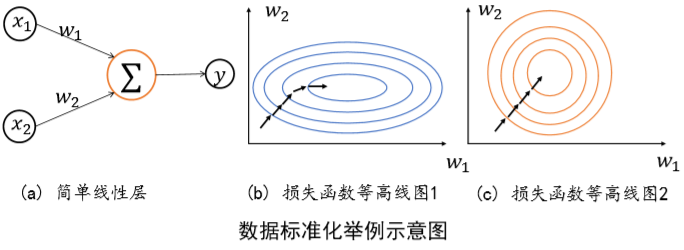
对于如下的线性模型。



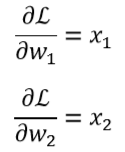
讨论如下 2 种输入分布下的优化问题：

1. 输入𝑥1∈[1,10]，𝑥2∈[1,10]
2. 输入𝑥1∈[1,10]，𝑥2∈[100,1000]

对这两种场景，函数的损失等高线图，图(b)是𝑥1∈[1,10]，𝑥2∈[1,10]时的某条优化轨迹线示意，图(c)是𝑥1∈[1,10]，𝑥2∈[100,1000]时的某条优化轨迹线示意，图中的圆环中心即为全局极值点。

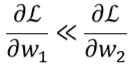


由于：



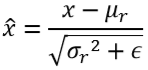
当𝑥1、𝑥2输入分布相近时，、偏导数值相当，函数的优化轨迹如上图(c)所示；

当𝑥1、𝑥2输入分布差距较大时，比如，则：



此时，损失函数等势线在𝑤2轴更加陡峭，某条可能的优化轨迹如上图(b)所示。对比2条优化轨迹线可以观察到，𝑥1、𝑥2相近时收敛更加快速，优化轨迹更理想（如上图c所示）。

通过上述的2个例子，我们能够经验性归纳出：网络层输入𝑥分布相近，并且分布在较小范围内时(如0附近)，更有利于函数的优化。数据标准化就是用于把输入输入𝑥变换为输入相近的一种方法，其可公式化描述为：



其中来自统计的所有数据的均值和方差，𝜖是为防止出现除0错误而设置的较小数 字，如1e – 8。

在基于Batch的训练阶段，可以把Batch 内部数据的均值和方差近似的看作，同时记录每个batch的和，从而计算在整个训练集的，在测试阶段需要使用整个训练集的，而不能使用某个Batch的均值和方差。

在标准化阶段不会引入新的待优化的变量，、和都由统计得到，不需要参与梯度更新。实际上，Batch Norm的作者为了提高Batch Norm的表达能力，还引入了“scale and shift”技巧，将𝑥 ̂变量再次映射变换：



其中𝛾参数实现对标准化后的𝑥 ̂再次进行缩放，𝛽参数实现对标准化的𝑥 ̂进行平移，𝛾、𝛽参数均由反向传播算法自动优化，实现网络层“按需”缩放平移数据的分布的目的。

参考：

龙书，TensorFlow深度学习。