Author：Liu Jian

Time：2020-05-16

机器学习Ⅲ-计算学习理论

> 参考文献：
>
> 1. pluskid - 机器学习物语(4)：PAC Learnability
> 2. Foundations of Machine Learning - Mehryar Mohri - 2018
> 3. An Introduction to Computational Learning Theory - Michael J. Kearns - 1994

西瓜书上本章内容晦涩难懂，且存在很多错误，建议以参考文献 [2] 为主。

一般情况下，我们这里讨论的都是**二分类的确定性问题**，所谓确定性是指对任意输入 $\boldsymbol{x} \sim \mathcal{D}$，其输出 y 是确定的，我们这里假设由目标概念 $c_{obj}(\boldsymbol{x})$ 给出，即 $y = c_{obj}(\boldsymbol{x})$；更一般的情况是输入输出服从某一联合概率分布：$(\boldsymbol{x}, y) \sim \mathcal{D}$，此时，给定 $\boldsymbol{x}$ 后，y 的值并不确定，服从一定的概率分布。

1 概念说明

1. 泛化误差、经验误差

某个具体函数 $h(\cdot) : \mathbb{X} \to \mathbb{Y}$ 的**泛化误差**：

$$R(h) = \mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}}(h(\boldsymbol{x}) \neq y) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}\{\mathbb{I}(h(\boldsymbol{x}) \neq y)\}$$

其中，$\mathcal{D}$ 为自变量 $\boldsymbol{x}$ 服从的真实分布；y 为 $\boldsymbol{x}$ 的真实标记，由某个未知的映射 $c_{obj}(\cdot)$ 给出，该映射也称目标映射；$\mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}}(h(\boldsymbol{x}) \neq y)$ 中 $\mathbb{P}$ 并不是指 $\boldsymbol{x}$ 的分布，$\boldsymbol{x}$ 分布的自变量显然不能为事件 $h(\boldsymbol{x}) \neq y$，因此这里是指 $h(\boldsymbol{x}) \neq y$ 这一事件发生时对应的 $\boldsymbol{x}$ 的概率之和，但显然这一概率大小依赖于 $\boldsymbol{x}$ 的分布。

分布 $\mathcal{D}$ 的一个独立同分布采样为 S，当然我们也会得到 S 中自变量的标记，S 的容量为 N，则 h 在数据集 S 上的**经验误差**：

$$\hat{R}_S(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(h(\boldsymbol{x}_i) \neq y_i)$$

我们需要说明的是，**下文中，有时候 S 只表示对输入 $\boldsymbol{x}$ 的采样 (西瓜书上称之为示例)，对于确定性问题，已知输入，对应标记也就唯一确定了；而有时候 S 还包括标记，即 $(\boldsymbol{x}, y)$，需根据具体语境确定。**

二者之间的关系：h 经验误差的期望等于泛化误差 $\mathbb{E}_{S \sim \mathcal{D}^N}\{\hat{R}_S(h)\} = R(h)$。因此，我们会用**经验误差去估计泛化误差，构建泛化误差界。**

2. 概念类、假设空间

假设空间 $\mathcal{H}$：我们使用某个模型对样本数据进行学习，以确定模型中的待定参数，但实际上这个模型含有未知参数，并不是一个具体的模型，而是代表了一类模型，机器学习或者说算法 $\mathcal{L}$ 做的就是从这一类模型中找到最优的那一个，我们把这一类模型张成的空间称为假设空间 $\mathcal{H}$。当 $\mathcal{H}$ 中模型的个数有限时，$\mathcal{H}$ 为**有限假设空间**；当 $\mathcal{H}$ 中模型的个数无限时，$\mathcal{H}$ 为**无限假设空间。**

概念类 $\mathcal{C}$：前面我们提到目标概念 (target concept) $c_{obj}(\cdot):\mathbb{X}\to\mathbb{Y}$ 是一个函数，给出了输入空间 $\mathbb{X}$ 中所有点的真实标记，即 $c(\boldsymbol{x})=y$，y 为 $\boldsymbol{x}$ 的真实标记。但是，和假设空间类似，我们只知道目标概念 c_{obj} 属于某一个模型空间，但具体是其中的哪一个模型我们并不知道，这一模型空间被称为概念类 $\mathcal{C}$。

当 $c_{obj}\in\mathcal{H}$ 时，也就是说 $\mathcal{H}$ 存在一个模型能将所有的自变量 $\boldsymbol{x}$ 正确地映射到其真实标记上，这种情况称为**可分的 (separable) 或 一致的 (consistent)**；当 $c_{obj}\notin\mathcal{H}$ 时，为**不可分的 (non-separable) 或不一致的 (non-consistent)。**

3. Hoeffding 不等式：若 $x^{(1)},x^{(2)},\cdots,x^{(n)}$ 为 n 个独立随机变量，且满足 $0\leqslant x^{(i)}\leqslant 1$，则对任意 $\epsilon>0$，有

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{x^{(i)}\}\geqslant\epsilon\right)\leqslant\exp\left(-2n\epsilon^2\right)$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}x^{(i)}-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{x^{(i)}\}\right|\geqslant\epsilon\right)\leqslant 2\exp\left(-2n\epsilon^2\right)$$

4. **机器学习的过程就是**：算法 $\mathcal{L}$ 根据数据集 S 从 $\mathcal{H}$ 中返回一个模型，记为 h_S。可以看到，h_S 除了依赖于数据集 S 还依赖于算法 $\mathcal{L}$，但 PAC 讨论的是问题的可学习性，并不依赖于具体的算法，因此为了简便起见，我们在 h_S 中忽略算法标记 $\mathcal{L}$，即不使用类似于 $h_{S,\mathcal{L}}$ 的标记。

可以看到，对于不同的采样 S，算法输出的假设 h_S 并不固定，因此，h_S 而是一个随机变量。前面我们说，对于一个给定的假设 (a fixed hypothesis) h，我们有 $\mathbb{E}_{S\sim\mathcal{D}^N}\{\hat{R}_S(h)\}=R(h)$，但对于随机假设 h_S，$R(h_S)$ 也是一个随机变量，而 $\mathbb{E}_{S\sim\mathcal{D}^N}\{\hat{R}_S(h_S)\}$ 是一个常数，二者显然不会相等，即 $\mathbb{E}_{S\sim\mathcal{D}^N}\{\hat{R}_S(h_S)\}\neq R(h_S)$，故对于 h_S 不能直接使用 Hoeffding 不等式。考虑到假设空间 $\mathcal{H}$ 中的所有假设都有可能成为 h_S，我们可以看到，在下文中，为了得到 h_S 的泛化误差界，我们在一定概率保证下，基于 Hoeffding 不等式推得了 $\mathcal{H}$ 中所有假设的泛化误差界。

2 PAC 可学习的定义

PAC 可学习的定义：

A concept class $\mathcal{C}$ is said to be **PAC-learnable** if there exists an algorithm $\mathcal{L}$ and a polynomial function $\mathrm{poly}(\cdot;\cdot;\cdot;\cdot)$ such that for any $\epsilon>0$ and $\delta>0$, for all distributions $\mathcal{D}$ on $\mathbb{X}$ and for any target concept $c\in\mathcal{C}$, the following holds for any sample size $N\geqslant\mathrm{poly}(1/\epsilon;1/\delta;\mathrm{size}(\boldsymbol{x});\mathrm{size}(c))$:

$$\mathbb{P}_{S\sim\mathcal{D}^N}(R(h_S)\leqslant\epsilon)\geqslant 1-\delta$$

If $\mathcal{L}$ further runs in $\mathrm{poly}(1/\epsilon;1/\delta;\mathrm{size}(\boldsymbol{x});\mathrm{size}(c))$, then $\mathcal{C}$ is said to be **efficiently PAC-learnable**. When such an algorithm $\mathcal{L}$ exists, it is called **a PAC-learning algorithm for** $\mathcal{C}$.

如果存在一个算法 $\mathcal{L}$ 和多项式函数 $\mathrm{poly}(\cdot;\cdot;\cdot;\cdot)$，使得对输入空间 $\mathbb{X}$ 上的任意分布 $\mathcal{D}$ 和目标概念 c_{obj} 为概念类 $\mathcal{C}$ 中的任何一个概念 c 时，当样本 $N\geqslant\mathrm{poly}(1/\epsilon;1/\delta;\mathrm{size}(\boldsymbol{x});\mathrm{size}(c))$，算法 $\mathcal{L}$ 在样本 S 上所学得模型的泛化误差小于任意给定 $\epsilon>0$ 的概率尽可能大，不低于 $1-\delta$，其中 δ 为任意给定的置信度：

$$\mathbb{P}_{S\sim\mathcal{D}^N}(R(h_S)\leqslant\epsilon)\geqslant 1-\delta$$

那么我们就说**概念类 $\mathcal{C}$ 是 PAC 可学习的**；进一步地，若这个算法的运行时间也是多项式时间 $\mathrm{poly}(1/\epsilon; 1/\delta; \mathrm{size}(\boldsymbol{x}); \mathrm{size}(c))$，那么概念类 $\mathcal{C}$ 为**高效 PAC 可学习的**，这样的**算法 $\mathcal{L}$ 被称为概念类 $\mathcal{C}$ 的 PAC 学习算法。**

- PAC 研究的是概念类 $\mathcal{C}$ 的可学习性，也就是说不论目标概念 c_{obj} 为 $\mathcal{C}$ 中的哪一个概念，是否存在一个算法均能"很好地"对其进行学习。注意，PAC 关心的是算法的存在性，并不关心这个算法具体是什么；
- "很好地学习"意味着当样本数量也就是样本复杂度 (sample complexity) $N \geqslant \mathrm{poly}(1/\epsilon; 1/\delta; \mathrm{size}(\boldsymbol{x}); \mathrm{size}(c))$ 时，会以很大的概率 $\geqslant 1-\delta$ 学得泛化误差足够小 $\leqslant \epsilon$ 的模型；
- $\mathrm{size}(\boldsymbol{x})$ 表示输入 $\boldsymbol{x}$ 的复杂度，比如 $\boldsymbol{x}$ 的维度 n；而 $\mathrm{size}(c)$ 表示概念类 $\mathcal{C}$ 中概念的最大复杂度；当考虑运行时间时，$\mathrm{size}(\boldsymbol{x}), \mathrm{size}(c)$ 就为对应的处理时间，比如 $\mathrm{size}(c)$ 就表示概念类 $\mathcal{C}$ 中的概念 c 输出一个样本标记所花的最大的时间。参考文献 [2] 中给出了两个学习问题，其目标概念是等价的，但表示形式不同，即对于相同输入，输出的标记相同，但计算输出所用时间不同，由此，虽然它们都是 PAC 可学习的，但一个是高效可学习的，一个则不是高效可学习的。可以看到，PAC 可学习性考察的是样本复杂度，而高效 PAC 可学习性还需进一步考察运行时间，这二者是不同的；
- 算法分析中，我们一般考察的是事件复杂度 (运行时间) 和空间复杂度 (占用内存)，而 PAC 更关注样本的复杂度 N。

我们可以根据假设空间 $\mathcal{H}$ 是否有限，目标概念 c_{obj} 是否属于假设空间 $\mathcal{H}$ 即是否可分，分情况讨论 PAC 可学习性：

- 不可分时 $c_{obj} \notin \mathcal{H}$，前面关于 PAC 学习的定义不再适用，比如，我们无法在有限的假设假设空间 $\mathcal{H}$ 中找到泛化误差无限小或为 0 的假设。此外，对于非确定性问题即 $(\boldsymbol{x}, y) \sim \mathcal{D}$，泛化误差也不可能达到 0。但我们可以放宽条件，试图去寻找假设空间中泛化误差最小的假设而不是泛化误差为 0 的假设。由此，我们可以对 PAC 学习的定义进行推广，引入了**不可知 PAC 学习 (agnostic PAC learning)，其定义如下**：

Let $\mathcal{H}$ be a hypothesis set. $\mathcal{L}$ is an agnostic PAC-learning algorithm if there exists a polynomial function $\mathrm{poly}(\cdot;\cdot;\cdot;\cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ over $\mathbb{X} \times \mathbb{Y}$, the following holds for any sample size $N \geqslant \mathrm{poly}(1/\epsilon; 1/\delta; \mathrm{size}(\boldsymbol{x}); \mathrm{size}(c))$:

$$\mathbb{P}_{S \sim \mathcal{D}^N}\left(R(h_S) - \min_{h \in \mathcal{H}} R(h) \leqslant \epsilon\right) \geqslant 1 - \delta$$

If $\mathcal{L}$ further runs in $\mathrm{poly}(1/\epsilon; 1/\delta; \mathrm{size}(\boldsymbol{x}); \mathrm{size}(c))$, then it is said to be an efficient agnostic PAC-learning algorithm.

可以看到 **PAC 学习是上述不可知 PAC 学习的特例，此时 $\min_{h \in \mathcal{H}} R(h) = 0$。因此，对于某个学习问题，我们可以直接讨论其不可知 PAC 可学习性，若是不可知 PAC 可学习的，进一步还有 $c_{obj} \in \mathcal{H}$，则为 PAC 可学习的。不过我们需要说明的是，PAC 可学习针对的是概念类 $\mathcal{C}$，即我们会说概念类 $\mathcal{C}$ 是 PAC 可学习的；而不可知 PAC 可学习针对的是假设空间 $\mathcal{H}$，即我们会说假设空间 $\mathcal{H}$ 是不可知 PAC 可学习的。虽然有此细微差别，但无妨大局，因此我们常不加区分地使用。**

- **假设空间无限时，我们需要引入 VC 维或者 Rademacher 复杂度来衡量假设空间的"大小"，更准确地说是假设空间表示能力的大小。**

接下来，我们将给出假设空间有限时，可分与不可分这两种情况下的 (不可知) PAC 可学习性的分析，以及假设空间无限时的处理方法。

3 假设空间有限可分

我们指出，有限可分一定是 PAC 可学习的，接下来我们来证明这一结论。

有限可分：假设空间 $\mathcal{H}$ 有限，且 $c_{obj} \in \mathcal{H}$ 。对于这种情况，在训练集 S 上出现标记错误的假设 h 一定不是目标概念 c_{obj} 。因此，我们的可以选择这样一个很简单的**算法，就是排除这些出现错误预测的假设，而只留下那些在训练集 S 上不出错的假设，并任意返回一个作为学得的模型。**为了判断这种情况下的 PAC 可学习性，接下来，问题即为推导需要多少样本 N 才能保证通过这种算法所得假设的泛化误差大于 ϵ 的概率小于 δ 。

记 $\mathcal{H}$ 中泛化误差大于 ϵ 的假设组成的集合为 $\mathcal{H}_\epsilon = \{h \in \mathcal{H} | R(h) > \epsilon\}$ 。我们并不知道算法会返回经验误差为 0 的假设中的哪一个，但是反过来，我们只需使 $\mathcal{H}_\epsilon$ 中的假设在样本上表现完美 (即经验误差为 0) 的情况出现的概率尽可能小，小于 δ 即可：

$$\mathbb{P}(\exists h \in \mathcal{H}_\epsilon : \hat{R}_S(h) = 0) \leqslant \delta$$

而

$$\begin{aligned}
&\mathbb{P}(\exists h \in \mathcal{H}_\epsilon : \hat{R}_S(h) = 0) = \mathbb{P}\left((\hat{R}_S(h_1) = 0) \vee (\hat{R}_S(h_2) = 0) \vee \cdots \vee (\hat{R}_S(h_{|\mathcal{H}_\epsilon|}) = 0)\right) \\
&\leqslant \sum_{h \in \mathcal{H}_\epsilon} \mathbb{P}(\hat{R}_S(h) = 0) \leqslant \sum_{h \in \mathcal{H}_\epsilon} (1-\epsilon)^N = |\mathcal{H}_\epsilon|(1-\epsilon)^N \leqslant |\mathcal{H}|(1-\epsilon)^N \leqslant |\mathcal{H}| \exp(-N\epsilon)
\end{aligned}$$

令 $\delta \geqslant |\mathcal{H}| \exp(-N\epsilon)$ ，解得：

$$N \geqslant \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

可以看到 $\frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$ 可视为 ϵ, δ 多项式函数，则**有限可分时，一定是 PAC 可学习的。**

最后，我们对算法输出假设 h_S 的泛化误差界进行解读， h_S 的泛化误差

$$R(h_S) \leqslant \frac{1}{N}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

的概率至少为 $1-\delta$ 。可以看到，在一定概率保证下，随着 N 的增大，算法输出假设的泛化误差上界以 $O(1/N)$ 的速率减小，这是一个很可观的收敛速率。此外，为了保证可分性，即 $c_{obj} \in \mathcal{H}$ ，我们一般会取一个较大的假设空间 $\mathcal{H}$ ，这会使得算法输出假设的泛化误差上界增大，但增大速率是对数的 $\ln|\mathcal{H}|$ ，并不太大。事实上， $\ln|\mathcal{H}|$ 可视为描述假设空间所需的奈特数。

4 假设空间有限不可分

我们指出，有限不可分一定是不可知 PAC 可学习的，接下来我们来证明这一结论。

有限不可分：假设空间 $\mathcal{H}$ 有限，但 $c_{obj} \notin \mathcal{H}$ 。**我们可以使用遵循经验风险最小化 (empirical risk minimization, EMR) 的算法，算法 $\mathcal{L}$ 根据样本 S ，返回使经验损失最小的假设：$h_S = \min_{h \in \mathcal{H}} \hat{R}_S(h)$ 。此外，我们无法直接计算不可知 PAC 可学习性定义中涉及到的泛化误差，为此，我们可以先使用经验误差构建泛化误差的界，再结合经验风险最小化进行推导。**

4.1 $\mathcal{H}$ 中所有假设的泛化误差界/ h_S 的泛化误差界

1. 由 Hoeffding 不等式可知，对任意给定的映射 h (并不是指算法输出的映射 h_S ，事实上映射 h 属不属于 $\mathcal{H}$ 下面的结论都成立，但这里既然我们讨论的是假设空间中的模型，不妨取 $h \in \mathcal{H}$)，其泛化误差 $R(h)$ 和经验误差 $\hat{R}_S(h)$ 满足如下的关系：

$$\mathbb{P}_{S \sim \mathcal{D}^N}\left(\left|R(h) - \hat{R}_S(h)\right| \leqslant \epsilon\right) \geqslant 1 - 2\exp\left(-2N\epsilon^2\right)$$

令 $\delta = 2\exp\left(-2N\epsilon^2\right)$ ，可知，对任意的 $\delta > 0$ ，如下关系以不低于 $1-\delta$ 的概率成立：

$$\left|R(h) - \hat{R}_S(h)\right| \leqslant \sqrt{\frac{\ln(2/\delta)}{2N}}$$

用数学语言表示如下：

$$(a)\ \forall h \in \mathcal{H},\ \mathbb{P}_{S\sim\mathcal{D}^N}\left(\left|R(h)-\hat{R}_S(h)\right| \leqslant \sqrt{\frac{\ln(2/\delta)}{2N}}\right) \geqslant 1-\delta$$

可以看到，**结论 (a) 在某一概率保证下，基于经验误差给出了 $\mathcal{H}$ 中某个假设的泛化误差界。** 顺便强调一下，N, ϵ, δ 这三个变量的自由度为二，它们间存在关系式 $\delta = (\geqslant) 2\exp\left(-2N\epsilon^2\right)$ 。结论 (a) 中概率不等式中只出现了 N, δ ，我们也可以给出只出现 N, ϵ ，或只出现 ϵ, δ 的概率不等式。

2. 在第一步中，我们得到了某个任意给定的假设 (a fixed hypothesis) $h \in \mathcal{H}$ 的泛化误差界。正如前文所述，采样所得样本 S 不同，算法返回的假设 h_S 也可能不同， h_S 是一个随机变量， $\mathcal{H}$ 中的每一个假设都可能被返回。**因此，第一步中所得的结论 (a) 是不够的，我们需要基于 Hoeffding 不等式为 $\mathcal{H}$ 中所有假设推导出一个统一的泛化误差界 (a uniform convergence bound)，我们首先给出结论**：

Let $\mathcal{H}$ be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds:

$$\forall h \in \mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right| \leqslant \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2N}}$$

用数学语言表述如下：

$$(b)\ \mathbb{P}_{S\sim\mathcal{D}^N}\left(\forall h \in \mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right| \leqslant \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2N}}\right) \geqslant 1-\delta$$

注意这里的结论 (b) 和第一步中结论 (a) 区别，这里的 $\forall h \in \mathcal{H}$ 在概率括号内，而上一步的 $\forall h \in \mathcal{H}$ 在括号外。可以看到，结论 (a) 中的 h 虽然可以任意给定，但在讨论概率时， h 是固定的 (fixed)；但 (b) 中讨论概率时， h 并不是固定的，针对的是 $\mathcal{H}$ 中所有的假设。结论 (a) 和 (b) 分别等价于：

$$\forall h \in \mathcal{H},\ \mathbb{P}_{S\sim\mathcal{D}^N}\left(\left|R(h)-\hat{R}_S(h)\right| \geqslant \sqrt{\frac{\ln(2/\delta)}{2N}}\right) \leqslant \delta$$

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(\exists h \in \mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right| \geqslant \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2N}}\right) \leqslant \delta$$

二者的区别类似于掷硬币，一共掷了 N 次 $X = (x_1, \cdots, x_N)$ ，正面记为 1 反面记为 -1 ，那么对于其中任意一次投掷来说，出现正面的概率为 $1/2$ ，但所有投掷结果均为正面的概率则为 $(1/2)^N$ ：

$$\forall x_i \in X,\ \mathbb{P}(x_i = 1) = \frac{1}{2}$$

$$\mathbb{P}(\forall x_i \in X,\ x_i = 1) = \left(\frac{1}{2}\right)^N$$

接下来，我们来证明结论 (b)。我们的目标是得到如下的关系式：

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(\forall h \in \mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right| \leqslant \epsilon\right) \geqslant 1-\delta$$

这等价于：

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(\exists h \in \mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right| \geqslant \epsilon\right) \leqslant \delta$$

而

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(\exists h\in\mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right)$$
$$=\mathbb{P}_{S\sim\mathcal{D}^N}\left(\left(\left|R(h_1)-\hat{R}_S(h_1)\right|\geqslant\epsilon\right)\vee\left(\left|R(h_2)-\hat{R}_S(h_2)\right|\geqslant\epsilon\right)\vee\cdots\vee\left(\left|R(h_{|\mathcal{H}|})-\hat{R}_S(h_{|\mathcal{H}|})\right|\geqslant\epsilon\right)\right)$$
$$\leqslant\sum_{i=1}^{|\mathcal{H}|}\mathbb{P}_{S\sim\mathcal{D}^N}\left(\left|R(h_i)-\hat{R}_S(h_i)\right|\geqslant\epsilon\right)\leqslant 2|\mathcal{H}|\exp\left(-2N\epsilon^2\right)\ \text{(Hoeffding 不等式)}$$

令 $\delta=2|\mathcal{H}|\exp\left(-2N\epsilon^2\right)$ 即可得结论 (b)。可以看到，**结论 (b) 在一定的概率保证下，基于经验误差给出了 $\mathcal{H}$ 中所有假设的泛化误差界。这样，虽然算法输出的假设 h_S 是一个随机变量，并不固定，可能为 $\mathcal{H}$ 中任意一个假设，但我们通过对 $\mathcal{H}$ 中所有假设求泛化误差界，还是得到了 h_S 的泛化误差界，即由结论 (b) 可得：**

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(\left|R(h_S)-\hat{R}_S(h_S)\right|\leqslant\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2N}}\right)\geqslant 1-\delta$$

和有限可分时一样，我们来对 h_S 的泛化误差界进行分析。在一定的概率保证下，我们有：

$$R(h_S)\leqslant\hat{R}_S(h_S)+O\left(\sqrt{\frac{\ln|\mathcal{H}|}{2N}}\right)$$

此时，泛化误差界随样本容量的增大而减小的速率为 $O(\sqrt{1/N})$。因此，**为了达到和有限可分时相同的泛化误差界，样本容量大大增加，需为有限可分时的平方。**事实上，对于 1.3 节有限可分的情况，我们也可以先证明其不可知 PAC 可学习性，再由 $c_{obj}\in\mathcal{H}$ 可得其 PAC 可学习，但显然 1.3 节中所得的泛化误差界更好。此外，对于泛化误差界 $\hat{R}_S(h_S)+O\left(\sqrt{\frac{\ln|\mathcal{H}|}{2N}}\right)$，**在第一项经验误差 $\hat{R}_S(h_S)$ 相等的前提下，较小 (也就是较简单) 的假设空间意味着更小的泛化误差界，这体现了 Occam's Razor principle**，即 All other things being equal, a simpler (smaller) hypothesis set is better.

4.2 泛化误差界 + EMR 推导可学习性

在得到 $\mathcal{H}$ 中所有假设的泛化误差界后，我们就可以结合经验风险最小化进行证明了。

记假设空间中泛化误差最小的假设为 $g=\arg\min_{h\in\mathcal{H}}R(h)$，经验风险最小化算法给出的假设 $h_S=\arg\min_{h\in\mathcal{H}}\hat{R}_S(h)$。我们将结论 (b) 表述为便于我们使用的形式：

$$\text{if}\ \ N\geqslant\frac{1}{2\epsilon^2}(\ln 2|\mathcal{H}|-\ln\delta),$$
$$\text{then}$$
$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(\forall h\in\mathcal{H},\quad\left|R(h)-\hat{R}_S(h)\right|\leqslant\epsilon\right)\geqslant 1-\delta$$

则，当 N 满足上述条件时，对于 $g\in\mathcal{H},h_S\in\mathcal{H}$，下面二式同时成立的概率不小于 $1-\delta$：

$$\left|R(g)-\hat{R}_S(g)\right|\leqslant\epsilon$$
$$\left|R(h_S)-\hat{R}_S(h_S)\right|\leqslant\epsilon$$

而上述二式成立时，我们有：

$$R(h_S)-R(g)\leqslant\hat{R}_S(h_S)+\epsilon-\left(\hat{R}_S(g)-\epsilon\right)$$
$$=\hat{R}_S(h_S)-\hat{R}_S(g)+2\epsilon\leqslant 2\epsilon$$

则，当 $N\geqslant\frac{1}{2\epsilon^2}(\ln 2|\mathcal{H}|-\ln\delta)$ 时，

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(R(h_S)-R(g)\leqslant 2\epsilon\right)\geqslant 1-\delta$$

取 $\epsilon \leftarrow 2\epsilon$，则当 $N \geqslant \dfrac{2}{\epsilon^2}(\ln 2|\mathcal{H}| - \ln \delta)$ 时，

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left(R(h_S) - R(g) \leqslant \epsilon\right) \geqslant 1-\delta$$

满足不可知 PAC 可学习的定义，证毕。

5 假设空间无限

当假设空间无限时，$|\mathcal{H}|$ 为无穷大，$|\mathcal{H}|$ 不能很准确地描述假设空间的复杂性或者说表示能力，假设空间有限时基于 $|\mathcal{H}|$ 关于 (不可知) PAC 可学习性的讨论已不再适用。为此，我们可以使用 VC 维、Rademacher 复杂度等指标以描述假设空间无限时的表示能力。**假设空间无限时 (不可知) PAC 可学习性有关结论的证明和假设空间有限时类似：先基于经验误差得到假设空间 $\mathcal{H}$ 中所有假设的泛化误差界，再结合经验风险最小化进行证明。**

5.1 增长函数与 VC 维

我们首先给出结论：**任何 VC 维有限的假设空间 $\mathcal{H}$ 都是 (不可知) PAC 可学习的。**

在引出 VC 维前，我们先介绍**增长函数 (growth function)。**The growth function $\Pi_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by:

$$\forall N \in \mathbb{N},\ \Pi_{\mathcal{H}}(N) = \max_{\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}} |\{(h(\boldsymbol{x}_1),\cdots,h(\boldsymbol{x}_N)) : h \in \mathcal{H}\}|$$

可见，$\Pi_{\mathcal{H}}(N)$ 表示假设空间对 N 个输入所能赋予标记的最大种数，而标记的每种可能结果称为一个**对分 (dichotomy)。**可以看到，This provides a measure of the richness of the hypothesis set $\mathcal{H}$. And this measure does not depend on the distribution, it is purely combinatorial.

基于增长函数，我们有如下结论：Let $\mathcal{H}$ be a family of functions taking values in $\{-1,1\}$. Then, for any $0<\epsilon<1$:

$$(c)\ \ \mathbb{P}_{S\sim\mathcal{D}^N}\left(\exists h\in\mathcal{H},\ \ \left|R(h)-\hat{R}_S(h)\right|>\epsilon\right) \leqslant 4\Pi_{\mathcal{H}}(2N)\exp\left(-\frac{N\epsilon^2}{8}\right)$$

上述结论 (c) 已很接近上一小节的结论 (b)，但我们不会直接基于结论 (c) 对 (不可知) PAC 可学习性进行推导，因为增长函数 $\Pi_{\mathcal{H}}(2N)$ 是假设空间表达能力在样本容量为 $2N$ 时的体现，其大小除了与假设空间 $\mathcal{H}$ 的表达能力有关，还与样本容量 N 有关，从其定义可以看出其计算一般比较困难，没有解析的表达式，直接令 $\delta = 4\Pi_{\mathcal{H}}(2N)\exp\left(-\dfrac{N\epsilon^2}{8}\right)$ 会因为 $\Pi_{\mathcal{H}}(2N)$ 的存在，而无法解得 N，也就无从判断 N 是否服从多项式函数。**为此，我们引入 VC 维这样一个标量 (不像增长函数一样是 N 的函数) 来直接表征假设空间的表达能力或者说复杂性。**

对**二分类**问题的 VC 维的定义：The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered (打散) by $\mathcal{H}$:

$$\mathrm{VC}(\mathcal{H}) = \max\{N : \Pi_{\mathcal{H}}(N) = 2^N\}$$

对某一 N 个输入的打散 (shattering)，是指对这 N 个输入，假设空间能输出所有的可能标记，这里是二分类问题，因此可能标记的个数为 2^N，而 VC 维就是假设空间能打散输入的最大个数。注意，若某个假设空间 $\mathcal{H}$ 的 VC 维为 d，这并不表示 $\mathcal{H}$ 能打散所有等于或小于 d 的输入，VC 维讨论的是存在性，而不是任意性，是说存在某种 d 个输入，假设空间 $\mathcal{H}$ 能将这种输入打散。可以看到，若假设空间 $\mathcal{H}$ 有限，由抽屉原理其 VC 维一定小于 $|\mathcal{H}|$。

Sauer's lemma 给出了 VC 维与增长函数之间的关系：若假设空间 $\mathcal{H}$ 的 VC维为 d，对任意的 $N \in \mathbb{N}$，有：

$$\Pi_{\mathcal{H}}(N) \leqslant \sum_{i=0}^{d} \binom{N}{i}$$

若 $N \geqslant d$，可推得：

$$\Pi_{\mathcal{H}}(N) \leqslant \left(\frac{eN}{d}\right)^{d}$$

代入结论 (c) 中，可得：

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^N}\left(\exists h\in\mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right|>\epsilon\right) &\leqslant 4\Pi_{\mathcal{H}}(2N)\exp\left(-\frac{N\epsilon^2}{8}\right)\\ &\leqslant 4\left(\frac{2eN}{d}\right)^{d}\exp\left(-\frac{N\epsilon^2}{8}\right)\end{aligned}$$

令 $\delta = 4\left(\frac{2eN}{d}\right)^{d}\exp\left(-\frac{N\epsilon^2}{8}\right)$ 解得 $\epsilon = \sqrt{\frac{8d\ln(2eN/d)+8\ln(4/\delta)}{N}}$，则：

$$(d)\ \ \mathbb{P}_{S\sim\mathcal{D}^N}\left(\forall h\in\mathcal{H},\quad \left|R(h)-\hat{R}_S(h)\right|\leqslant\epsilon\right)\geqslant 1-\delta$$

结论 (d) 就对应了上一小节的结论 (b)，它在某一概率保证下，基于 VC 维给出了无限假设空间 $\mathcal{H}$ 所有假设的泛化误差界。可以看到，这一泛化误差界 ϵ 只与样本容量 N 有关，收敛速率为 $O(\frac{1}{\sqrt{N}})$，而与分布 $\mathcal{D}$ 和具体的采样数据 S 无关 (分布无关 distribution-free，数据独立 data-independent；当然，经验误差 $\hat{R}_S(h)$ 的计算总是依赖于采样数据 S 的，不过我们这里关注的是 ϵ)。

接下来，仿照上一小节，结合 EMR 即可证明本小节开头给出的结论。

5.2 Rademacher 复杂度

基于 VC 维的泛化误差界 ϵ 是分布无关、数据独立的，但若考虑数据分布，比如使用本小节将要介绍的 Rademacher 复杂度，那么我们就会得到更“紧”的泛化误差界。类似于极大似然估计使用具体的样本数据去估计统计量，Rademacher 复杂度使用具体的样本数据得到了一个更“紧”的 ϵ 。西瓜书上本节内容写得很混乱，相比之下参考文献 [2] 《Foundations of Machine Learning》就写得通俗易懂。

接下来我们将介绍有关 Rademacher 复杂度的一些概念和结论，并给出基于 Rademacher 复杂度的泛化误差界。和前面类似，有了泛化误差界，就可以对 (不可知) PAC 可学习性进行证明，这一部分不再赘述。

给定某个损失函数 $L:\mathbb{Y}\times\mathbb{Y}\to\mathbb{R}$，那么在损失函数 L 和假设空间 $\mathcal{H}$ 的基础上我们可以定义映射集合 $\mathcal{F}_{\mathcal{H}}$，其元素是这样的一些映射 $f_h = L(h(\boldsymbol{x}),y):\mathbb{X}\times\mathbb{Y}\to\mathbb{R}$。我们强调，$L$ 为某个给定的损失函数，是不变的，而 $h\in\mathcal{H}$ 是假设空间中任意的假设，是变化的，假设空间中的每个假设 $h\in\mathcal{H}$ 就对应了一个映射 f_h：

$$\mathcal{F}_{\mathcal{H}} = \{f_h:(\boldsymbol{x},y)\to L(h(\boldsymbol{x}),y):h\in\mathcal{H}\}$$

我们记 $\boldsymbol{z}=(\boldsymbol{x},y)$，则 f_h 就是从 $\mathbb{Z}$ 到 $\mathbb{R}$ 的映射。

(1) 我们首先以函数集合 $\mathcal{F}_{\mathcal{H}}$ 为例介绍 Rademacher 复杂度的概念：

- **函数集合 $\mathcal{F}_{\mathcal{H}}$ 的经验 Rademacher 复杂度**：对某个容量为 N 的固定样本 $S=(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N)$，函数集合 $\mathcal{F}_{\mathcal{H}}$ 关于样本 S 的经验 Rademacher 复杂度为：

$$\hat{\mathscr{R}}_S(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f_h\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f_h(\boldsymbol{z}_i)\right\}$$
$$= \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f_h\in\mathcal{F}}\frac{\boldsymbol{\sigma}^T\boldsymbol{f}_{h,S}}{N}\right\}$$

其中，$\boldsymbol{\sigma} = \langle\sigma_1,\cdots,\sigma_N\rangle^T$，$\sigma_i$ 为相互独立的算计变量，为 $\{-1,+1\}$ (注意不是区间，而是二值) 上的均匀分布，被称为 Rademacher 变量，表示随机噪声；$\boldsymbol{f}_{h,S} = \langle f_h(\boldsymbol{z}_1),\cdots,f_h(\boldsymbol{z}_N)\rangle^T$；**二者的内积 $\boldsymbol{\sigma}^T\boldsymbol{f}_{h,S}$ 衡量了 $\boldsymbol{f}_{h,S}$ 与随机噪声向量 $\boldsymbol{\sigma}$ 的相关程度。**Thus, the empirical Rademacher complexity measures on average how well the function class $\mathcal{F}_{\mathcal{H}}$ correlates with random noise on S. This describes the richness of the family $\mathcal{F}_{\mathcal{H}}$: richer or more complex families $\mathcal{F}_{\mathcal{H}}$ can generate more vectors $\boldsymbol{f}_{h,S}$ and thus better correlate with random noise, on average.

- 上述经验 Rademacher 复杂度针对的是某个固定的样本 S，若对样本 S 求期望，即可得**函数集合 $\mathcal{F}_{\mathcal{H}}$ 的 Rademacher 复杂度**：

$$\mathscr{R}_N(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_{S\sim\mathcal{D}^N}\left\{\hat{\mathscr{R}}_S(\mathcal{F}_{\mathcal{H}})\right\}$$
$$= \mathbb{E}_{S\sim\mathcal{D}^N}\left\{\mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f_h\in\mathcal{F}}\frac{\boldsymbol{\sigma}^T\boldsymbol{f}_{h,S}}{N}\right\}\right\} = \mathbb{E}_{S,\boldsymbol{\sigma}}\left\{\sup_{f_h\in\mathcal{F}}\frac{\boldsymbol{\sigma}^T\boldsymbol{f}_{h,S}}{N}\right\}$$

- 我们强调，**上述 Rademacher 复杂度的定义并不只是针对 $\mathcal{F}_{\mathcal{H}}$ 这种形式的函数集合，我们这里只是以 $\mathcal{F}_{\mathcal{H}}$ 为例来介绍 Rademacher 复杂度，我们可以求任意一个函数集合的 Rademacher 复杂度，只需将 f_h 换成对应的映射即可。此外，某个函数集合的 Rademacher 复杂度为 $\mathscr{R}_N(\cdot)$，而其经验 Rademacher 复杂度为 $\hat{\mathscr{R}}_S(\cdot)$，其中 $\cdot$ 表示该函数集合，因为 Rademacher 复杂度与具体样本 S 无关，而与样本容量 N 有关，因此其下标由经验 Rademacher 复杂度中的 S 变为 N。**

(2) 介绍完上述基本概念，我们有如下结论：

- **我们仍以函数集合 $\mathcal{F}_{\mathcal{H}}$ 为例，给出关于 Rademacher 复杂度的一般结论，即下述结论不仅限于 $\mathcal{F}_{\mathcal{H}}$ 这种形式的函数集合，将函数集合 $\mathcal{F}_{\mathcal{H}}$ 换成任意其他的函数集合也是成立的，我们只是以 $\mathcal{F}_{\mathcal{H}}$ 为例给出结论：若 $\mathcal{F}_{\mathcal{H}}$ 中的函数将 $\mathbb{Z}$ 映射到区间 $[0,1]$ 上，那么对任意的 $\delta>0$，我们有**：

$$\mathbb{P}\left(\forall f_h\in\mathcal{F}_{\mathcal{H}},\ \mathbb{E}_{\boldsymbol{z}}\{f_h(\boldsymbol{z})\} \leqslant \frac{1}{N}\sum_{i=1}^{N}f_h(\boldsymbol{z}_i) + 2\mathscr{R}_N(\mathcal{F}_{\mathcal{H}}) + \sqrt{\frac{\ln(1/\delta)}{2N}}\right) \leqslant 1-\delta$$
$$\mathbb{P}\left(\forall f_h\in\mathcal{F}_{\mathcal{H}},\ \mathbb{E}_{\boldsymbol{z}}\{f_h(\boldsymbol{z})\} \leqslant \frac{1}{N}\sum_{i=1}^{N}f_h(\boldsymbol{z}_i) + 2\hat{\mathscr{R}}_S(\mathcal{F}_{\mathcal{H}}) + 3\sqrt{\frac{\ln(2/\delta)}{2N}}\right) \leqslant 1-\delta$$

- 进一步，对于**二分类确定性问题**，即 $h\in\mathcal{H}$ 将 $\boldsymbol{x}$ 映射到 $\{-1,+1\}$ 上，并且取函数集合 $\mathcal{F}_{\mathcal{H}}$ 中的损失函数为 $0-1$ 损失，即 $\mathcal{F}_{\mathcal{H}} = \{(\boldsymbol{x},y)\to\mathbb{I}(h(\boldsymbol{x})\neq y) : h\in\mathcal{H}\}$，我们有：

$$\mathbb{E}_{\boldsymbol{z}}\{f_h(\boldsymbol{z})\} = \mathbb{E}_{\boldsymbol{z}}\{\mathbb{I}(h(\boldsymbol{x})\neq y)\} = \mathbb{E}_{\boldsymbol{x}}\{\mathbb{I}(h(\boldsymbol{x})\neq y)\} = R(h)$$
$$\frac{1}{N}\sum_{i=1}^{N}f_h(\boldsymbol{z}_i) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(h(\boldsymbol{x})\neq y) = \hat{R}_S(h)$$
$$\hat{\mathscr{R}}_S(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{f_h\in\mathcal{F}}\frac{\sum_{i=1}^{N}\sigma_i f_h(\boldsymbol{z}_i)}{N}\right\} = \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left\{\sup_{h\in\mathcal{H}}\frac{\sum_{i=1}^{N}\sigma_i h(\boldsymbol{x}_i)}{N}\right\} = \frac{1}{2}\hat{\mathscr{R}}_{S_{\mathbb{X}}}(\mathcal{H})$$
$$\mathscr{R}_N(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_{S,\boldsymbol{\sigma}}\left\{\sup_{f_h\in\mathcal{F}}\frac{\sum_{i=1}^{N}\sigma_i f_h(\boldsymbol{z}_i)}{N}\right\} = \frac{1}{2}\mathbb{E}_{S_{\mathbb{X}},\boldsymbol{\sigma}}\left\{\sup_{h\in\mathcal{H}}\frac{\sum_{i=1}^{N}\sigma_i h(\boldsymbol{x}_i)}{N}\right\} = \frac{1}{2}\mathscr{R}_N(\mathcal{H})$$

其中，样本 $S = (\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N) = ((\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_N,y_N))$，其在 $\mathbb{X}$ 上的投影 $S_{\mathbb{X}} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)$；第三个式子给出了函数集合 $\mathcal{F}_{\mathcal{H}}$ 在样本 S 上的经验 Rademacher 复杂度和函数集合 $\mathcal{H}$ 在样本 $S_{\mathbb{X}} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)$ 上的经验 Rademacher 复杂度的关系，而第四个式子则给出了函数集合 $\mathcal{F}_{\mathcal{H}}$ 关于 $\boldsymbol{z}$ 分布的 Rademacher 复杂度与函数集合 $\mathcal{H}$ 关于 $\boldsymbol{x}$ 分布的 Rademacher 复杂度之间的关系。**由此，我**

们得到了 $\mathcal{H}$ 中所有假设的两种泛化误差界：

$$\mathbb{P}\left(\forall h \in \mathcal{H},\ \ R(h) \leqslant \hat{R}_S(h) + \mathscr{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2N}}\right) \leqslant 1-\delta$$

$$\mathbb{P}\left(\forall h \in \mathcal{H},\ \ R(h) \leqslant \hat{R}_S(h) + \hat{\mathscr{R}}_{S_{\mathbb{X}}}(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2N}}\right) \leqslant 1-\delta$$

其中，$\mathscr{R}_N(\mathcal{H})$ 与分布相关，而 $\hat{\mathscr{R}}_{S_{\mathbb{X}}}(\mathcal{H})$ 则与具体的采样数据相关。However, the computation of the empirical Rademacher complexity $\hat{\mathscr{R}}_{S_{\mathbb{X}}}(\mathcal{H})$ is NP-hard for some hypothesis sets，而 $\mathscr{R}_N(\mathcal{H})$ 显然比 $\hat{\mathscr{R}}_{S_{\mathbb{X}}}(\mathcal{H})$ 更难计算。

基于 Rademacher 复杂度的泛化误差界 $\epsilon_r = \mathscr{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2N}}$ 因包含分布的信息，会比基于 VC 维的泛化误差界 $\epsilon_{vc} = \sqrt{\frac{8d\ln(2eN/d) + 8\ln(4/\delta)}{N}}$ 更"紧"。事实上，假设空间 $\mathcal{H}$ 的 Rademacher 复杂度与增长函数有如下关系：

$$\mathscr{R}_N(\mathcal{H}) \leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(N)}{N}}$$

则：

$$\begin{aligned}\epsilon_r = \mathscr{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2N}} &\leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(N)}{N}} + \sqrt{\frac{\ln(1/\delta)}{2N}} \\ &\leqslant \sqrt{\frac{2d\ln(eN/d)}{N}} + \sqrt{\frac{\ln(1/\delta)}{2N}} \triangleq \epsilon'_{vc}\ \ (\text{Sauer's lemma})\end{aligned}$$

我们得到了另一种基于 VC 维的泛化误差界 ϵ'_{vc} 。

5.3 算法稳定性

本小节思路如下，具体细节不再赘述：

基于 VC 维的泛化误差界分布无关，数据独立，而基于 Rademacher 复杂度的泛化误差界则考虑了分布和采样数据，因此相比前者会更"紧"。可以看到，上述二者均不涉及具体的学习算法，但若对某个学习问题，存在某种学习算法 $\mathcal{L}$ 在采样数据 S 发生变化时，输出的假设不会发生太大变化，即算法具有稳定性，那么我们也可以利用算法的稳定性推得泛化误差界，进而可得如下结论：**若学习算法 $\mathcal{L}$ 满足经验风险最小化且稳定，则假设空间 $\mathcal{H}$ 可学习。**