# 第十二章 计算学习理论

汇报人：陈程

2024.11.24

# 目录

# 1.基础知识

**计算学习理论（Computational learning theory）**:关于通过“计算”来进行“学习”的理论，即关于机器学习的理论基础，其目的是分析学习任务的困难本质，为学习算法提供理论保证，并根据分析结果指导算法设计。

# 1.基础知识

给定样例集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, $\boldsymbol{x}_i \in \mathcal{X}$, 本章主要讨论二分类问题, 若无特别说明, $y_i \in \mathcal{Y} = \{-1, +1\}$. 假设 $\mathcal{X}$ 中的所有样本服从一个隐含未知的分布 $\mathcal{D}$, $D$ 中所有样本都是独立地从这个分布上采样而得, 即独立同分布 (independent and identically distributed, 简称 *i.i.d.*) 样本.

令 $h$ 为从 $\mathcal{X}$ 到 $\mathcal{Y}$ 的一个映射, 其泛化误差为

$$E(h; \mathcal{D}) = P_{\boldsymbol{x} \sim \mathcal{D}}\big(h(\boldsymbol{x}) \neq y\big) \ , \tag{12.1}$$

泛化误差

$h$ 在 $D$ 上的经验误差为

$$\widehat{E}(h; D) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\big(h(\boldsymbol{x}_i) \neq y_i\big) \ . \tag{12.2}$$

经验误差

由于 $D$ 是 $\mathcal{D}$ 的独立同分布采样, 因此 $h$ 的经验误差的期望等于其泛化误差. 在上下文明确时, 我们将 $E(h; \mathcal{D})$ 和 $\widehat{E}(h; D)$ 分别简记为 $E(h)$ 和 $\widehat{E}(h)$. 令 $\epsilon$ 为 $E(h)$ 的上限, 即 $E(h) \leqslant \epsilon$; 我们通常用 $\epsilon$ 表示预先设定的学得模型所应满足的误差要求, 亦称“误差参数”.

# 2.PAC学习

**PAC（Probably Approximately Correct，概率近似正确）**框架是机器学习中一个用于分析学习算法的理论框架。目的是为了形式化地研究在有限样本情况下学习算法的性质和界限。PAC框架为理解学习算法在统计意义上的行为提供了一个数学基础。

令 $c$ 表示“概念”(concept), 这是从样本空间 $\mathcal{X}$ 到标记空间 $\mathcal{Y}$ 的映射, 它决定示例 $\boldsymbol{x}$ 的真实标记 $y$, 若对任何样例 $(\boldsymbol{x}, y)$ 有 $c(\boldsymbol{x}) = y$ 成立, 则称 $c$ 为目标概念; 所有我们希望学得的目标概念所构成的集合称为“概念类”(concept class), 用符号 $\mathcal{C}$ 表示.

给定学习算法 $\mathfrak{L}$, 它所考虑的所有可能概念的集合称为“假设空间”(hypothesis space), 用符号 $\mathcal{H}$ 表示. 由于学习算法事先并不知道概念类的真实存在, 因此 $\mathcal{H}$ 和 $\mathcal{C}$ 通常是不同的, 学习算法会把自认为可能的目标概念集中起来构成 $\mathcal{H}$, 对 $h \in \mathcal{H}$, 由于并不能确定它是否真是目标概念, 因此称为“假设”(hypothesis). 显然, 假设 $h$ 也是从样本空间 $\mathcal{X}$ 到标记空间 $\mathcal{Y}$ 的映射.

**目标概念**

**概念类**

**假设空间**

# 2.PAC学习

**定义 12.1** **PAC 辨识** (PAC Identify): 对 $0 < \epsilon, \delta < 1$, 所有 $c \in \mathcal{C}$ 和分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$, 其输出假设 $h \in \mathcal{H}$ 满足

$$P(E(h) \leqslant \epsilon) \geqslant 1 - \delta \,, \tag{12.9}$$

则称学习算法 $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$.

对于一个概念类 C和目标概念 c∈C，如果存在一个算法 L，对于任意的 ϵ>0和 δ>0，算法 L 能够使用大小为 m 的样本 S以至少 1−δ的概率找到一个假设 h∈C，使得 h的误差小于 ϵ，那么我们说 c 是PAC可辨识的。

若ϵ=0.05，δ=0.1，则称算法L有90%的概率将模型训练到95%的正确率。

# 2.PAC学习

**定义 12.2 PAC 可学习** (PAC Learnable): 令 $m$ 表示从分布 $\mathcal{D}$ 中独立同分布采样得到的样例数目, $0 < \epsilon, \delta < 1$, 对所有分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$ 和多项式函数 $\text{poly}(\cdot, \cdot, \cdot, \cdot)$, 使得对于任何 $m \geqslant \text{poly}(1/\epsilon, 1/\delta, \text{size}(\boldsymbol{x}), \text{size}(c))$, $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$, 则称概念类 $\mathcal{C}$ 对假设空间 $\mathcal{H}$ 而言是 PAC 可学习的, 有时也简称概念类 $\mathcal{C}$ 是 PAC 可学习的.

**样本数量和训练速度、训练准确率、模型复杂度在多项式级别相关。**

# 2.PAC学习

- 有了上面两个定义，就可以定义PAC学习算法：
- 若存在$\mathcal{L}$使得$\mathcal{C}$ 对 $\mathcal{H}$ PAC可学习，且$\mathcal{L}$的运行时间是多项式函数$poly\left(\frac{1}{\epsilon},\frac{1}{\delta},size(\boldsymbol{x}),size(c)\right)$ ⇒ 称 $\mathcal{C}$是高效PAC可学习的，$\mathcal{L}$为$\mathcal{C}$的PAC学习算法
- 显然，$\mathcal{C}$ 对 $\mathcal{H}$ PAC可学习意味着$\forall m \geq poly\left(\frac{1}{\epsilon},\frac{1}{\delta},size(\boldsymbol{x}),size(c)\right)$，因此PAC学习算法实际上意味着 $\mathcal{L}$ 的运行时间和样本条数有关。
- 我们称满足条件的最小的$m$为$\mathcal{L}$的样本复杂度。

# 3.有限假设空间

PAC中一个关键问题是假设空间H的复杂度，如果H有限，我们称其为有限假设空间，否则为无限假设空间。

有限假设空间有两种情况：
一种是存在c∈H，即目标概念在假设空间内，此时存在假设能够对样本数据完美学习，这称为可分情形。
另一种是c ∉H，即任何一个假设都出现或多或少的错误，这称为不可分情形。

到底需多少样例才能学得目标概念 $c$ 的有效近似呢？对 PAC 学习来说，只要训练集 $D$ 的规模能使学习算法 $\mathfrak{L}$ 以概率 $1-\delta$ 找到目标假设的 $\epsilon$ 近似即可.

# 3.有限假设空间——可分情形

我们先估计泛化误差大于 $\epsilon$ 但在训练集上仍表现完美的假设出现的概率. 假定 $h$ 的泛化误差大于 $\epsilon$, 对分布 $\mathcal{D}$ 上随机采样而得的任何样例 $(\boldsymbol{x}, y)$, 有

$$\begin{aligned} P(h(\boldsymbol{x})=y) &= 1-P(h(\boldsymbol{x})\neq y) \\ &= 1-E(h) \\ &< 1-\epsilon\,. \end{aligned} \tag{12.10}$$

由于 $D$ 包含 $m$ 个从 $\mathcal{D}$ 独立同分布采样而得的样例, 因此, $h$ 与 $D$ 表现一致的概率为

$$\begin{aligned} P\big((h(\boldsymbol{x}_1)=y_1)\wedge\ldots\wedge(h(\boldsymbol{x}_m)=y_m)\big) &= \big(1-P\left(h\left(\boldsymbol{x}\right)\neq y\right)\big)^m \\ &< (1-\epsilon)^m\,. \end{aligned} \tag{12.11}$$

因为是可分情形，c∈H，所以H中至少存在一个h=c，考虑到PAC定义，不需要相等，只要满足误差要求即可，那么所有假设泛化误差可写成 $P\big(h\in\mathcal{H}: E(h)>\epsilon\wedge\hat{E}(h)=0\big)$

考虑到h相互独立，$P\big(h\in\mathcal{H}: E(h)>\epsilon\wedge\hat{E}(h)=0\big)=\sum_{i}^{|\mathcal{H}|}P\big(E(h_i)>\epsilon\wedge\hat{E}(h_i)=0\big)$

事实上，$P\big(E(h_i)>\epsilon\wedge\hat{E}(h_i)=0\big)$ 就是前提h经验误差为0，泛化误差大于 ϵ的概率。

# 3.有限假设空间——可分情形

我们事先并不知道学习算法 $\mathfrak{L}$ 会输出 $\mathcal{H}$ 中的哪个假设, 但仅需保证泛化误差大于 $\epsilon$, 且在训练集上表现完美的所有假设出现概率之和不大于 $\delta$ 即可:

$$P\left(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0\right) < |\mathcal{H}|(1-\epsilon)^m$$
$$< |\mathcal{H}| e^{-m\epsilon} , \tag{12.12}$$

令式(12.12)不大于 $\delta$, 即

$$|\mathcal{H}| e^{-m\epsilon} \leqslant \delta , \tag{12.13}$$

可得

$$m \geqslant \frac{1}{\epsilon}\left(\ln |\mathcal{H}| + \ln \frac{1}{\delta}\right). \tag{12.14}$$

由此可知, 有限假设空间 $\mathcal{H}$ 都是 PAC 可学习的, 所需的样例数目如式(12.14)所示, 输出假设 $h$ 的泛化误差随样例数目的增多而收敛到 0, 收敛速率为 $O(\frac{1}{m})$.

**不可分的时候，c ∉H，也就是说H中的任何一个假设在样本集中都没法达到100%的准确率，那么如何处理?**

# 3.有限假设空间——不可分情形

**霍夫丁不等式**:

- Hoeffding 不等式 [Hoeffding, 1963]: 若 $x_1, x_2, \ldots, x_m$ 为 $m$ 个独立随机变量, 且满足 $0 \leqslant x_i \leqslant 1$, 则对任意 $\epsilon > 0$, 有

$$P\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i) \geqslant \epsilon\right) \leqslant \exp(-2m\epsilon^2)\ , \tag{12.5}$$

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i)\right| \geqslant \epsilon\right) \leqslant 2\exp(-2m\epsilon^2)\ . \tag{12.6}$$

**引理 12.1** 若训练集 $D$ 包含 $m$ 个从分布 $\mathcal{D}$ 上独立同分布采样而得的样例, $0 < \epsilon < 1$, 则对任意 $h \in \mathcal{H}$, 有

$$P\big(\widehat{E}(h) - E(h) \geqslant \epsilon\big) \leqslant \exp(-2m\epsilon^2)\ , \tag{12.15}$$

$$P\big(E(h) - \widehat{E}(h) \geqslant \epsilon\big) \leqslant \exp(-2m\epsilon^2)\ , \tag{12.16}$$

$$P\Big(\big|E(h) - \widehat{E}(h)\big| \geqslant \epsilon\Big) \leqslant 2\exp(-2m\epsilon^2)\ . \tag{12.17}$$

# 3.有限假设空间——不可分情形

**推论 12.1** 若训练集 $D$ 包含 $m$ 个从分布 $\mathcal{D}$ 上独立同分布采样而得的样例, $0<\epsilon<1$, 则对任意 $h \in \mathcal{H}$, 式(12.18)以至少 $1-\delta$ 的概率成立:

$$\widehat{E}(h)-\sqrt{\frac{\ln(2/\delta)}{2m}} \leqslant E(h) \leqslant \widehat{E}(h)+\sqrt{\frac{\ln(2/\delta)}{2m}} . \qquad (12.18)$$

**推论 12.1** 若训练集 $D$ 包含 $m$ 个从分布 $\mathcal{D}$ 上独立同分布采样而得的样例, $0<\epsilon<1$, 则对任意 $h \in \mathcal{H}$, 式(12.18)以至少 $1-\delta$ 的概率成立:

$$\widehat{E}(h)-\sqrt{\frac{\ln(2/\delta)}{2m}} \leqslant E(h) \leqslant \widehat{E}(h)+\sqrt{\frac{\ln(2/\delta)}{2m}} . \qquad (12.18)$$

**也就是说，样本数量越多，经验误差越接近泛化误差。**

# 3.有限假设空间——不可分情形

**定理 12.1** 若 $\mathcal{H}$ 为有限假设空间, $0<\delta<1$, 则对任意 $h\in\mathcal{H}$, 有

$$P\left(\left|E(h)-\widehat{E}(h)\right|\leqslant\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right)\geqslant 1-\delta\,. \tag{12.19}$$

**证明** 令 $h_1,h_2,\ldots,h_{|\mathcal{H}|}$ 表示假设空间 $\mathcal{H}$ 中的假设, 有

$$\begin{aligned}
&P\left(\exists h\in\mathcal{H}:\left|E(h)-\widehat{E}(h)\right|>\epsilon\right)\\
=&P\left(\left(\left|E_{h_1}-\widehat{E}_{h_1}\right|>\epsilon\right)\vee\ldots\vee\left(\left|E_{h_{|\mathcal{H}|}}-\widehat{E}_{h_{|\mathcal{H}|}}\right|>\epsilon\right)\right)\\
\leqslant&\sum_{h\in\mathcal{H}}P\left(\left|E(h)-\widehat{E}(h)\right|>\epsilon\right)\,,
\end{aligned}$$

由式(12.17)可得

$$\sum_{h\in\mathcal{H}}P\left(\left|E(h)-\widehat{E}(h)\right|>\epsilon\right)\leqslant 2|\mathcal{H}|\exp(-2m\epsilon^2)\,,$$

于是, 令 $\delta=2|\mathcal{H}|\exp(-2m\epsilon^2)$ 即可得式(12.19).

显然在不可分情形下，L无法学习到c，但是在argminE(h)下我们能够得到一个泛化误差最小的作为解，这即是不可分情形下的学习。

# 3.有限假设空间——不可分情形

**定义 12.5** 不可知 PAC 可学习 (agnostic PAC learnable): 令 $m$ 表示从分布 $\mathcal{D}$ 中独立同分布采样得到的样例数目, $0 < \epsilon, \delta < 1$, 对所有分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$ 和多项式函数 $\text{poly}(\cdot,\cdot,\cdot,\cdot)$, 使得对于任何 $m \geqslant \text{poly}(1/\epsilon, 1/\delta, \text{size}(\boldsymbol{x}), \text{size}(c))$, $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中输出满足式(12.20)的假设 $h$:

$$P\big(E(h) - \min_{h' \in \mathcal{H}} E(h') \leqslant \epsilon\big) \geqslant 1 - \delta\,, \tag{12.20}$$

则称假设空间 $\mathcal{H}$ 是不可知 PAC 可学习的.

定义：在不可知PAC可学习的情况下，学习算法不知道完整的概念类，但仍然能够在满足PAC学习标准的情况下学习。也就是说，算法能够以高概率输出一个**近似正确的假设**，而这个假设来自于一个未知的或者部分未知的概念类。

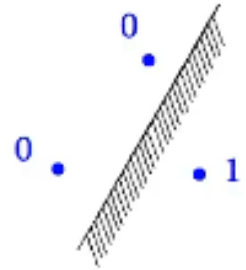
与标准PAC学习的区别：
1.标准PAC学习假设概念类是已知的，而不可知PAC学习放宽了这个假设，使其更加符合实际情况。
2.在标准PAC学习中，学习算法的目标是找到概念类中的一个概念，而在不可知PAC学习中，学习算法的目标是找到一个能够泛化到未知概念的良好假设。

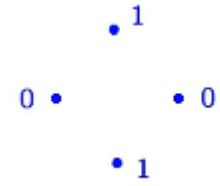
# 4.VC维

## H无限时如何处理?

**VC维**是指一个假设空间H能够“**打散**”（shatter）的最大数据点集的大小。具体来说，如果存在一个数据点集S，其大小为d，并且H 中的每一个函数都能将S 中的点以所有可能的方式分开（即对于S 中的任何子集，都存在一个假设能够正确地标记该子集的所有点为正例，而其余点为负例），那么H 的VC维至少为d。

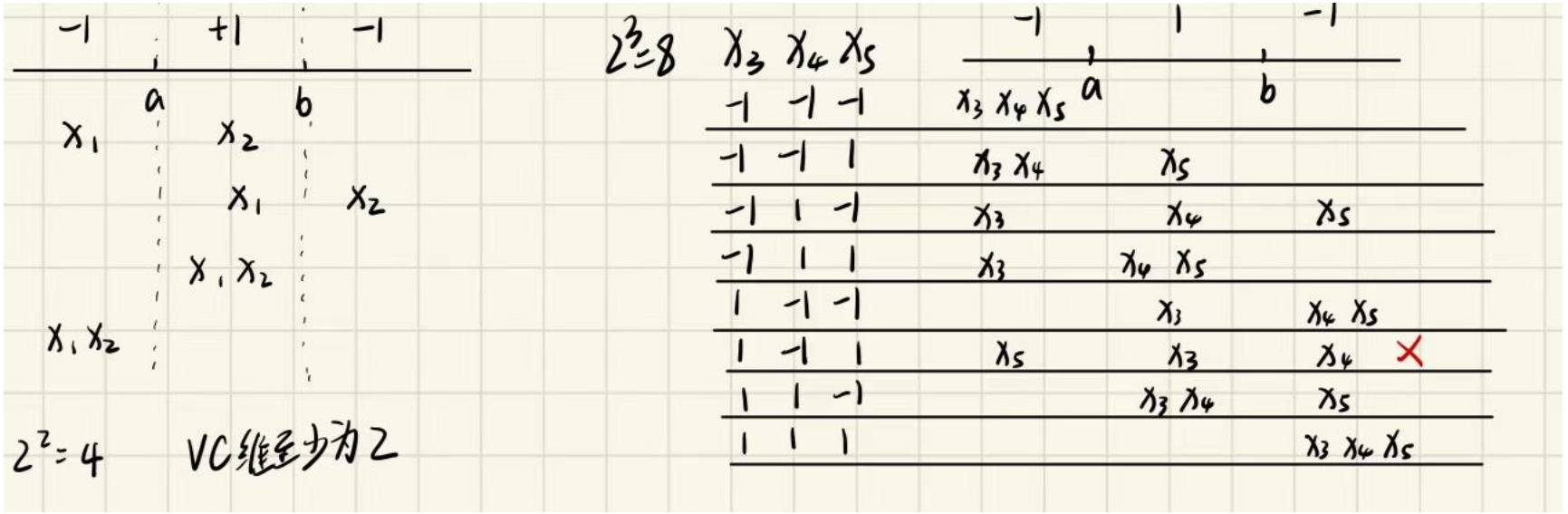
- 考虑一个简单的二维数据点集 $S$，包含3个点。假设空间 $H$ 是所有可能的线性分类器。对于这3个点，我们可以有以下几种标记方式：
  - 所有3个点都是正例。
  - 第1个点是正例，其余两个是负例。
  - 第2个点是正例，其余两个是负例。
  - 第3个点是正例，其余两个是负例。
  - 所有3个点都是负例。
- 如果存在一个线性分类器 $h \in H$，对于上述每一种标记方式，都能正确地将点分类，那么我们说 $H$ 能够打散 $S$。



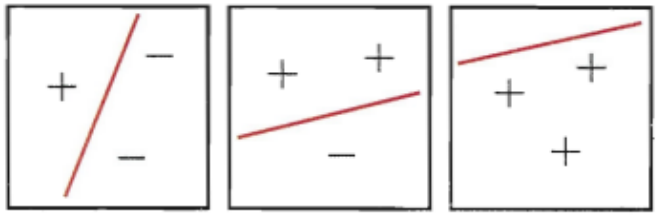
# 4.VC维

**例 12.1** 实数域中的区间 $[a,b]$: 令 $\mathcal{H}$ 表示实数域中所有闭区间构成的集合 $\{h_{[a,b]} : a,b \in \mathbb{R}, a \leqslant b\}$, $\mathcal{X} = \mathbb{R}$. 对 $x \in \mathcal{X}$, 若 $x \in [a,b]$, 则 $h_{[a,b]}(x) = +1$, 否则 $h_{[a,b]}(x) = -1$. 令 $x_1 = 0.5$, $x_2 = 1.5$, 则假设空间 $\mathcal{H}$ 中存在假设 $\{h_{[0,1]}, h_{[0,2]}, h_{[1,2]}, h_{[2,3]}\}$ 将 $\{x_1, x_2\}$ 打散, 所以假设空间 $\mathcal{H}$ 的 VC 维至少为 2; 对任意大小为 3 的示例集 $\{x_3, x_4, x_5\}$, 不妨设 $x_3 < x_4 < x_5$, 则 $\mathcal{H}$ 中不存在任何假设 $h_{[a,b]}$ 能实现对分结果 $\{(x_3,+),(x_4,-),(x_5,+)\}$. 于是, $\mathcal{H}$ 的 VC 维为 2.



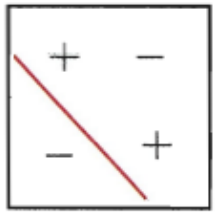
# 4.VC维

**例 12.2** 二维实平面上的线性划分: 令 $\mathcal{H}$ 表示二维实平面上所有线性划分构成的集合, $\mathcal{X} = \mathbb{R}^2$. 由图 12.1 可知, 存在大小为 3 的示例集可被 $\mathcal{H}$ 打散, 但不存在大小为 4 的示例集可被 $\mathcal{H}$ 打散. 于是, 二维实平面上所有线性划分构成的假设空间 $\mathcal{H}$ 的 VC 维为 3.



存在这样的集合, 其 $2^3 = 8$ 种对分均可被线性划分实现

(a) 示例集大小为 3



对任何集合, 其 $2^4 = 16$ 种对分中至少有一种不能被线性划分实现

(b) 示例集大小为 4

**图 12.1** 二维实平面上所有线性划分构成的假设空间的 VC 维为 3

如果一个假设空间能够打散所有大小为d 的数据点集，那么它的VC维至少为d。如果存在一个大小为d+1 的数据点集，假设空间不能打散它，那么假设空间的VC维至多为d。

VC维的计算是**基于数据点集的分类能力**，而不是假设空间的大小。因此，即使假设空间是无限的，只要存在足够大的数据点集，VC维仍然可以被计算出来。

如果一个假设空间能够打散的数据点集越大，它的VC维越高，其表达能力越强，但同时也可能导致过拟合问题，即模型在训练数据上表现很好，但在新数据上表现不佳。

# 5.Rademacher复杂度

**定义**：

- Rademacher复杂度是指在给定数据集上，一个函数类 $F$ 中所有函数值与随机正交基向量的内积的平均值。
- 具体来说，对于一个函数类 $F$，其Rademacher复杂度 $R(F)$ 定义为：

$$R(F) = \mathbb{E}_{x,\sigma}\left[\sup_{f \in F} \sum_{i=1}^{n} \sigma_i f(x_i)\right]$$

其中 $x_1, x_2, \ldots, x_n$ 是数据点，$\sigma_1, \sigma_2, \ldots, \sigma_n$ 是独立的随机变量，每个变量在 $\pm 1$ 中取值，$f(x_i)$ 是函数类 $F$ 中的函数在点 $x_i$ 上的值。

**性质**：

- Rademacher复杂度与函数类的VC维有关，但通常比VC维更容易计算。
- Rademacher复杂度可以用来估计函数类在未见数据上的泛化误差。

# 谢谢！