偏差与复杂性的权衡

第二章我们看到，除非我们非常谨慎，否则训练集会误导学习者，导致过拟合问题。为了克服这点，我们限制搜索空间到假定集H。这个假定集可以被看做反映了一些学习者对于任务拥有的先验的知识，即一个假设集的成员是该任务的低误差模型。比方说，在我们的苦瓜模型中，根据我们来自其他水果的经验，我们可以假定一些矩形在颜色-硬度平面可以预测苦瓜味道。

先验知识真的对于成功的学习很重要么？是否可能有一类通用的学习方法，可以不需要任何关于任务先验知识，并可以挑战所有的任务呢？让我们详细讨论这个问题。对一个特殊的学习任务（由位置的$X \times Y$的分布D定义）其风险$L_D(h)$足够小。此处，该问题变成了是否存在训练集大小m,学习算法A,使得对于每个分布D，只要A接受m个分布D的实例，就有很高概率能输出有很低风险的预测器h。

该章节第一部分正式引出这个问题。天下没有白吃午餐 定理将会告诉我们，这个通用的学习器并不存在。更准确地说，这个定理说对于二分类问题，对于每个学习器，都会存在一个分布使其（该学习器）失效。我们说，这个学习器失效，当其接受从分布中获取的 独立分布的 例子，它的输出分布有很大风险，如≥ 0.3使得对于同样的分布，存在一个其他的学习器使得能够输出更低风险的假定。换句话说，这个定理说明没有学习者可以胜任全部学习任务，每个学习器都有一些自己不行而其他学习器可以胜任的任务。

因此，当接触特定学习任务时，我们应该有一些对分布D的先验的知识。其中一种先验知识即D来自于一些特定参数。在24章，我们会学到在该假定下的学习方法。另一个类型的先验知识是，我们假定当定义PAC学习模型是存在h在一些先前定义的假定集H，$L_D(h) = 0$。一个更温和的先验知识是假设$\min L_D(h)$非常小。这样，这个弱假设是使用不可知PAC模型的前提，因为我们需要输出假定集的风险不会大于$\min L_D(h)$

在本章节第二部分，我们会学到使用该假定集作为先验知识的好处和坏处。我们将会分解ERM算法的误差为两部分。第一部分，反映了我们先验知识的质量，由最小风险$\min L_D(h)$ 衡量。这部分又叫做*近似误差*，或算法选择假定的*偏差*。第二部分是由过拟合导致的误差，依赖于H的复杂度大小，叫做*估计误差*。这俩属于表明选择一个更复杂H的博弈（减少偏差但可能带来额外的过拟合的风险）抑或更简单的H（增加偏差但可能减少过拟合）。

5.1 天下没有白吃的午餐 定理

这一部分，我们可以证明并不存在万能的学习者。我们将展示没有学习者可以在所有任务上一马平川。

定理5.1 假设学习算法A对于二项分类任务01损失在域X。让训练集的大小m为任意小于X/2的数字，那么，存在分布D在$X \times \{0,1\}$使得：

1. 存在函数$f : X \to \{0,1\}$ 使得$L_D(f) = 0$
2. 有至少1/7的概率选择$S \sim D^m$，我们有$L_D(A(S)) \geq 1/8$

这个定理说明，对于每个学习者，都存在一个任务使其失效，但这个任务可以被其他学习者学习到。事实上，这个成功的学习者即假定集为H={f}的ERM学习者。更广义地说，对于任何有限的假定集，包含f，大小满足$m \geq 8\log(7|H|/6)$

证明：设C为 大小为2m的X 的子集。该证明的想法是对于任意学习算法，只学习C中一半的实例并不能推断出另一半实例的标签。因此，存在“现实”，即，存在一些目标函数f，会违背A(S)预测的另一半未学习的实例。

注意，有$T = 2^m$个可能的从C到{0,1}的函数。记这些函数为f1,f2,...,fT. 对于每个函数，令D_i是$C \times \{0,1\}$上的分布：

$$D_i(\{(x,y)\}) = \begin{cases} 1/|C| & \text{若 } y = f_i(x) \\ 0 & \text{其他} \end{cases} \tag{1}$$

即当y是真是的x的标签时，选出(x,y)的概率为1/|C|。若不是，则为0。显然$L_{D_i}(f_i) = 0$

我们将展示，对于任何算法，A接收来自$C \times \{0,1\}$的m个例子，返回函数$A(S) : C \to \{0,1\}$，满足

$$\max_{i \in [T]} E[L_{D_i}(A(S))] \geq 1/4 \tag{2}$$

可以清楚的看到，这意味着对于每个算法A'，接受来自$X \times \{0,1\}$ 的m个例子，存在函数$f : X \to \{0,1\}$与分布D，使得$L_D(f) = 0$并且

$$E[L_D(A'(S))] \geq 1/4 \tag{3}$$

很容易验证，这些足以说明$P[L_D(A'(S)) \geq 1/8] \geq 1/7$，即我们所要证明的。（见练习1）

我们现在需要证明方程2。C有$k = (2m)^m$种可能的序列。记为$S_1, \ldots, S_k$，且若$S_j = (x_1, \ldots, x_m)$，我们记S_j^i为标签函数为f_i的序列S_j，即$S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$。若分布为$D_i$，则可能的训练集A可以接受$S_1^i, \ldots S_k^i$ 且所有的训练集有相同的概率被抽样。因此，

$$E[L_{D_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \tag{4}$$

因为最大值>平均值>最小值，故

$$\begin{aligned} \max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) \\ &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) \\ &\geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) \end{aligned} \tag{5}$$

下一步，定下一些$j \in [k]$，记$S_j = (x_1, \ldots, x_m)$并且设$v_1, \ldots, v_p$为没有出现在S_j中但在C中的例子。显然，$p \geq m$。因此，对于每个函数$h : C \to \{0,1\}$以及每个i我们都有

$$\begin{aligned} L_{\mathcal{D}_i}(h) &= \frac{1}{2m} \sum_{x \in C} 1_{[h(x) \neq f_i(x)]} \\ &\geq \frac{1}{2m} \sum_{r=1}^{p} 1_{[h(v_r) \neq f_i(v_r)]} \\ &\geq \frac{1}{2p} \sum_{r=1}^{p} 1_{[h(v_r) \neq f_i(v_r)]} \end{aligned}$$

因此

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}\left(A\left(S_j^i\right)\right) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}1_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}1_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\\ &\geq \frac{1}{2}\cdot\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}1_{[A(S_j^i)(v_r)\neq f_i(v_r)]}\end{aligned} \tag{6}$$

接下来，对于一些$r\in[p]$我们可以把函数$f_1,\ldots,f_T$分为T/2个对，每一对$(f_i,f_{i'})$，当且仅当$c=v_r$，我们有$c\in C, f_i(c)\neq f_{i'}(c)$. 由于对于这样的每一对，必须有$S_j^i=S_j^{i'}$,满足

$$1_{[A(S_j^i)(v_r)\neq f_i(v_r)]}+1_{[A(S_j^{i'})(v_r)\neq f_{i'}(v_r)]}=1 \tag{7}$$

可得

$$\frac{1}{T}\sum_{i=1}^{T}1_{[A(S_j^i)(v_r)\neq f_i(v_r)]}=1/2 \tag{8}$$

将这个与方程6, 5,以及方程4结合，我们可以得到方程2 即证明完成。

5.1.1 没白午 与 先验知识

天下没有白吃的午餐定理（下面简称《没白午》定理）又是如何能与先验知识联系在一起的呢？让我们考虑基于函数f先验集H的ERM预测器。这类缺少先验知识：每个来自从域到标签集的可能函数都被考虑为候选者。根据《没吃午》定理，任何算法选择输出，都会失败。因此，这类并非PAC可学习的。可以导致下面的推论：

推论5.2 若X为无限域集，H为所有从X到{0,1}的函数集合。则H不是PAC可学习的。

证明 假设，（反证）若类是可学习的。
选择$\epsilon<1/8,\delta<1/7$. 根据PAC可学习性的定义，必然有一些学习算法A及整数$m=m(\epsilon,\delta)$ 使得对于所有数据生成分布，若对一些函数$f:X\to\{0,1\}, L_D(f)=0$，随着概率大于$1-\delta,\ \ L_D(A(S))\leq\epsilon$. 然而，根域《没吃午》定理。由于$|X|>2m$，每个学习算法（尤其是学习算法A）都存在一个分布D使得有$1/7>\delta, L_D(A(S))>1/8>\epsilon$，导致矛盾。□

我们怎么组织这般失败呢？我们可以危险预知通过《没白午》定理，使用我们对学习任务的先验知识去避免可能导致我们失败的分布。该先验知识可以限制我们的先验集。

但是我们应该选择好的先验集么？一方面，我们希望类包含的先验没有误差（PAC设定），或者要有最小的误差而不是相对较少（不可知PAC设定）。另一方面，我们不能简单的选择最大的集——一个包含所有函数的集。因此，权衡与博弈即将到来。

5.2 误差分解

为了回答这个问题，我们可以将ERM误差分解为两部分。令h_S为ERM先验。

$$L_D(h_S)=\epsilon_{\text{近似}}+\epsilon_{\text{估计}}\ \ \text{其中}\epsilon_{\text{近似}}=\min L_D(h), \epsilon_{est}=L_D(h_S)-\epsilon_{\text{近似}} \tag{9}$$

- 近似误差 - 假定集中的预测器能实现的最小风险。这一项描述了在限定假定集中的风险。该项并不依赖样本的多少，而是取决于假定集的大小。扩大假定集有助于减少近似误差。
- 估计误差 - 近似误差与ERM预测器误差的差值。由于经验风险是真实风险的估计值，因此最小化经验风险就相当于最小化真实风险的估计值。
  这一项的大小取决于训练集的样本数，以及假定集的大小、复杂度。如前文，对于有限的假定集，$\epsilon_{\text{估计}}$ 随着假定集大小$|H|$增大，而随着样本数m减少。我们可以认为，H的大小描述了假定集复杂度。在后面，我们会详细定义假定集的复杂度。

由于我们的目标是最小化总风险，我们不免面临着一个权衡问题。即偏差与复杂度的权衡。一方面，选择一个非常庞大的H有助于减少近似误差，但是会增加估计误差，因为庞大的H会导致过拟合。当然，最好的选择是只存在一个元素的H——贝叶斯最优预测器。但显然，这个依赖于分布D的贝叶斯最优预测器并不现实（我们要是知道D还学习个啥玩意）。

学习理论研究的是在保持一个合理的估计误差限制下，H可以有多大。大多数情况下，经验研究基于给出一个巧妙地假定集。这里，“巧妙”指的是一个近似误差不会特别大的集。想法是，我们虽然并不擅长，也不必知道如何构造最优的预测器，我们仍然可以利用手中已有的一点先验知识，构造出一个近似误差与估计误差都不太大的预测器。回到我们苦瓜的例子，我们并不知道苦瓜的色泽、硬度如何决定其味道。但是我们首先能确定——苦瓜是一种水果。因此，借助我们对水果的先验知识，我们可以判断，用色泽-硬度空间会是一个不错的预测器。

5.3 总结

没白午定理指出没有一个普适的学习器。每个学习器都应该基于特定的任务，并基于一些先验知识来确保成功。目前为止，我们建立的先验模型都是从一个大假定集中选取特定的假定子集。但选定这个假定集时，我们面临着权衡，是选择一个大的复杂的假定集来保证近似误差很小，还是选择一个有更多限制的小假定集来使估计误差更小？在下一章，我们将会更详细的探索估计误差。在第7章，我们将会用另一种方式讨论先验知识。