Lecture 13　Hazard of overfitting

整理者: LobbyBoy*　2020年2月26日

1. What is Overfitting?

先看一个关于bad generalization的例子。我们的训练数据集由5个样本组成，输入空间与输出空间均为一维实数域，因此我们要解决的是一个回归问题。这5个样本数据是由某一个二阶多项式+noise产生的，即：

$$y_n = ax_n^2 + bx_n + c + noise, \quad n = 1, 2, 3, 4, 5, \quad a \neq 0$$

假设我们使用4阶多项式转换来拟合这一组数据，则一定可以在$\mathcal{Z}$空间中找到一个hyperplane，恰好通过这5个样本(设计矩阵可逆)。对应回$\mathcal{Z}$空间，也就是能够找到一条4次曲线，通过所有样本点。这样的hypothesis的$E_{in} = 0$。如下图所示，蓝线是我们希望学到的function，红线是我们用4th order polynomial transformation学到的hypothesis：

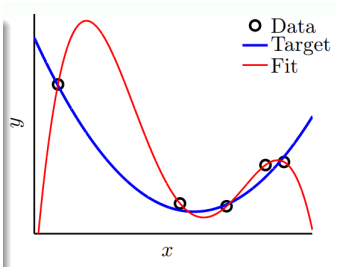

图 1: Bad generalization

肉眼可见红色曲线的E_{out}一定很大。对于这种E_{in}很小，E_{out}很大的情况，我们将之称为“Bad generalization”，即训练出的hypothesis虽然拟合训练数据的能力很好，但是泛化到新数据上的能力很差。

下面这张图我们已见过多次。当假说集合的VC dimension为d_{VC}^*时，泛化上界最小。当我们选择的假说集合过于复杂时，如$d_{VC} = 1126$，那么我们学到的hypothesis即使具有很小的E_{in}，但泛化上界却会很大，因此称之为bad generalization。当我们从$d_{VC} = d_{VC}^*$切

*本笔记根据台湾大学林轩田教授于线上教育平台Coursera开设的“机器学习基石”课程整理而成(课程内容见: https://www.coursera.org/learn/ntumlone-mathematicalfoundations/home/welcome)。笔记内的大多数图片来自于林老师的课程slides。感谢林老师能够将如此精彩的课程通过线上平台同所有人分享，thanks!

换到$d_{VC} = 1126$时，发生的事情是E_{in}降低+E_{out}上升，我们将这一过程称为“overfitting”(过拟合)；当我们从$d_{VC} = d^*_{VC}$切换到$d_{VC} = 1$时，发生的事情是E_{in}上升+E_{out}上升，我们将这一过程称为“underfitting”(欠拟合)。

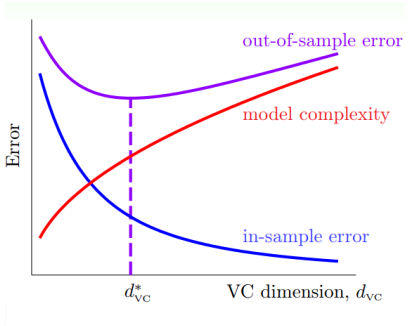

图 2: VC-Error

总而言之，bad generalization是一种状态，即E_{in}低而E_{out}高的一种状态；而overfitting是一个过程，该过程中E_{in}下降但E_{out}却在上升。

最后，我们将发生“过拟合”比喻成现实中“出车祸”。出车祸的原因可能有很多种：第一，可以是车速太快，这对应于使用大VC的假说集合容易发生过拟合；第二，可能是道路太颠簸，这对应于训练数据的噪声太大；第三，可能是对路况的观察与熟悉不够，这对应于训练数据太少。在下一讲中，我们主要探讨训练数据中的噪声会对训练产生怎样的影响。

2. The Role of Noise and Data Size

本节我们将探讨样本量较小时的情形。依然是1维回归问题，考虑两组样本：第一组样本的输出来自10次多项式+一些噪声，第二组样本的输出来自50次多项式且无噪声，如下图所示：

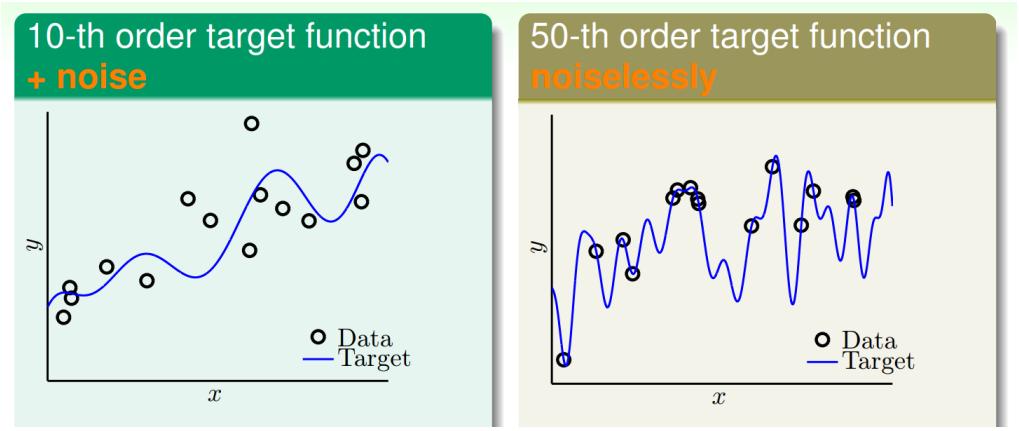

图 3: Two groups of sample

首先，我们分别用2次多项式转换与10次多项式转换拟合第一组样本(10th poly+noise)。结果如下图所示：

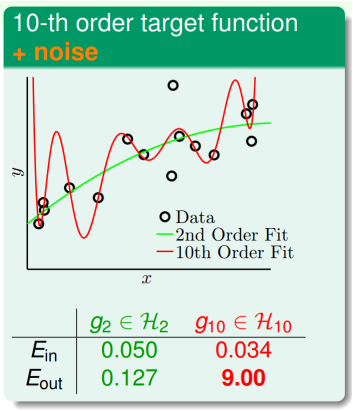

图 4: 10th poly+noise

上图中绿色曲线表示用2次转换拟合的结果，红色曲线表示用10次转换拟合的结果。结合图与表格中的数据，我们知道，从2次转换switch到10次转换，发生了过拟合，即经验误差下降而泛化误差上升。

接着我们分别用2次多项式转换与10次多项式转换拟合第二组样本(50th poly)。结果如下图所示：

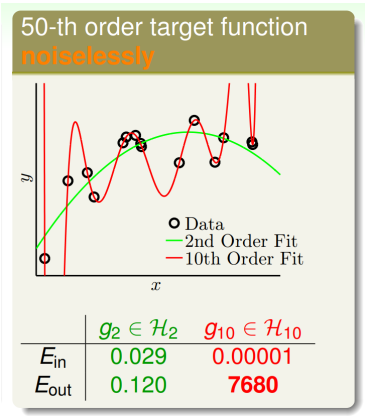

图 5: 50th poly

上图中绿色曲线表示用2次转换拟合的结果，红色曲线表示用10次转换拟合的结果。结合图与表格中的数据，我们知道，从2次转换switch到10次转换，仍然发生了过拟合，即经验误差下降而泛化误差上升。

综上，在两个问题中，表现较好的假说集合均为更简单的$\mathcal{H}_2$。这好像有悖于直觉：因为第一组样本中的输出就是10次多项式产生的，为什么用10次多项式的假说集合去拟合它效果反而更差？如果是因为有noise的原因的话，那么为什么在第二组没有noise的样本中，10次多项式的拟合效果依然很差？

我们首先回答第一个问题：为什么用10次多项式拟合第一组样本效果更差。我们可以画出两个假说集合的learning curve(回忆regression那一章中两条曲线的表达式)：

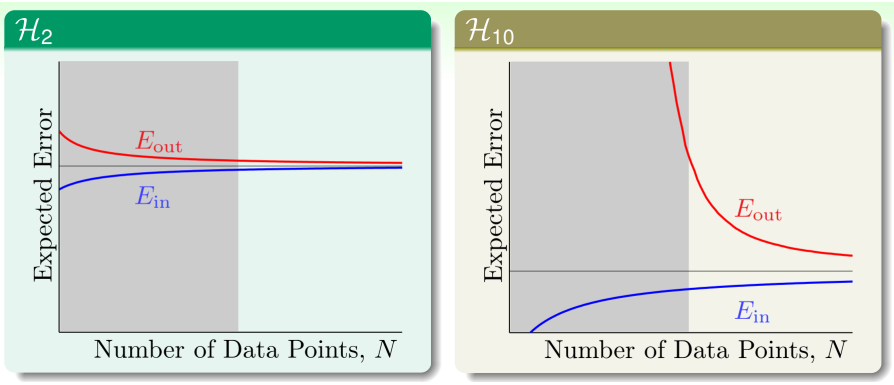

图 6: 50th poly

我们先看两个假说集合学习曲线中经验误差与泛化误差的收敛值(随着样本数增大)。$\mathcal{H}_2$的收敛水平值较大，这是因为target function是10次的，而$\mathcal{H}_2$拟合得再优秀，与10次多项式还是有差距的，因此最后收敛值较大；而$\mathcal{H}_{10}$在样本有很多的时候，因为target function本来就是10次的，所以能够拟合得很好，但不是完美的，因为有noise。但我们之前说过，我们现在的情形是样本数很少，对应于上图中的灰色区域。可见，此时用$\mathcal{H}_{10}$训练出的结果，虽然经验误差很小，但是泛化误差却由于模型复杂度的原因而很大$(1+\frac{1+d}{N})$；而$\mathcal{H}_2$训练出的结果，虽然经验误差较大，但是VC bound较紧，因此泛化误差也大，但是离经验误差不远。在实践中，我们可以将其解释为，模型过于复杂而样本较少，因此模型将noise也拟合得很好，导致泛化误差变大，这就是高复杂度伴随的代价。当然，从学习曲线中可以看出，当样本容量逐渐增大时，泛化误差呈指数级缩减。

至于No noise的第二组样本，为什么10次多项式的拟合效果依然很差？其实是有noise的，但这个noise是来自target complexity——50次目标多项式产生的样本很复杂，就像是noise。

3. Deterministic Noise

对于回归问题，我们假设输入是由一个Q_f次多项式的target function与一个均值为0、方差为σ^2的高斯噪声构成的，即：

$$y = f(x) + \epsilon \quad \sim \quad Gaussian\Big(\sum_{q=0}^{Q_f} \alpha_q x^q, \sigma^2\Big)$$

并将样本量记作N。我们依然有两个假说集合：$\mathcal{H}_2$代表2次多项式的hypotheses，$\mathcal{H}_{10}$代表10次多项式的hypotheses；记$\mathcal{H}_2$的训练结果为g_2，$\mathcal{H}_{10}$的训练结果为g_{10}。我们用：

$$E_{out}(g_{10}) - E_{out}(g_2)$$

度量overfit的程度：该值为正，说明从$\mathcal{H}_2$切换到$\mathcal{H}_{10}$会导致泛化误差的增加，而经验误差一定是下降的($E_{in}(g_{10}) \leq E_{in}(g_2)$)，这就是发生了过拟合。我们分别尝试“控制$Q_f$+变动$N$与$\sigma^2$”和“控制$\sigma^2$+变动$N$与$Q_f$”，得到下面的两幅图：

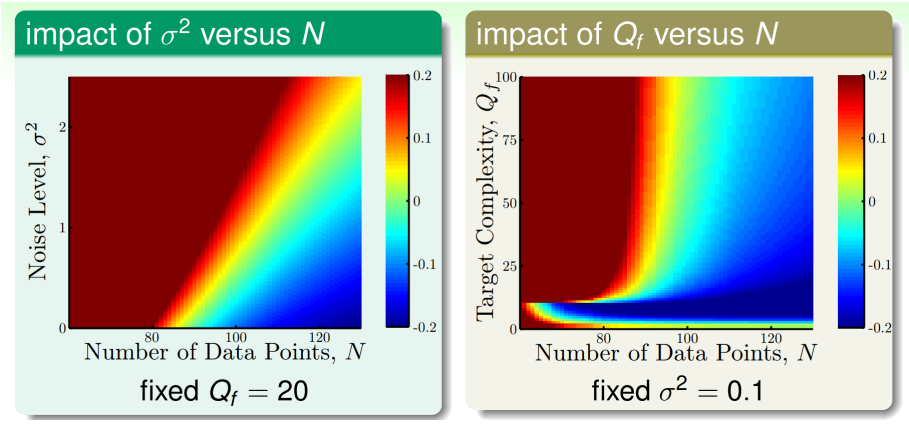

图 7: The results

可以看出，当①样本量小，②随机误差(stochastice noise)大，③确定性误差(deterministic noise)大(目标函数非常复杂)时，从简单model切换到复杂model会发生overfitting。

最后，我们再讨论一下确定性误差(deterministic noise)。例如，f表示一个复杂的目标函数(如50次多项式)，$\mathcal{H}$表示一个简单的假说集合(如2次多项式的集合)，因此用$\mathcal{H}$去学f是一定无法得到$E_{out} = 0$的hypothesis的，因为$f \notin \mathcal{H}$。我们将$\mathcal{H}$中最最接近f的hypothesis记作h^*，则h^*与f的差异就称之为deterministic noise，即下图中的灰色区域：

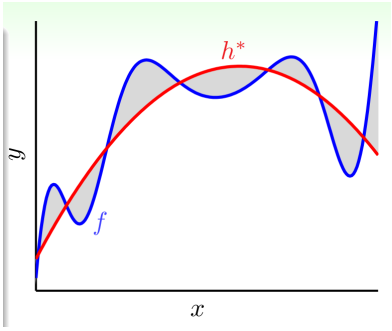

图 8: Deterministic noise - gray area

用deterministic noise解释2次多项式较10次多项式而言对于50次多项式产生的无噪声样本拟合效果更好：对$\mathcal{H}_2$与$\mathcal{H}_{10}$来说，50次多项式都是一个无法完美拟合的目标，因此学习时面临的用deterministic noise很大(即使没有stochastic noise)。在样本量很少的情况下，使用复杂的假说集合，也无法避免这个确定性的噪声，同时还会带来模型复杂度高的代

价。

4. Dealing with Overfitting

本节介绍一些缓解overfitting的方法，都很直观：

learning	driving
overfit	commit a car accident
use excessive d_{vc}	'drive too fast'
noise	bumpy road
limited data size N	limited observations about road condition
start from simple model	drive slowly
data cleaning/pruning	use more accurate road information
data hinting	exploit more road information
regularization	put the brakes
validation	monitor the dashboard

图 9: Dealing with overfitting