## 一 理解欠拟合-过拟合

欠拟合与过拟合是机器学习的永恒的话题。我们评价一个模型的优劣，从根本上来讲，其实就是要看这个模型平衡欠拟合与过拟合的能力。

欠拟合与过拟合说的是模型的容量、数据模式的复杂度以及数据量之间的比较关系对模型的泛化性能的影响。

如果模型容量比数据模式的复杂度小很多，也就是说，模型的“表征能力”比较弱，则模型无法完全或接近完全地学到数据中的规律（举例：用线性函数拟合二次规律的数据），这就造成欠拟合。

如果模型容量比数据模式的复杂度大很多，也就是模型过参数化，这时模型经过充分学习（使训练集上的准确率非常高）后，相当于把数据中的扰动当做一种规律学到了（举例：用高次函数去拟合线性规律的数据），这就导致模型泛化性变差。

这里需要仔细理解“扰动”的含义。所谓“扰动”，就是说我们拿到的数据，都可以被认为是从一个“总体函数”中采样得到的。这个“总体函数”的均值就是我们梦寐以求的“God model”，但是对每个输入，其输出毕竟是一个随机变量，其服从一个分布，并且有“方差”的概念，也就是说，我们采集的数据必然可能会偏离God model，这就是“扰动”的含义。

（注意，并不是说数据越少，扰动信息越多，数据越多，扰动信息越少。而是说，模型容量相对于数据容量越大，越容易学到数据中的扰动信息（无关乎多少），注意这是由相对容量决定的）

欠拟合-过拟合问题有些类似于求解多元线性方程组的问题——变量的个数跟方程个数之间的比较关系决定了方程组是欠定还是超定。

但是注意！模型的欠拟合与过拟合还跟数据量有关。当模型的容量和数据规律的复杂度确定时，数据量越大，模型越不容易欠拟合或者过拟合——也就是说，若模型欠拟合，则数据量越大，欠拟合的程度越低；若模型过拟合，则数据量越大，过拟合的程度越低。

何以如此呢？我们拿过拟合问题来分析。直觉上看，数据越多，模型的过参数化就越不严重，因此越不易过拟合。从另外一方面来说，数据越少就意味着关于真实的总体分布的信息越少，数据越多就意味着关于真实的总体分布的信息越多。而模型做的事情可以看成是从数据中获取信息。所以数据越多，模型就越能接近真实的总体分布。

举例来说的话，就是二次函数去拟合线性规律的数据：少量数据时，二次项系数比较大；大量数据时，二次项系数比较小，甚至接近0。

一般来谈的话，首先，扰动信息指的是数据集中的样本偏离真实的拟合关系的程度。由于数据集是从真实的拟合关系对应的分布采样出来的，因此必然存在扰动。

模型的目标一般是最小化数据集上的所有样本的预测损失（如MSE），是要考虑所有的样本的一个整体情况。模型容量相对于数据规律的复杂度比较大时，容易比较精确地对训练集所有样本进行预测，因此容易学到数据上的扰动信息；当模型容量相对于数据规律的复杂度比较小时，模型无法比较精确地对所有样本进行预测，只能尽量将整体损失降到最小。那么关键问题来了，这时，怎么知道模型学到的不是扰动，而是更接近真实拟合关系呢？ 因为数据是从真实拟合关系（God model）对应的分布采样出来的，因此真实拟合关系能够使整体损失最小，所以模型会尽量地接近真实拟合关系。

从这里可以看出，我们如此努力地避免过拟合，其实就是为了接近真实的拟合关系。因为越接近真实的拟合关系，模型的泛化性能才会越好。

在机器学习中，树模型和神经网络模型的容量都能搞得很大，因此欠拟合一般都不是问题，在实际中主要关心过拟合的问题。

## 二 如何避免过拟合

模型的过拟合跟模型的结构、模型的策略（目标函数）、模型的训练过程等都有关系。

### 1.数据

从数据角度来看，数据量越大，越不易过拟合（当然，要保证数据的质量，即数据的分布尽可能地与真实分布一致）。因此可以通过增加数据量来缓解过拟合。比如多采样一些数据，或者使用数据增强。

### 2.模型

(1) 适当减小模型的容量（比如减小模型的参数量，具体来说，决策树的剪枝、神经网络的weight tying等）。

但是注意，理论上来讲，减小模型的容量并不绝对能缓解过拟合，因为减小模型的容量，相当于减小了模型的参数空间，但减小之后的参数空间中的参数，可能都跟真实拟合关系的参数相去甚远。所以还要看具体如何减小模型的容量。

(2) 选取更优秀的模型结构（比如，集成树模型比单棵决策树模型更不容易过拟合）。

### 3.训练过程

(1) 提前终止

提前终止是为了避免模型在训练集上充分学习，以此减轻过拟合。

(2) 交叉验证

交叉验证是为了选取在该问题上泛化性能比较好的超参数，以此减轻过拟合。

### 4.正则化

正则化是机器学习中的一个重要的研究课题。

狭义的正则化是指在损失函数上加上正则化项（或者说约束项，或者说罚项），构成新的目标函数。这样可以使模型不仅向减少损失函数的方向优化，还向正则化的目标方向进行优化，从而使学到的模型不仅仅是最小化训练集上的损失，还兼顾了正则化的目标。

而广义的正则化，不一定是通过在损失函数上加正则项的方式实现，也可以有其它的方式，但目的都是让模型不局限于关注训练集上的数据的信息，而是还考虑“正则”所提供的信息，从而提升泛化性能。如神经网络中的BN，也具有正则化的作用。

正则化的本质是给模型融入“先验信息”。

我们知道，有监督学习，简单地说就是拟合数据中的关系。我们采集到的数据必然是有限的、离散的，而我们要拟合的关系通常都是无限的、连续的。如果没有任何先验信息，那么我们就完全无法对训练集之外的数据作出预测。而“模型”本身就提供了一种先验信息（更正式的说法是“归纳偏好”，具体来说，这里的先验是指，模型是对总体分布的一种“描述”，或者说它给出了总体分布的“形式”，并给定一些未知参数，告诉人们：我这些参数中必然有一组能够给出真实的分布，我认为真实分布对应的那组参数就是能对训练集有最好预测结果的那组参数。）。有了模型提供的先验信息，我们就可以对任意数据作出预测。而正则化是在模型上再加上一些先验信息的有效手段。

那么正则化为什么能够缓解过拟合呢？

一般来讲，正则化跟防止过拟合并没有必然的联系。如前所述，“缓解过拟合”的本质目标，是让模型“接近真实拟合关系”；而正则化是给模型添加先验的一种手段。因此，如果先验信息是符合实际情况的，那么就能使模型更接近真实的拟合关系。

坊间流传的正则化能够减小过拟合的原因——是因为正则化减小了模型的容量或者减小了模型的参数空间，这我不敢苟同。因为一般来讲，正则化并不一定能够减小模型参数空间，而且上文已经谈到过，即使减小模型的参数空间，也不一定能起到缓解过拟合的效果。

所以关于正则化缓解过拟合的原因，还是应该从正则化的真正作用——加先验来解释。我们给模型加正则化，一般都是考虑了真实拟合关系的特点，因此能够使加了正则化后的模型更接近真实的拟合关系。

从另一个略有差异的角度考虑，前文说过，过拟合的模型容易学到训练数据中的扰动信息，这是因为模型的容量相对于数据的规律的复杂度来说过大。实际上也可以这样理解：模型容易学到数据中的扰动信息，也是因为关于真实拟合关系的先验信息太少。从这个角度来看，正则化施加了有效的先验，就能让模型尽量少学到训练数据中的扰动信息，从而缓解了过拟合。

关于“减小模型参数空间未必能防止过拟合”，以下有一个例子，可以用来思考：

比如对于一堆线性规律的数据，用二次函数进行拟合，这就过拟合了。这时考虑加正则化，减小模型的参数空间。

如果用L1正则化，很大可能把二次项系数给变成0，这就起到避免过拟合的作用；但是用L2正则化，只是让系数变小，但还是二次函数，这样也能缓解过拟合吗？这需要仔细思考。。