模型优化的风向标：偏差与方差

原创 石头 机器学习算法那些事 2018-10-13

前面讲到测试集的泛化性能是衡量学习模型优劣的金标准，实际工作中，我们会遇到两个不能忽视的问题：

（一）训练数据集D给定的情况下，不同的测试集T会有不同的测试精度；
（二）不同训练数据集D构建的最优模型f不同。

因此，仅仅通过测试误差来评价模型的泛化性能是存在偏差的。已知训练数据集D和测试数据集T的概率分布，用测试数据集的期望泛化误差来评价模型的泛化性能模型是最佳评价方案，其中，测试数据集期望泛化误差包括偏差，方差和噪声。作者认为偏差与方差的最重要应用是对自己设计的学习模型有一个更深入的了解，并指导自己怎么去优化学习模型。

本文首先介绍了期望泛化误差的推导过程，并引入了偏差与方差的定义；然后介绍了偏差与方差的应用以及偏差与方差的估计方法。最后对本文进行总结。

1、期望泛化误差

测试集的期望泛化误差是评价模型泛化性能的金标准，已知训练数据集D的概率分布，对测试样本**x**，令y_D为**x**在数据集中的标记，y为**x**的真实标记，f(**x**;D)为训练集D上学得模型f在**x**上的预测输出，不同的训练集D构建不同的最优模型f。
以回归任务为例，模型的泛化能力用E(f;D)表示。
学习算法的期望预测为：

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)] \ ,$$

使用样本数相同的不同训练集产生的方差为：

$$var(\boldsymbol{x}) = \mathbb{E}_D \left[\left(f\left(\boldsymbol{x}; D\right) - \bar{f}\left(\boldsymbol{x}\right) \right)^2 \right]$$

噪声为：

$$\varepsilon^2 = \mathbb{E}_D\left[(y_D - y)^2\right]$$

假定噪声期望为零：

$$\mathbb{E}_D[y_D - y] = 0$$

期望输出与真实标记的差别称为偏差（bias），即

$$bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$$

期望泛化方差分解：

$$\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&\quad + \mathbb{E}_D\left[2\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y + y - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)^2\right] + \mathbb{E}_D\left[\left(y - y_D\right)^2\right] \\
&\quad + 2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)\left(y - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \left(\bar{f}(\boldsymbol{x}) - y\right)^2 + \mathbb{E}_D\left[\left(y_D - y\right)^2\right],
\end{aligned}$$

由最后的等式可知：

$$E(f;D) = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \varepsilon^2$$

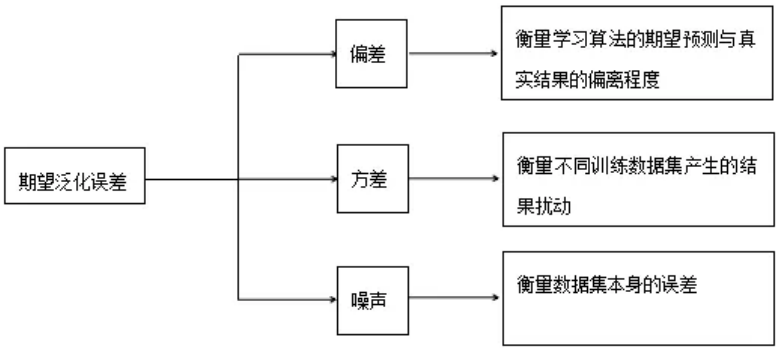
也就是说，泛化误差可分解为偏差、方差与噪声之和。

期望泛化误差 = 偏差 + 方差 + 噪声

偏差表示学习算法的期望预测与真实结果的偏离程度，即刻画了学习算法本身的拟合能力。训练模型越复杂，则偏差越小，如线性回归可以通过增加参数个数来提高模型的复杂度，神经网络可以通过增加隐单元个数来提高模型的复杂度。

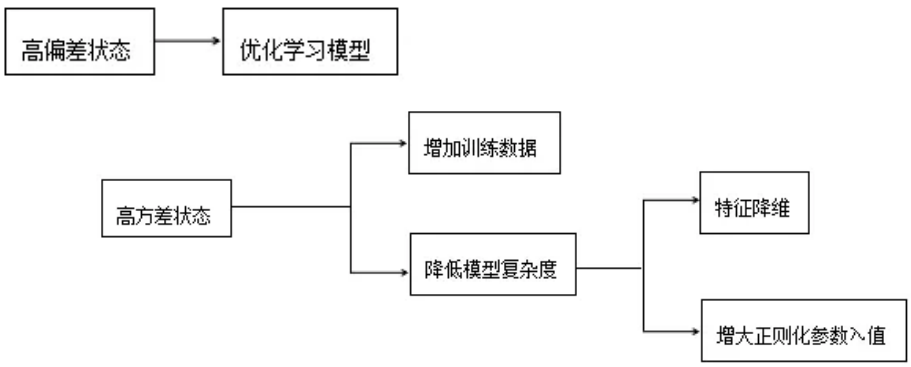
方差表示同样大小的训练数据集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响。若模型处于过拟合状态，不同训练数据集产生的学习模型相差较大。

噪声表示任何学习模型所能达到期望泛化误差的下界，即刻画了学习问题本身的难度。若噪声比较大，即使是最优模型，泛化误差也比较大。

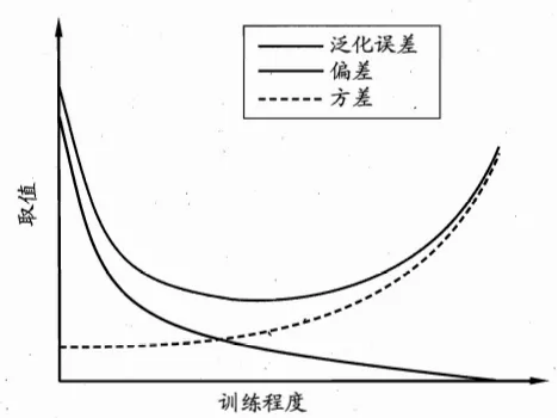


2、偏差-方差应用

我们在设计学习模型时需要对该模型有一个清晰的思路，若当前学习模型处于高偏差状态，则需要优化学习模型，增加训练数据并不能优化模型；若当前学习模型处于高方差状态，不同的训练数据集构建的模型有比较大的差异，需要增加训练数据或降低模型的复杂度（如下图）。



偏差与方差是一对矛盾的量，称偏差-方差窘境（bias –variance dilemma）。当学习模型的训练程度低时，学习器的拟合能力不强，模型处于高偏差状态，训练数据集的扰动不足以使学习器发生显著变化，模型具有低方差属性；当模型训练程度高时，模型复杂度高，处于低偏差状态，训练数据集的扰动使学习器发生了显著变化，模型具有高方差属性。



3、偏差与方差估计

随机抽样的训练数据集是有限的，按照第二节偏差与方差的公式是无法计算偏差与方差，因为无法知道训练数据集D和测试数据集T的分布情况，只能通过现有的训练数据集来估计偏差与方差。本文介绍了两种偏差与方差估计方法，第一种是对训练集分成多组的偏差与方差估计，第二种是对训练集未分组的偏差与方差估计，一般采用第二种偏差与方差的估计方法。

3.1 训练数据集分组的偏差与方差估计

我们处理的数据都是随机抽样的样本数据，期望是均值的无估计偏差，若对采样的数据进行分组，可以估计模型的偏差与方差。抽样数据集共分为L组，每组抽样数据集包含N个采样点，由上节的偏差与方差公式，可推算偏差与方差的公式如下：

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N}\{\overline{y}(x_n) - h(x_n)\}^2$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \overline{y}(x_n)\right\}^2$$

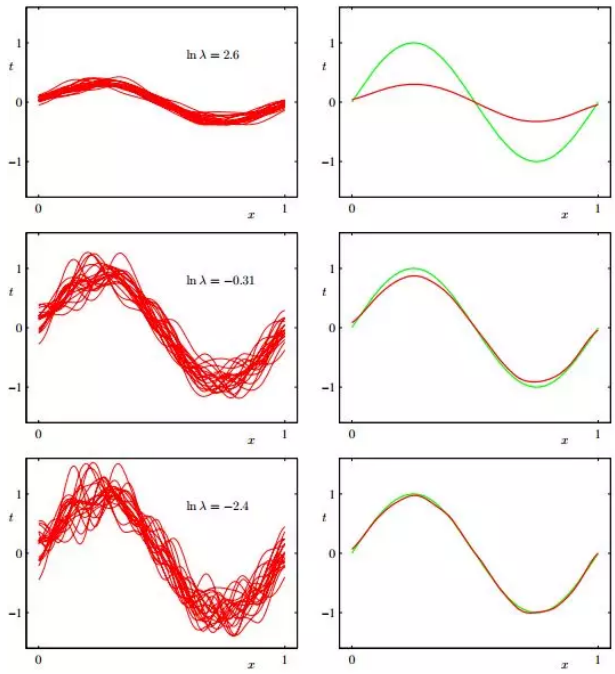
其中，期望预测：

$$\overline{y}(x) = \frac{1}{L}\sum_{l=1}^{L}y^{(l)}(x)$$

h(xn)为样本输入为的真实标记。

【例】拟合正弦函数h(x)=sin(2*pi*x)，对正弦函数独立采样100组数据集，每组数据集包含25个点，模型采用正则化的多元线性回归拟合，假设线性回归参数个数确定，求正则化参数λ与偏差、方差的关系。

解：正则化参数λ与偏差、方差的关系如下图：



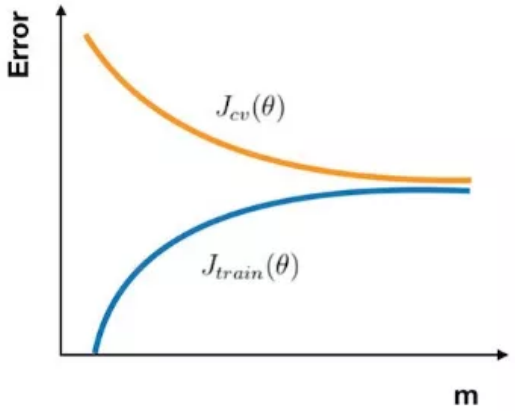
如上图所示，左图表示100组数据集的拟合情况，可评估模型的方差。右图是对左图100组数据集每点求平均后的数据拟合情况，绿线表示理论模型，红线表示数据集平均后的结果，可评估模型的偏差。

由图可知，方差与偏差是一对矛盾量，λ越大，模型复杂度越低，偏差越大，方差越小；反之，偏差越小，方差越大。

3.2 训练数据集不分组的偏差与方差估计

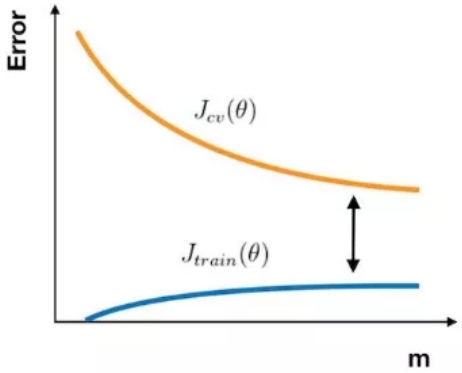
为了构建最优模型，尽可能使用更多的训练数据。在实际工作中，训练数据集不分组，对一组训练数据集构建最优学习模型并评估模型的偏差与方差。
学习曲线可以估计学习模型的偏差与方差，训练数据集划分为训练集和验证集，学习曲线是衡量样本数与训练集误差、交叉验证集误差的关系，横坐标是样本数，纵坐标是训练数据集误差$J_{train}(\theta)$和交叉验证集误差$J_{cv}(\theta)$。

高偏差：



在高偏差的情形下，样本数增加，训练集误差和交叉验证集误差十分接近，但是误差很大。因此，当模型处于高偏差的情形下，首先考虑的应该是优化模型。

高方差：



在高方差的情形下，样本数增加，训练集误差缓慢增加，交叉验证集误差减小。因此，当模型处于高方差的情形下，增加样本数可能会减小交叉验证集误差。

4、总结

测试数据集的期望泛化误差是衡量模型性能的金标准，期望泛化误差包括偏差、方差以及噪声。模型处于高偏差低方差的状态时，则需优化模型；模型处于低偏差高方差的状态时，则需增加训练数据或降低模型复杂度。因此，判断当前模型处于何种状态是进一步优化模型的前提，如何判断可参考第三节。

参考：

周志华 《机器学习》

Christopher M. Bishop <<Pattern Recognition and Machine Learning>>

https://blog.csdn.net/hertzcat/article/details/80035330

END

长按二维码关注

机器学习算法那些事
微信：beautifulife244

砥砺前行 不忘初心