综上分析可知，我们对直接法计算规则进行了改善，在工程力学教学过程中，使得我们对静力学和材料力学未知力的计算有了一个比较统一的计算规则和简便计算方法，即等号右侧已知力（或力矩）的和，对正负号的规定，采用了“同方向（转向）为负，反方向（转向）为正”确定方法[1]，统一确定直接法计算未知力的正负号规则，扩大了计算未知力的范围，和以往工程力学教学内容比较，对工程力学中未知力计算的直接法，不仅仅是停留在对梁的弯矩和剪力的计算，采用此法之后，对约束力（支座反力）、轴力、扭矩的计算，直接法都是简洁而适用的。通过这样的教学方法，极大地改善了教学效果，提高了学生对工程力学计算掌握的能力，使得复杂问题得到简化，有效提高解决问题的速度和正确率。

参考文献

[1]张慧珍.工程力学I.武汉：武汉理工大学出版社，2016年

[2]杨力彬，赵平.建筑力学I.北京：机械工业出版社，2004年

[3]王赞.“同为负反为正”在工程力学计算中应用[J].工程技术（文摘版）·建筑，2016（7）：00064-00064.

作者简介：

王赞（1969.12-），男，汉族，内蒙古人，硕士，副教授，从事工程力学，建筑力学，普通物理，建筑装饰材料与施工等课程教学与研究。

统计机器学习中的过拟合问题

陈克根

（十堰市职业技术（集团学校） 湖北 十堰 442000）

[摘 要]在20世纪80年代的时候，符号学习还是机器学习的主流，而自从20世纪九十年代以来，就一直是统计机器学习的天下。机器学习从纯粹的理论研究和模型研究发展到了解决现实生活中实际问题为目的的应用研究。机器学习算法赋予了机器学习的能力，学习的好坏可以用预设的误差函数来衡量。机器学习中可能存在过拟合或者欠拟合的问题，这影响着机器学习算法的好坏，本文对过拟合问题进行了研究，分析出了两种可以解决过拟合问题的方法。分别是使用充分的数据集，和使用合适复杂度的函数集。通过对统计机器学习基础理论的研究，得出了有关机器学习模型的建立的启示，在以后设计机器学习模型时可以作为参考。

[关键词]人工智能；统计机器学习；过拟合；欠拟合

【DOI】10.12252/j.issn.2096-627X.2019.11.1103

一、统计机器学习问题

机器学习是一门，致力于研究如何通过计算的手段，利用经验来改善系统自身的性能。在计算机系统中，“经验”通常以“数据”形式存在，因此，机器学习所研究的主要内容，是关于在计算机上从数据中产生“模型”的算法（“模型”泛指从数据中学得的结果），即学习算法。有了学习算法，再把经验数据提供给学习算法，它就能基于这些数据产生模型；在面对新的情况时，模型会给我们提供相应的判断。如果说计算机科学是研究关于“算法”的学问，那么机器学习就是研究关于“学习算法”的学问。

二、机器学习模型

要进行机器学习，首先要有数据。每一条数据相当于一条记录，这组记录的集合称为一个数据集，其中每条记录是关于一个事件或对象的描述，称为一个示例或样本。反映事件或对象在某方面的表现或性质的事项称为属性或特征。属性的取值称为特征值。属性张成的空间称为属性空间、样本空间或输入空间。每一个数据记录就在属性空间中对应一个坐标点。

其次，机器学习只有示例数据是不够的，要判断一个模型的好坏，还需要训练样本的结果信息即标签。拥有了标记信息的示例则称为样例，所有标记的集合称为标记空间或输出空间。

最后，机器学习算法要学得的，是反映数据集关系的假设。从数据中学得模型的过程称为学习或训练，这个过程通过执行某种学习算法来进行，使用训练数据学得的模型对应了关于数据的某种潜在的规律，因此被称为假设。我们可以把学习过程看作一个在所有假设组成的空间中进行搜索的过程，搜索的目标是找到与训练集匹配（拟合）的假设。

机器学习的目标是使学得的模型很好的适用于新样本，而不是仅仅在训练样本上工作得很好，训练样本只是全体样本空间的一个很小的采样。这种学得模型适用于新样本的能力，称为泛化能力，具有强泛化能力的模型能很好地适用于整个样本空间。于是，尽管训练集只是样本空间的一部分，算法仍然要求它能反映整个空间的特性，否则很难在训练集上学得在整个模型上效果好的模型。

三、过拟合与欠拟合

以机器学习的分类学习为例，把分类错误的样本数占样本总数的比例称为错误率，即如果在m个样本中有n个样本分类错误，则错误率$E=n/m$。学习器的实际预测输出与样本的真实输出之间的差异称为误差。在训练集上的误差称为训练误差或经验误差，在新样本上的误差称为泛化误差。显然，目标是获得泛化误差小的学习器。但是学习器事先并不知道新样本是什么样，实际能做的是努力使经验误差最小化，在很多情况下，我们可以学得一个经验误差很小的模型，甚至在样本上分类全部正确，错误率为零，但这样的模型多数情况下都不好。

实际希望是在新样本上能表现很好的学习器，为了达到这个目标，应该从训练样本中尽可能学出适用于所有潜在样本的普遍规律，这样才能在遇到新样本时做出正确的判别。如果学习算法对训练样本太匹配，很可能就把训练样本自身的一些特点当作了所有潜在样本都有的一般性质，这样就会导致泛化性能下降。这种现象在机器学习中称为过拟合[1]，与过拟合相对的是欠拟合。

经过理论论证，有多种因素导致过拟合问题。最常见的原因是，训练集的数量级和模型的复杂度不匹配，训练集的数量级小于模型的复杂度。比如要拟合一个二阶多项式ax^2+bx+c，只使用两个坐标点$(x_1,y_1),(x_2,y_2)$是不能反映出二阶多项式的两个零点位置信息的，也就无法学习出真实的二阶多项式系数。还有的原因就是训练集和测试集的特征分布不一致，无法反映整个样本空间。或者样本里存在噪声数据，干扰过大以至于模型记住了噪音特征，忽略了真实的输入输出间的关系。

四、避免过拟合的方法

首先，最直接的办法是数据增强。理论上讲，所有的过拟合无非是训练数据的缺乏和训练参数的增加。想要获得更复杂的模型，需要更多的参数，现在的神经网络模型因此也越来越深。但是训练样本的特征多样性如果无法表示出多样性，再多的训练参数也毫无意义，反而会造成过拟合。训练的模型泛化能力也会很差。大量数据带来的特征多样性有助于充分利用所有的训练参数。数据增强的手段一般有：

1.收集更多数据；

2.对已有数据加入高斯噪声，可能具有较低的失真水平；

3.使用条件对抗神经网络（Conditional GANs）来产生对抗数据。

其次，可使用提前终止方式训练模型。提前停止其实是另一种正则化方法，在训练集和验证集上，一次迭代之后计算各自的错误率，当在验证集上的错误率最小，在没开始增大之前停止训练，因为如果接着训练，训练集上的错误率一般是会继续减小的，但验证集上的错误率会上升，这就说明模型的泛化能力开始变差了，出现过拟合问题，及时停止能获得泛化更好的模型。

最后，在神经网络中最流行的dropout法，dropout法的简单阐述就是在神经网络的神经元之间传播时，让某个神经元以一定的概率停止工作，可以使模型的泛化能力更强。因为dropout使得模型不会太依赖某些局部的特征。例如下图，一个完整的三层的全连接网络中，虚线代表临时dropout的神经元，激活值只从实线传播。

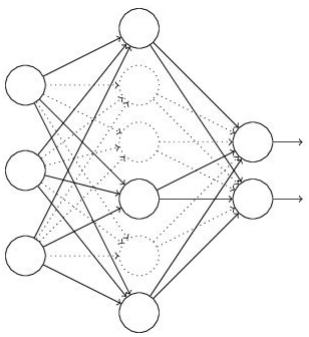

图1 dropout示意图

五、结论

能够避免拟合噪声的机器学习是健壮的算法。过度拟合和欠拟合是一个根本问题，即使是经验丰富的数据分析人员也会感到不满意。有些模型看起来不错，但问题是他们甚至从未使用过测试集，更不用说验证集。

过拟合的方法总结下来就是，增强数据和减少不必要的模型特征多样性。在往后的机器学习中，多使用减轻过拟合的方法，使用测试集、验证集对算法进行测试。这样才能做出鲁棒的机器学习算法。

参考文献

[1]Shalev-Shwartz, Shai.Understanding Machine Learning: From Theory to Algorithms[M].Cambridge University Press, 2014.