

正则化策略 \Rightarrow 减少测试误差 / 误化误差

数据预处理

① 制除错误数据

- 对已知样本进行简单处理、变换，从而增大数据集，额外生产假训练数据，减小过拟合风险。
- 就像旋转/裁取/扭曲。

- 和全训练数据相比，提供的信息没那么多，且需要算法验证得到的样本依然表示原对象。

- 优点是简单，成本低。

L1/L2 正则化

L1和L2是最常见的正则化方法。

它们在代价函数 (cost function)

中增加一个正则项，由于添加了这个正则项，权重矩阵的值减小。

具有更小权重矩阵的神经网络导致更简单的模型。因此它可在一定程度上减少过拟合。

$$J(w) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m} ||w||_2^2$$

这里的 λ 是正则化参数，它是一个需要优化的超参数。

L1正则化

$$J(w) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m} ||w||_1$$

这里，我们惩罚权重矩阵的绝对值。L1对于过拟合很有用。其它情况下，一般选择优先选择L2正则化。

$$J(w) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m} ||w||_2^2$$

这里的 λ 是正则化参数，它是一个需要优化的超参数。

L1正则化

$$J(w) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m} ||w||_1$$

这里，我们惩罚权重矩阵的绝对值。L1对于过拟合很有用。其它情况下，一般选择优先选择L2正则化。

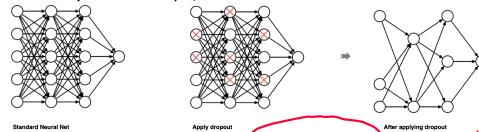
早停

当验证集性能越来越差时，性能不再提升时立即停止，代价函数可能不够小。



Dropout 随机失活 \Rightarrow 最常用

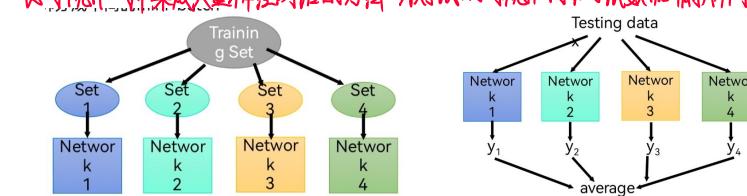
每个迭代过程中随机选择某些节点删除前后连接



训练阶段：学习参数开始之前 \Rightarrow 每个神经元以概率失活重新获得新网络结构，再使用新网络结构训练。
对每一批数据先恢复所有神经元再重新选择失活神经元。

测试阶段：所有权重乘上 $1 - p$ (无dropout)

权可视作一种集成大量神经网络的方法，测试时可视作将测试数据输入不同网络得输出再平均。



集成模型一般优于单一模型，可捕获更多的随机性。
dropout 也使神经网络模型优于正常模型。

如何缓解过拟合问题

Dropout 可以看作是集成大量神经网络的方法

不同的网络可能产生不同的过拟合，取平均则有可能让一些“相反”的拟合互相抵消

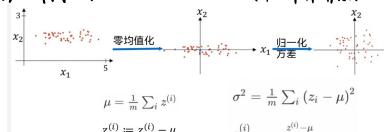
Dropout 能够减少神经元之间复杂的共适应 (co-adaptation) 关系

Dropout每次丢弃的神经元是随机选择的，网络权值的更新不会依赖于隐藏层之间的固定关系即网络中每个神经元不会对另一个特定神经元的激活非常敏感。这使得网络能够学习到一些更加泛化的特征。

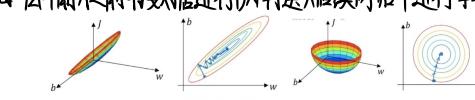
* dropout+maxout
drop connect
annealed dropout
standout

Batch Normalization 批归一化

利用网络训练时的一个 minibatch 数据计算输入的均值和方差，归一化后重构



在每一层输出之前将数据进行BN再送入后端网络中进行学习



从计算结果的分布对激活函数很重要 \Rightarrow 对数据值分布集中于要特征的数据能有效改善



函数增长最快的方向

~~梯度下降法~~ 利用损失函数梯度寻找最小损失函数的方法

梯度方向 $w_{ij}^{new} = w_{ij}^{old} - \eta \frac{\partial E}{\partial w_{ij}}$ 学习率 η 太大跳过最优点 \Rightarrow 动态调整

样本 $x^{(i)}, y^{(i)}$ 的增长 \Rightarrow 整个数据集上损失

梯度下降法 定义

缺点

优点

$\theta = \theta - \eta \Delta \theta$ (BGD)

梯度下降法 (批量梯度下降)

$\theta = \theta - \eta \Delta \theta$ (SGD)

梯度下降法 (随机梯度下降)

$\theta = \theta - \eta \Delta \theta$ (MBGD)

梯度下降法 (小批量梯度下降)

① 梯度台阶困难

② 不反映所有特征

③ 易被困于鞍点

收敛更快

利用高维优化矩阵加速度简过程

SGD, BGD, MBGD

梯度

所有特征

全局最佳

全局最佳