第 4 章　泛化界

对于学习算法来说, 判断其性能好坏的依据是泛化误差, 即学习算法基于训练集学习得到的模型在未见数据上的预测能力. 由第 2 章介绍的 PAC 学习理论可知, 泛化误差依赖于学习算法所考虑的假设空间及训练集的大小, 这使得评估学得模型的泛化误差较为困难. 一般来说, 泛化误差与学习算法 $\mathfrak{L}$ 所考虑的假设空间 $\mathcal{H}$、训练集大小 m 以及数据分布 $\mathcal{D}$ 有关. 到底是如何相关的呢? 本章就来讨论这一重要问题. 下面将按照泛化误差上界和下界分别展开讨论.

4.1 泛化误差上界

4.1.1 有限假设空间

对于假设空间 $\mathcal{H}$, 由 2.1 和 2.2 节的内容可知 $\mathcal{H}$ 分为 **有限假设空间** 和 **无限假设空间**, 根据目标概念 c 是否在 $\mathcal{H}$ 中又可以分为 **可分** 与 **不可分** 情形.

可分情形

对于可分的有限假设空间 $\mathcal{H}$, 目标概念 $c \in \mathcal{H}$, 任何在训练集 D 上犯错的假设都肯定不是要找的目标概念. 因此可以剔除这些在训练集 D 上犯错的假设, 最终留下与 D 一致的假设, 目标概念一定存在于这些一致的假设中. 如果 D 足够大, 则最终剩下的一致假设会很少, 从而能够以较大的概率找到目标概念的近似. 然而由于实际中数据集 D 通常只包括有限数量的样本, 所以假设空间 $\mathcal{H}$ 中会剩下不止一个与 D 一致的“等效”假设, 这些“等效”假设无法通过数据集 D 再进行区分. 一般来说, 无法强求通过训练集 D 能精确找到目标概念 c. 在 PAC 学习理论中, 只要训练集 D 的规模能使学习算法 $\mathfrak{L}$ 以至少 $1-\delta$ 的概率找到目标概念的 ϵ 近似即可. 当 $\mathcal{H}$ 为可分的有限假设空间时, 有下面的定理成立:

参见 2.1 节.

关于“等效”和 PAC 学习理论参见 2.2 节.

定理 4.1　令 $\mathcal{H}$ 为可分的有限假设空间, D 为从 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的训练集, 学习算法 $\mathfrak{L}$ 基于训练集 D 输出与训练集一致的假设 $h \in \mathcal{H}$, 对于 $0 < \epsilon, \delta < 1$, 若 $m \geqslant \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$, 则有

$$P(E(h) \leqslant \epsilon) \geqslant 1-\delta\,, \tag{4.1}$$

即 $E(h) \leqslant \epsilon$ 以至少 $1-\delta$ 的概率成立.

证明 学习算法 $\mathfrak{L}$ 输出与训练集一致的假设 $h \in \mathcal{H}$, 该假设的泛化误差依赖于训练集 D, 我们希望能够以较大的概率找到与目标概念 ϵ 近似的假设. 若 h 的泛化误差大于 ϵ 且与训练集一致, 则这样的假设出现的概率可以表示为

$$P(\exists\, h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0)\ . \tag{4.2}$$

$E(h) > \epsilon$ 意味着 h 犯错的概率大于 ϵ, 即预测正确的概率小于等于 $1-\epsilon$. $\hat{E}(h) = 0$ 表示对 m 个独立同分布采样得到的样本 h 均预测正确, 该事件发生概率不大于 $(1-\epsilon)^m$.

下面只需证明这一概率至多为 δ 即可. 通过计算可知

$$\begin{aligned} P\left(\exists\, h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0\right) &\leqslant \sum_{h \in \mathcal{H}} P\left(E(h) > \epsilon \wedge \widehat{E}(h) = 0\right) \\ &< |\mathcal{H}|(1-\epsilon)^m\ . \end{aligned} \tag{4.3}$$

因此只需要保证 (4.3) 最右端不大于 δ 即可. 由于 $(1-\epsilon)^m \leqslant e^{-\epsilon m}$, 若 $m \geqslant \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$, 则有

$$|\mathcal{H}|(1-\epsilon)^m \leqslant |\mathcal{H}|e^{-\epsilon m} \leqslant \delta\ . \tag{4.4}$$

从而可知 $P(E(h) > \epsilon) \leqslant \delta$, 即 $P(E(h) \leqslant \epsilon) \geqslant 1-\delta$, 定理得证. □

参见 2.2 节, 不难发现这里的 m 是关于 $1/\delta$ 和 $1/\epsilon$ 的多项式, 因此有限可分的假设空间 $\mathcal{H}$ 是 PAC 可学的.

这一定理表明假设空间 $\mathcal{H}$ 是有限可分时, 学习算法 $\mathfrak{L}$ 输出假设的泛化误差依赖于假设空间的大小 $|\mathcal{H}|$ 和训练集的大小 m, 随着训练集中样本数目的逐渐增加, 泛化误差的上界逐渐趋近于 0, 收敛率是 $O(1/m)$.

不可分情形

在不可分情形中, 目标概念不在假设空间中, 假设空间中的每个假设都会或多或少地出现分类错误, 我们不再奢望找到目标概念的 ϵ 近似, 而是希望找到假设空间中泛化误差最小假设的 ϵ 近似. 对于学习算法输出的假设 h 来说, 泛化误差是其在未见数据上的预测能力, 无法直接观测得到, 但其在训练集上的经验误差是可以直接观测得到的. 定理 2.1 探讨了泛化误差与经验误差之间的关系, 表明当训练集中样本数目 m 较大时, h 的经验误差是泛化误差的较好近似. 基于这一关系, 可以给出下面的定理.

定理 4.2 令 $\mathcal{H}$ 为有限假设空间, D 为从 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的训练集, $h \in \mathcal{H}$, 对于 $0 < \delta < 1$ 有

$$P\left(\left|E(h)-\widehat{E}(h)\right| \leqslant \sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right) \geqslant 1-\delta\ . \tag{4.5}$$

证明 将 $\mathcal{H}$ 中的有限假设记为 $h_1, h_2, \ldots, h_{|\mathcal{H}|}$, 通过计算可得

$$\begin{aligned}
&P\left(\exists\, h\in\mathcal{H}: \left|\widehat{E}(h)-E(h)\right|>\epsilon\right)\\
&=P\left(\left(\left|\widehat{E}(h_1)-E(h_1)\right|>\epsilon\right)\vee\cdots\vee\left(\left|\widehat{E}(h_{|\mathcal{H}|})-E(h_{|\mathcal{H}|})\right|>\epsilon\right)\right)\\
&\leqslant \sum_{h\in\mathcal{H}} P\left(\left|\widehat{E}(h)-E(h)\right|>\epsilon\right)\ .
\end{aligned} \tag{4.6}$$

联合界不等式 (1.19).

基于引理 2.1, 令 $2\exp(-2m\epsilon^2)=\delta/|\mathcal{H}|$ 可得

$$\begin{aligned}
&\sum_{h\in\mathcal{H}} P\left(\left|\widehat{E}(h)-E(h)\right|>\epsilon\right)\\
&\leqslant \sum_{h\in\mathcal{H}} \delta/|\mathcal{H}| \leqslant |\mathcal{H}|\cdot\delta/|\mathcal{H}| = \delta\ .
\end{aligned} \tag{4.7}$$

由 $2\exp(-2m\epsilon^2)=\delta/|\mathcal{H}|$ 可求解 $\epsilon=\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}$, 从而定理得证. □

由定理 4.2 可知 $E(h)\leqslant\widehat{E}(h)+\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}$ 以至少 $1-\delta$ 的概率成立. 由于 $\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}=O(1/\sqrt{m})$, 所以在有限不可分情形下, 泛化误差的收敛率为 $O(1/\sqrt{m})$.

4.1.2 无限假设空间

对于无限假设空间, 需要从 VC 维和 Rademacher 复杂度的角度来分析其泛化误差界.

有限 VC 维假设空间的泛化误差界

在 3.1 节介绍了增长函数和 VC 维, 定理 3.1 表明 VC 维与增长函数密切相关. 接下来, 我们首先介绍关于增长函数与泛化误差之间关系的引理 4.1 [Vapnik and Chervonenkis, 1971].

引理 4.1 对于假设空间 $\mathcal{H}$, $h\in\mathcal{H}$, $m\in\mathbb{N}$ 和 $0<\epsilon<1$, 当 $m\geqslant 2/\epsilon^2$ 时有

$$P\left(\left|E(h)-\widehat{E}(h)\right|>\epsilon\right)\leqslant 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)\ . \tag{4.8}$$

证明 考虑两个大小均为 m 且分别从 $\mathcal{D}$ 独立同分布采样得到的训练集 D 和 D'. 首先证明

$$P\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\geqslant\frac{1}{2}P\left(\sup_{h\in\mathcal{H}}\left|E(h)-\widehat{E}_D(h)\right|>\epsilon\right)\ . \tag{4.9}$$

令 Q 表示集合

$D\sim\mathcal{D}^m$ 表示 D 大小为 m 且从 $\mathcal{D}$ 独立同分布采样得到.

$$\left\{D\sim\mathcal{D}^m\middle|\sup_{h\in\mathcal{H}}\left|E(h)-\widehat{E}_D(h)\right|>\epsilon\right\}\ , \tag{4.10}$$

计算可得

$$\begin{aligned}
&P\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\\
&=\mathbb{E}_{D,D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\qquad\text{概率与期望之间的转化.}\\
&=\mathbb{E}_{D\sim\mathcal{D}^m}\left[\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\right]\\
&\geqslant\mathbb{E}_{D\in Q}\left[\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\right]\ .\qquad\text{$Q\subset\mathcal{D}^m$.}
\end{aligned}\tag{4.11}$$

根据 Q 的定义可知, 对于任意 $D\in Q$, 存在一个假设 $h_0\in\mathcal{H}$ 使得 $\left|E(h_0)-\widehat{E}_D(h_0)\right|>\epsilon$. 对于 h_0, 计算可得

$$\begin{aligned}
&\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\\
&\geqslant\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\left|\widehat{E}_D(h_0)-\widehat{E}_{D'}(h_0)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\\
&=\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\left|\widehat{E}_D(h_0)-E(h_0)-\left(\widehat{E}_{D'}(h_0)-E(h_0)\right)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\\
&\geqslant\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\left|\widehat{E}_D(h_0)-E(h_0)\right|-\left|\left(\widehat{E}_{D'}(h_0)-E(h_0)\right)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\ .\qquad\text{基于 $|A-B|\geqslant|A|-|B|$.}
\end{aligned}\tag{4.12}$$

注意 $\left|E(h_0)-\widehat{E}_D(h_0)\right|>\epsilon$, 若 $\left|\widehat{E}_{D'}(h_0)-E(h_0)\right|\leqslant\frac{1}{2}\epsilon$, 则 $\left|\widehat{E}_D(h_0)-E(h_0)\right|-\left|\left(\widehat{E}_{D'}(h_0)-E(h_0)\right)\right|\geqslant\frac{1}{2}\epsilon$ 成立. 从而基于 (4.12) 可得

$$\mathbb{E}_{D'\sim\mathcal{D}^m}\left[\mathbb{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]$$

$$
\begin{aligned}
&\geqslant \mathbb{E}_{D' \sim \mathcal{D}^m}\left[\mathbb{I}\left(\left|\widehat{E}_D(h_0)-E(h_0)\right|-\left|\left(\widehat{E}_{D'}(h_0)-E(h_0)\right)\right| \geqslant \frac{1}{2}\epsilon\right)\right] \\
&\geqslant \mathbb{E}_{D' \sim \mathcal{D}^m}\left[\mathbb{I}\left(\left|\widehat{E}_{D'}(h_0)-E(h_0)\right| \leqslant \frac{1}{2}\epsilon\right)\right] \\
&= P\left(\left|\widehat{E}_{D'}(h_0)-E(h_0)\right| \leqslant \frac{1}{2}\epsilon\right) \\
&= 1-P\left(\left|\widehat{E}_{D'}(h_0)-E(h_0)\right| > \frac{1}{2}\epsilon\right) . \qquad (4.13)
\end{aligned}
$$

概率与期望之间的转化.

再由 Chebyshev 不等式 (1.21) 可得

$$
\begin{aligned}
&P\left(\left|\widehat{E}_{D'}(h_0)-E(h_0)\right| > \frac{1}{2}\epsilon\right) \\
\leqslant &\frac{4\left(1-E(h_0)\right)E(h_0)}{\epsilon^2 m} \\
\leqslant &\frac{1}{\epsilon^2 m} . \qquad (4.14)
\end{aligned}
$$

当 $m \geqslant 2/\epsilon^2$ 时, $P\left(\left|\widehat{E}_{D'}(h_0)-E(h_0)\right| > \frac{1}{2}\epsilon\right) \leqslant 1/2$. 于是可得

$$
\begin{aligned}
&P\left(\sup_{h \in \mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right| \geqslant \frac{1}{2}\epsilon\right) \\
\geqslant &\mathbb{E}_{D \in Q}\left[\frac{1}{2}\right] \\
= &\frac{1}{2}P\left(\sup_{h \in \mathcal{H}}\left|E(h)-\widehat{E}_D(h)\right| > \epsilon\right) . \qquad (4.15)
\end{aligned}
$$

至此, (4.9) 成立. 由于 D 和 D' 均为从 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的训练集, 则 D 和 D' 一共包含 $2m$ 个样本 (这 $2m$ 个样本有可能重复). 若令 T_i 表示这 $2m$ 个样本上的置换, 则有 $(2m)!$ 个 T_i. 令 T_iD 表示 $2m$ 个样本经过置换 T_i 的前 m 个样本, T_iD' 表示这 $2m$ 个样本经过置换 T_i 的后 m 个样本, 则对于 D, D', T_iD 和 T_iD' 有

集合 A 到自身的映射称为 A 的一个变换, 如果 A 是有限集且变换是一一变换 (双射), 则称这一变换为 A 的一个置换. 置换可以看作调换集合中样本顺序的一种方式, 对于包含 m 个样本的集合 A, 一共有 $m!$ 个置换.

$$
\begin{aligned}
&P\left(\sup_{h \in \mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right| \geqslant \frac{1}{2}\epsilon\right) \\
= &P\left(\sup_{h \in \mathcal{H}}\left|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)\right| \geqslant \frac{1}{2}\epsilon\right) . \qquad (4.16)
\end{aligned}
$$

因此有

概率与期望之间的转化.

$$
\begin{aligned}
&P\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\\
&=\mathbb{E}_{D,D'}\left[\frac{1}{(2m)!}\sum_{i=1}^{(2m)!}\mathbb{I}\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\\
&=\mathbb{E}_{D,D'}\left[\frac{1}{(2m)!}\sum_{i=1}^{(2m)!}\sup_{h\in\mathcal{H}}\mathbb{I}\left(\left|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]\\
&\leqslant\mathbb{E}_{D,D'}\left[\sum_{h\in\mathcal{H}_{|D+D'}}\frac{1}{(2m)!}\sum_{i=1}^{(2m)!}\mathbb{I}\left(\left|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\right]. \qquad (4.17)
\end{aligned}
$$

其中 $\mathcal{H}_{|D+D'}$ 为 $\mathcal{H}$ 在训练集 $D+D'$ 上的限制. 接下来考虑

$$
\sum_{i=1}^{(2m)!}\mathbb{I}\left(\left|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)\right|\geqslant\frac{1}{2}\epsilon\right). \qquad (4.18)
$$

(4.18) 表示对于给定假设 h 满足 $|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)|\geqslant\frac{1}{2}\epsilon$ 的置换数目. 令 l 表示 h 在 $D+D'$ 上预测正确的样本数目, 有

k 表示 T_iD 中被 h 预测正确的样本数目, $m-k$ 表示 T_iD 中被 h 预测错误的样本数目, $\binom{l}{k}$ 表示从 l 个预测正确的样本中选择 k 个样本的种数, $\binom{2m-l}{m-k}$ 表示从 $2m-l$ 个预测错误的样本中选择 $m-k$ 个样本的种数, $\binom{2m}{m}$ 表示从 $2m$ 个样本中选择 m 个样本构成 T_iD 的种数. 若 $|(m-k)/m-(m-l+k)/m|\geqslant\epsilon/2$, 即 $|2k/m-l/m|\geqslant\epsilon/2$, 则 $|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)|\geqslant\frac{1}{2}\epsilon$.

$$
\begin{aligned}
&\frac{1}{(2m)!}\sum_{i=1}^{(2m)!}\mathbb{I}\left(\left|\widehat{E}_{T_iD}(h)-\widehat{E}_{T_iD'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\\
&=\sum_{\substack{k\in[l]\\ \text{s.t. }|2k/m-l/m|\geqslant\epsilon/2}}\frac{\binom{l}{k}\binom{2m-l}{m-k}}{\binom{2m}{m}}\\
&\leqslant 2\exp\left(-\frac{\epsilon^2 m}{8}\right). \qquad (4.19)
\end{aligned}
$$

结合 (4.17) 可得

$$
P\left(\sup_{h\in\mathcal{H}}\left|\widehat{E}_D(h)-\widehat{E}_{D'}(h)\right|\geqslant\frac{1}{2}\epsilon\right)\leqslant 2|\mathcal{H}_{|D+D'}|\exp\left(-\frac{\epsilon^2 m}{8}\right). \qquad (4.20)
$$

由增长函数的定义 (3.2) 可知 $\left|\mathcal{H}_{|D+D'}\right|\leqslant\Pi_{\mathcal{H}}(2m)$. 再结合 (4.9), 对于任意假

设 $h \in \mathcal{H}$ 可得

$$\begin{aligned} & P\left(|E(h)-\widehat{E}_D(h)|>\epsilon\right) \\ \leqslant & P\left(\sup_{h\in\mathcal{H}}|E(h)-\widehat{E}(h)|>\epsilon\right) \\ \leqslant & 4\Pi_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8}) , \end{aligned} \tag{4.21}$$

从而定理得证. □

基于引理 4.1, 再结合关于 VC 维与增长函数之间关系的定理 3.1, 有下面的定理.

定理 4.3 若假设空间 $\mathcal{H}$ 的有限 VC 维为 d, $h \in \mathcal{H}$, 则对 $m > d$ 和 $0 < \delta < 1$ 有

$$P\left(\left|E(h)-\widehat{E}(h)\right| \leqslant \sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1-\delta . \tag{4.22}$$

证明 由定理 3.1 可知

$$4\Pi_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8}) \leqslant 4\left(\frac{2em}{d}\right)^d \exp(-\frac{m\epsilon^2}{8}) . \tag{4.23}$$

令 $4\left(\frac{2em}{d}\right)^d \exp(-\frac{m\epsilon^2}{8}) = \delta$ 可得

$$\epsilon = \sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}} . \tag{4.24}$$

将 (4.24) 带入引理 4.1, 定理得证. □

这里忽略 δ 和常数项.

由定理 4.3 可知 $E(h) \leqslant \widehat{E}(h) + O\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$ 以至少 $1-\delta$ 的概率成立, 泛化误差的收敛率为 $O\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$. 对于有限 VC 维的假设空间, 泛化误差的收敛率与 VC 维的大小有关, VC 维越大, 假设空间越复杂, 泛化误差的收敛率也越慢. 其次, 有限 VC 维的不可分假设空间比有限不可分假设空间更难收敛, 这也是无限假设空间与有限假设空间的区别.

基于 Rademacher 复杂度的泛化误差界

关于 Rademacher 复杂度参见 3.2 节 (3.29) 和 (3.30).

对于从 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的训练集 Z, 函数空间 $\mathcal{F}$ 关于 Z 的经验 Rademacher 复杂度和关于 $\mathcal{D}$ 的 Rademacher 复杂度分别是 $\widehat{\mathfrak{R}}_Z(\mathcal{F})$ 和 $\mathfrak{R}_m(\mathcal{F})$, 基于 $\widehat{\mathfrak{R}}_Z(\mathcal{F})$ 和 $\mathfrak{R}_m(\mathcal{F})$ 可以分析关于函数空间 $\mathcal{F}$ 的泛化误差界 [Mohri et al., 2018].

定理 4.4 对于实值函数空间 $\mathcal{F}: \mathcal{Z} \mapsto [0,1]$, 从分布 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的训练集 $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$, $\boldsymbol{z}_i \in \mathcal{Z}$, $f \in \mathcal{F}$ 和 $0 < \delta < 1$, 以至少 $1-\delta$ 的概率有

$$\mathbb{E}\left[f(\boldsymbol{z})\right] \leqslant \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i) + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}\,, \tag{4.25}$$

$$\mathbb{E}\left[f(\boldsymbol{z})\right] \leqslant \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i) + 2\widehat{\mathfrak{R}}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}\,. \tag{4.26}$$

证明 令

$$\widehat{E}_Z(f) = \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i)\,, \tag{4.27}$$

$$\Phi(Z) = \sup_{f\in\mathcal{F}}\left(\mathbb{E}[f] - \widehat{E}_Z(f)\right)\,, \tag{4.28}$$

Z' 为与 Z 仅有一个样本不同的训练集, 不妨设 $\boldsymbol{z}_m \in Z$ 和 $\boldsymbol{z}'_m \in Z'$ 为不同样本, 可得

$$\begin{aligned}
&\Phi(Z') - \Phi(Z)\\
=&\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f] - \widehat{E}_{Z'}(f)\right) - \sup_{f\in\mathcal{F}}\left(\mathbb{E}[f] - \widehat{E}_Z(f)\right)\\
\leqslant&\sup_{f\in\mathcal{F}}\left(\widehat{E}_Z(f) - \widehat{E}_{Z'}(f)\right)\\
=&\sup_{f\in\mathcal{F}}\frac{f(\boldsymbol{z}_m) - f(\boldsymbol{z}'_m)}{m} \leqslant \frac{1}{m}.
\end{aligned} \tag{4.29}$$

同理可得

$$\Phi(Z) - \Phi(Z') \leqslant \frac{1}{m}\,. \tag{4.30}$$

从而可知

$$\left|\Phi(Z)-\Phi(Z')\right| \leqslant \frac{1}{m}\,. \tag{4.31}$$

根据 McDiarmid 不等式 (1.32) 可知, 对于 $0<\delta<1$,

$$\Phi(Z) \leqslant \mathbb{E}_Z\left[\Phi(Z)\right] + \sqrt{\frac{\ln(1/\delta)}{2m}} \tag{4.32}$$

以至少 $1-\delta$ 的概率成立. 下面估计 $\mathbb{E}_Z[\Phi(Z)]$ 的上界

$$\begin{aligned}
&\mathbb{E}_Z\left[\Phi(Z)\right] \\
&= \mathbb{E}_Z\left[\sup_{f\in\mathcal{F}}\Big(\mathbb{E}[f]-\widehat{E}_Z(f)\Big)\right] \\
&= \mathbb{E}_Z\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{Z'}\left[\widehat{E}_{Z'}[f]-\widehat{E}_Z[f]\right]\right] \\
&\leqslant \mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\Big(\widehat{E}_{Z'}[f]-\widehat{E}_Z[f]\Big)\right] && \text{Jensen 不等式 (1.11)}. \\
&= \mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\big(f(\boldsymbol{z}_i')-f(\boldsymbol{z}_i)\big)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},Z,Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(f(\boldsymbol{z}_i')-f(\boldsymbol{z}_i)\big)\right] && \boldsymbol{\sigma}=(\sigma_1;\cdots;\sigma_m),\ \sigma_i\in\{-1,+1\}. \\
&\leqslant \mathbb{E}_{\boldsymbol{\sigma},Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{z}_i')\right] + \mathbb{E}_{\boldsymbol{\sigma},Z}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i f(\boldsymbol{z}_i)\right] \\
&= 2\mathbb{E}_{\boldsymbol{\sigma},Z}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{z}_i)\right] \\
&= 2\mathfrak{R}_m(\mathcal{F})\,.
\end{aligned} \tag{4.33}$$

由 (4.27)、(4.28)、(4.32) 和 (4.33) 可知 (4.25) 得证. 根据 (3.29) 和 (4.29) 可知替换训练集中的一个样本后经验 Rademacher 复杂度最多改变 $1/m$, 即 $\left|\widehat{\mathfrak{R}}_Z(\mathcal{F})-\widehat{\mathfrak{R}}_{Z'}(\mathcal{F})\right| \leqslant 1/m$. 再根据 McDiarmid 不等式 (1.33) 可知

$$\mathfrak{R}_m(\mathcal{F}) \leqslant \widehat{\mathfrak{R}}_Z(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2m}} \tag{4.34}$$

$\mathbb{E}\left[\widehat{\mathfrak{R}}_Z(\mathcal{F})\right]=\mathfrak{R}_m(\mathcal{F})$.

以至少 $1-\delta/2$ 的概率成立. 由 (4.32) 可知

$$\Phi(Z) \leqslant \mathbb{E}_Z[\Phi(Z)] + \sqrt{\frac{\ln(2/\delta)}{2m}} \tag{4.35}$$

以至少 $1-\delta/2$ 的概率成立. 由 (4.33)~(4.35)和联合界不等式 (1.19) 得

$$\Phi(Z) \leqslant 2\widehat{\mathfrak{R}}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \tag{4.36}$$

以至少 $1-\delta$ 的概率成立, 从而 (4.26) 得证. □

定理 4.4 的适用范围是实值函数空间 $\mathcal{F}: \mathcal{Z} \mapsto [0,1]$, 一般用于回归问题. 对于分类问题有下面的定理.

定理 4.5 对于假设空间 $\mathcal{H}: \mathcal{X} \mapsto \{-1,+1\}$, 从分布 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的训练集 $D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$, $\boldsymbol{x}_i \in \mathcal{X}$, $h \in \mathcal{H}$ 和 $0<\delta<1$, 以至少 $1-\delta$ 的概率有

$$E(h) \leqslant \widehat{E}(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}\,, \tag{4.37}$$

$$E(h) \leqslant \widehat{E}(h) + \widehat{\mathfrak{R}}_D(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}\,. \tag{4.38}$$

证明 对于二分类问题的假设空间 $\mathcal{H}$, 令 $\mathcal{Z}=\mathcal{X}\times\{-1,+1\}$, $\mathcal{H}$ 中的假设 h 可以变形为 $f_h(\boldsymbol{z})=f_h(\boldsymbol{x},y)=\mathbb{I}\left(h(\boldsymbol{x})\neq y\right)$. 于是值域为 $\{-1,+1\}$ 的假设空间 $\mathcal{H}$ 转化为值域为 $[0,1]$ 的函数空间 $\mathcal{F}_{\mathcal{H}}=\{f_h: h\in\mathcal{H}\}$. 由 (3.29) 可知

$$\begin{aligned}
\widehat{\mathfrak{R}}_Z(\mathcal{F}_{\mathcal{H}}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_h\in\mathcal{F}_{\mathcal{H}}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f_h(\boldsymbol{x}_i,y_i)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \mathbb{I}\left(h(\boldsymbol{x}_i)\neq y_i\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \frac{1-y_i h(\boldsymbol{x}_i)}{2}\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i + \sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}(-y_i\sigma_i h(\boldsymbol{x}_i))\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}(-y_i\sigma_i h(\boldsymbol{x}_i))\right]
\end{aligned}$$

由于 $y_i \in \{-1,+1\}$, $-y_i\sigma_i$ 等价于 σ_i.

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}(\sigma_i h(\boldsymbol{x}_i))\right]$$
$$= \frac{1}{2}\widehat{\mathfrak{R}}_D(\mathcal{H}). \tag{4.39}$$

同时对上式两边取期望可得

$$\mathfrak{R}_Z(\mathcal{F}_{\mathcal{H}}) = \frac{1}{2}\mathfrak{R}_D(\mathcal{H})\ . \tag{4.40}$$

将 (4.40) 代入定理 4.4, 定理得证. □

4.2 泛化误差下界

基于 VC 维也可以分析泛化误差下界, 这一下界的意义在于指出对于任何学习算法存在某一数据分布, 当样本数目有限时, 学习算法不能以较大概率输出目标概念的近似, 其中的要点是如何找到这样一种数据分布. 下面将针对假设空间可分与不可分这两种情形分别进行讨论.

可分情形

首先分析可分情形下的泛化误差下界 [Ehrenfeucht et al., 1988].

定理 4.6 若假设空间 $\mathcal{H}$ 的 VC 维 $d>1$, 则对任意 $m>1$ 和学习算法 $\mathfrak{L}$, 存在分布 $\mathcal{D}$ 和目标概念 $c\in\mathcal{H}$ 使得

D 从 $\mathcal{D}$ 独立同分布采样得到, 这里泛化误差与具体的目标概念 c 相关, 因此将其记为 $E(h_D,c)$, 而不是 $E(h_D)$.

$$P\left(E(h_D,c) > \frac{d-1}{32m}\right) \geqslant \frac{1}{100}\ , \tag{4.41}$$

其中 h_D 为学习算法 $\mathfrak{L}$ 基于大小为 m 的训练集 D 输出的假设.

证明 由于 VC 维 $d>1$, 不妨令 $S=\{\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{d-1}\}\subset\mathcal{X}$ 表示能被 $\mathcal{H}$ 打散的集合. 对于 $\epsilon>0$, 下面将构造一种数据分布 $\mathcal{D}$ 使得概率质量集中在 S 上, 并且较高的概率质量 $(1-8\epsilon)$ 集中在 $\boldsymbol{x}_0$ 上, 而其余的概率质量平均分配在其他点上.

$$P_{\mathcal{D}}(\boldsymbol{x}_0) = 1-8\epsilon \ \wedge\ P_{\mathcal{D}}(\boldsymbol{x}_i) = \frac{8\epsilon}{d-1} \quad (i\in[d-1])\ . \tag{4.42}$$

目标概念可能是 $\mathcal{H}$ 中任一假设, 且 S 能够被 $\mathcal{H}$ 打散, 所以 S 中任一样本的真实标记可以是任意标记, 这样对于 S 中未出现的样本, $\mathfrak{L}$ 都可能预测错误, 因此不会优于随机结果.

根据以上构造过程, 从分布 $\mathcal{D}$ 采样得到的数据集主要包含 $\boldsymbol{x}_0$. 由于 $\mathcal{H}$ 将 S 打散, 因此对于 S 中未出现的样本, 任意学习算法 $\mathfrak{L}$ 的预测不会优于随机结果.

不失一般性, 假设学习算法 $\mathfrak{L}$ 在 $\boldsymbol{x}_0$ 上预测正确. 对于大小为 m 的训练集 D, 令 $\bar{D}$ 表示出现在 $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{d-1}\}$ 中的样本集合, $A=$

$\left\{D \sim \mathcal{D}^m \middle| \left(|D| = m\right) \wedge \left(|\bar{D}| \leqslant (d-1)/2\right)\right\}$. 对于给定的 $D \in A$, 考虑来自均匀分布 $\mathcal{U}$ 的目标概念 $c : S \mapsto \{-1, +1\}$. 由于 S 可以被 $\mathcal{H}$ 打散, 根据如上构造可得

$$\begin{aligned}
\mathbb{E}_{\mathcal{U}}\left[E(h_D, c)\right] &= \sum_{c}\sum_{\boldsymbol{x}\in S} \mathbb{I}\left(h_D(\boldsymbol{x}) \neq c(\boldsymbol{x})\right) P_{\boldsymbol{x}\sim\mathcal{D}}(\boldsymbol{x}) P_{c\sim\mathcal{U}}(c) \\
&\geqslant \sum_{c}\sum_{\boldsymbol{x}\in S-\bar{D}-\{\boldsymbol{x}_0\}} \mathbb{I}\left(h_D(\boldsymbol{x}) \neq c(\boldsymbol{x})\right) P_{\boldsymbol{x}\sim\mathcal{D}}(\boldsymbol{x}) P_{c\sim\mathcal{U}}(c) \\
&= \sum_{\boldsymbol{x}\in S-\bar{D}-\{\boldsymbol{x}_0\}} \left(\sum_{c} \mathbb{I}\left(h_D(\boldsymbol{x}) \neq c(\boldsymbol{x})\right) P_{c\sim\mathcal{U}}(c)\right) P_{\boldsymbol{x}\sim\mathcal{D}}(\boldsymbol{x}) \\
&= \frac{1}{2} \sum_{\boldsymbol{x}\in S-\bar{D}-\{\boldsymbol{x}_0\}} P_{\boldsymbol{x}\sim\mathcal{D}}(\boldsymbol{x}) \\
&\geqslant \frac{1}{2}\frac{d-1}{2}\frac{8\epsilon}{d-1} = 2\epsilon\,.
\end{aligned} \tag{4.43}$$

计算时只考虑样本 $\boldsymbol{x} \in S - \bar{D} - \{\boldsymbol{x}_0\}$ 所产生的分类错误, 而未考虑所有的样本.

对 S 中未出现的样本, $\mathfrak{L}$ 的预测不会优于随机结果.

考虑分布 $\mathcal{D}$ 的构造和 $\bar{D}$ 的大小, $P_{\mathcal{D}}(\boldsymbol{x}_i) = \frac{8\epsilon}{d-1}$, $|S - \bar{D} - \{\boldsymbol{x}_0\}| \geqslant \frac{d-1}{2}$, $1 \leqslant i \leqslant d-1$.

由于 $\mathbb{E}_{\mathcal{U}}\left[E(h_D, c)\right] \geqslant 2\epsilon$ 对于任意 $D \in A$ 均成立, 因此关于 A 的期望也成立, 即 $\mathbb{E}_{D\in A}\left[\mathbb{E}_{\mathcal{U}}\left[E(h_D, c)\right]\right] \geqslant 2\epsilon$. 根据 Fubini 定理 [Stein and Shakarchi, 2009]: 若函数 $f(x, y)$ 的期望 $\mathbb{E}_{x,y}\left[|f(x, y)|\right] < \infty$, 则

$$\mathbb{E}_x\left[\mathbb{E}_y\left[f(x, y)\right]\right] = \mathbb{E}_y\left[\mathbb{E}_x\left[f(x, y)\right]\right]\,, \tag{4.44}$$

可知交换期望计算顺序不等式依然成立, 即有

$$\mathbb{E}_{\mathcal{U}}\left[\mathbb{E}_{D\in A}\left[E(h_D, c)\right]\right] \geqslant 2\epsilon\,. \tag{4.45}$$

由于期望的下界为 2ϵ, 必定存在一个目标概念 $c^* \in \mathcal{H}$ 满足 $\mathbb{E}_{D\in A}\left[E(h_D, c^*)\right] \geqslant 2\epsilon$, 其中 $E(h_D, c^*) = \mathbb{E}_{\mathcal{D}}\mathbb{I}(h_D(\boldsymbol{x}) \neq c^*(\boldsymbol{x}))$. 下面将该期望按照 $E(h_D, c^*)$ 的取值分解成如下两部分

$$\begin{aligned}
&\mathbb{E}_{D\in A}\left[E(h_D, c^*)\right] \\
&= \sum_{D:E(h_D,c^*)>\epsilon} E(h_D, c^*)P(D) + \sum_{D:E(h_D,c^*)\leqslant\epsilon} E(h_D, c^*)P(D) \\
&\leqslant P_{\boldsymbol{x}\sim\mathcal{D}}\left(\boldsymbol{x} \in (S - \{\boldsymbol{x}_0\})\right) P_{D\in A}\left(E(h_D, c^*) > \epsilon\right) \\
&\quad + \epsilon\left(1 - P_{D\in A}\left(E(h_D, c^*) > \epsilon\right)\right) \\
&= 8\epsilon P_{D\in A}\left(E(h_D, c^*) > \epsilon\right) + \epsilon\left(1 - P_{D\in A}\left(E(h_D, c^*) > \epsilon\right)\right)
\end{aligned}$$

前一项放大至 $S - \{\boldsymbol{x}_0\}$ 全部预测错误, 即 $E(h_D, c^*) \leqslant P_{\boldsymbol{x}\sim\mathcal{D}}(\boldsymbol{x} \in (S - \{\boldsymbol{x}_0\})) = 8\epsilon$, 后一项将 $E(h_D, c^*)$ 放大至 ϵ.

$$= \epsilon + 7\epsilon P_{D\in A}\left(E(h_D, c^*) > \epsilon\right) . \tag{4.46}$$

基于 (4.46) 和 $\mathbb{E}_{D\in A}\left[E(h_D, c^*)\right] \geqslant 2\epsilon$, 可求解出

$$P_{D\in A}\left(E(h_D, c^*) > \epsilon\right) \geqslant \frac{1}{7\epsilon}(2\epsilon - \epsilon) = \frac{1}{7} . \tag{4.47}$$

因此, 在所有大小为 m 的样本集合 $\mathcal{D}^m$ 中, 满足 $E(h_D, c^*) > \epsilon$ 的样本集出现的概率为

$$\begin{aligned} & P_{D\sim\mathcal{D}^m}\left(E(h_D, c^*) > \epsilon\right) \\ \geqslant\ & P_{D\in A}(E(h_D, c^*) > \epsilon) P_{D\sim\mathcal{D}^m}(D \in A) \\ \geqslant\ & \frac{1}{7} P_{D\sim\mathcal{D}^m}(D \in A) . \end{aligned} \tag{4.48}$$

接下来, 只要找到 $P_{D\sim\mathcal{D}^m}(D \in A)$ 的下界即可证明定理.

令 l_m 表示从 S 中按分布 $\mathcal{D}$ 独立同分布采样 m 个样本落在 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d-1}\}$ 中的数目, 根据 Chernoff 不等式 (1.26) 可知, 对于 $\gamma > 1$, 有

$8\epsilon m$ 为样本落在 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d-1}\}$ 中的期望数目.

$$P_{D\sim\mathcal{D}^m}\left(l_m \geqslant 8\epsilon m\left(1+\gamma\right)\right) \leqslant e^{-8\epsilon m\frac{\gamma^2}{3}} . \tag{4.49}$$

令 $\epsilon = (d-1)/(32m), \gamma = 1$, 可得

$A = \{D \sim \mathcal{D}^m \big| (|D| = m) \wedge (|\bar{D}| \leqslant \frac{d-1}{2})\}$, $\bar{D}$ 表示出现在 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d-1}\}$ 中的样本集合.

$$\begin{aligned} & 1 - P_{D\sim\mathcal{D}^m}(D \in A) \\ =\ & P_{D\sim\mathcal{D}^m}\left(l_m \geqslant \frac{d-1}{2}\right) \\ \leqslant\ & e^{-(d-1)/12} \\ \leqslant\ & e^{-1/12} . \end{aligned} \tag{4.50}$$

令 $e^{-1/12} \leqslant 1 - 7\delta$, 可得 $P_{D\sim\mathcal{D}^m}(D \in A) \geqslant 7\delta$, 再根据

$$P_{D\sim\mathcal{D}^m}\left(E(h_D, c^*) > \epsilon\right) \geqslant \frac{1}{7} P_{D\sim\mathcal{D}^m}(D \in A) \tag{4.51}$$

可知

$$P_{D\sim\mathcal{D}^m}\left(E(h_D, c^*) > \epsilon\right) \geqslant \delta , \tag{4.52}$$

取 $\delta = \frac{1}{100}$, 从而定理得证. □

定理 4.6 表明对于任意学习算法 $\mathfrak{L}$, 必存在一种“坏”分布 $\mathcal{D}$ 以及一个目标概念 c^*, 使得 $\mathfrak{L}$ 输出的假设 h_D 总会以较高概率 (至少 1 %) 产生 $O(\frac{d}{m})$ 的错误. 需要注意的是, 定理 4.6 中数据分布 $\mathcal{D}$ 是与学习算法 $\mathfrak{L}$ 无关的, 只与假设空间 $\mathcal{H}$ 有关.

不可分情形

对于不可分假设空间的泛化误差下界, 主要比较学习算法 $\mathfrak{L}$ 的泛化误差与**贝叶斯最优分类器** (Bayes' classifier) 泛化误差之间的关系. 首先, 需要先给出两个引理 [Mohri et al., 2018].

分布 $\mathcal{D}$ 上取得最小泛化误差的分类器称为贝叶斯最优分类器, 参见 6.1 节.

引理 4.2 令 σ 为服从 $\{-1,+1\}$ 上均匀分布的随机变量, 对于 $0<\alpha<1$ 构造随机变量 $\alpha_\sigma=\frac{1}{2}+\frac{\alpha\sigma}{2}$, 基于 σ 构造 $X\sim\mathcal{D}_\sigma$, 其中 $\mathcal{D}_\sigma$ 为伯努利分布 Bernoulli(α_σ), 即 $P(X=1)=\alpha_\sigma$. 令 $S=\{X_1,\ldots,X_m\}$ 表示从分布 $\mathcal{D}_\sigma^m$ 独立同分布采样得到的大小为 m 的集合, 即 $S\sim\mathcal{D}_\sigma^m$, 则对于函数 $f: X^m\mapsto\{-1,+1\}$ 有

$$\mathbb{E}_\sigma\left[P_{S\sim D_\sigma^m}\left(f(S)\neq\sigma\right)\right]\geqslant\Phi\left(2\lceil m/2\rceil,\alpha\right)\ , \tag{4.53}$$

其中 $\Phi(m,\alpha)=\frac{1}{4}\left(1-\sqrt{1-\exp\left(-\frac{m\alpha^2}{1-\alpha^2}\right)}\right)$.

引理 4.2 的证明过程参阅 [Slud, 1977], 可以用抛硬币实验进行解释. 有两种硬币分别记为 A 和 B, 当 σ 取值为 -1 时选取 A 硬币; 当 σ 取值为 $+1$ 时选取 B 硬币, 硬币经过抛掷后正面向上的概率为 α_σ, 函数 f 要从 m 次抛掷硬币的结果 S 推断 σ 的取值.

根据引理 4.2 进一步推导可知, 为了确定 σ 的取值, m 至少应为 $\Omega(\frac{1}{\alpha^2})$. 此外, 还需要下面的引理以在推导过程中进行放缩.

引理 4.3 令 Z 为取值范围为 $[0,1]$ 的随机变量, 对于 $\gamma\in[0,1)$ 有

$$P(Z>\gamma)\geqslant\frac{\mathbb{E}[Z]-\gamma}{1-\gamma}\geqslant\mathbb{E}[Z]-\gamma\ . \tag{4.54}$$

证明 要点在于将 Z 的取值范围按照 γ 进行划分并分别进行放缩, 考虑随机变量 Z 的期望

$$\begin{aligned}\mathbb{E}[Z]&=\sum_{z\leqslant\gamma}P(Z=z)z+\sum_{z>\gamma}P(Z=z)z\\&\leqslant\sum_{z\leqslant\gamma}P(Z=z)\gamma+\sum_{z>\gamma}P(Z=z)\\&=\gamma P(Z\leqslant\gamma)+P(Z>\gamma)\\&=\gamma(1-P(Z>\gamma))+P(Z>\gamma)\\&=(1-\gamma)P(Z>\gamma)+\gamma\ .\end{aligned} \tag{4.55}$$

前一项将 z 放大至 γ, 后一项将 z 放大至 1.

整理化简可得

$$P(z > \gamma) \geqslant \frac{\mathbb{E}[Z] - \gamma}{1 - \gamma} \geqslant \mathbb{E}[Z] - \gamma \,. \tag{4.56}$$

□

基于引理 4.2 和引理 4.3 可以分析不可分情形下的泛化误差下界 [Anthony and Bartlett, 2009].

定理 4.7 若假设空间 $\mathcal{H}$ 的 VC 维 $d > 1$, 则对任意 $m > 1$ 和学习算法 $\mathfrak{L}$, 存在分布 $\mathcal{D}$ 使得

$$P_{Z \sim \mathcal{D}^m}\left(E(h_Z) - \inf_{h \in \mathcal{H}} E(h) > \sqrt{\frac{d}{320m}}\right) \geqslant \frac{1}{64} \,. \tag{4.57}$$

其中 h_Z 为学习算法 $\mathfrak{L}$ 基于大小为 m 的训练集 Z 输出的假设.

证明 令 $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\} \subset \mathcal{X}$ 表示能被 $\mathcal{H}$ 打散的集合. 对于 $\alpha \in [0, 1]$ 和向量 $\boldsymbol{\sigma} = (\sigma_1; \cdots; \sigma_d) \in \{-1, +1\}^d$, 在 $S \times \mathcal{Y}$ 上构造如下分布 $\mathcal{D}_{\boldsymbol{\sigma}}$

推断 $\boldsymbol{x}_i$ 的标记需要估计 σ_i, 基于引理 4.2 可知至少需要 $\Omega(\frac{1}{\alpha^2})$ 次采样才能准确估计 σ_i 的取值.

$$P_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{z} = (\boldsymbol{x}_i, +1)) = \frac{1}{d}\left(\frac{1}{2} + \frac{\sigma_i \alpha}{2}\right) \quad (i \in [d]) \,, \tag{4.58}$$

$$P_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{z} = (\boldsymbol{x}_i, -1)) = \frac{1}{d}\left(\frac{1}{2} - \frac{\sigma_i \alpha}{2}\right) \quad (i \in [d]) \,. \tag{4.59}$$

令 $\inf_{h \in \mathcal{H}} E(h)$ 表示假设空间 $\mathcal{H}$ 所能达到的最优误差. 不妨考虑一种极端情形, 将 $\mathcal{H}$ 放松到所有假设空间, 即达到贝叶斯最优分类器 $h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}$ 的泛化误差, 其中 $h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = \arg\max_{y \in \{-1,+1\}} P(y|\boldsymbol{x}_i) = \operatorname{sign}(\mathbb{I}(\sigma_i > 0) - 1/2)$, $i \in [d]$. 因为 S 能被 $\mathcal{H}$ 打散, 可知 $h^*_{\mathcal{D}_{\boldsymbol{\sigma}}} \in \mathcal{H}$.

对于 $h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}$ 计算可得

$$\begin{aligned}
E(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}) &= \sum_{\boldsymbol{x}_i \in S} \Big(P_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{z} = (\boldsymbol{x}_i, +1)) \mathbb{I}(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = -1) \\
&\quad + P_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{z} = (\boldsymbol{x}_i, -1)) \mathbb{I}(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = +1) \Big) \\
&= \sum_{\boldsymbol{x}_i \in S} \Big(P_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{z} = (\boldsymbol{x}_i, +1)) \mathbb{I}(\sigma_i < 0) + P_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{z} = (\boldsymbol{x}_i, -1)) \mathbb{I}(\sigma_i > 0) \Big) \\
&= \sum_{\boldsymbol{x}_i \in S} \frac{1}{d}\left(\frac{1}{2} - \frac{\alpha}{2}\right) = \frac{1}{2} - \frac{\alpha}{2} \,.
\end{aligned} \tag{4.60}$$

对于任意 $h \in \mathcal{H}$ 计算可得

$$\begin{aligned}
E(h) = \sum_{\boldsymbol{x}_i \in S} & \Big(P_{\mathcal{D}_{\boldsymbol{\sigma}}}\big(\boldsymbol{z} = (\boldsymbol{x}_i, +1)\big) \mathbb{I}(h(\boldsymbol{x}_i) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)) \mathbb{I}\big(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = +1\big) \\
& + P_{\mathcal{D}_{\boldsymbol{\sigma}}}\big(\boldsymbol{z} = (\boldsymbol{x}_i, +1)\big) \mathbb{I}(h(\boldsymbol{x}_i) = h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)) \mathbb{I}\big(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = -1\big) \\
& + P_{\mathcal{D}_{\boldsymbol{\sigma}}}\big(\boldsymbol{z} = (\boldsymbol{x}_i, -1)\big) \mathbb{I}(h(\boldsymbol{x}_i) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)) \mathbb{I}\big(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = -1\big) \\
& + P_{\mathcal{D}_{\boldsymbol{\sigma}}}\big(\boldsymbol{z} = (\boldsymbol{x}_i, -1)\big) \mathbb{I}(h(\boldsymbol{x}_i) = h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)) \mathbb{I}\big(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i) = +1\big) \\
= \sum_{\boldsymbol{x}_i \in S} & \Big(\frac{1+\alpha}{2d} \mathbb{I}\big(h(\boldsymbol{x}_i) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)\big) + \frac{1-\alpha}{2d} \mathbb{I}\big(h(\boldsymbol{x}_i) = h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)\big) \Big) \\
= & \frac{\alpha}{d} \sum_{\boldsymbol{x}_i \in S} \mathbb{I}\big(h(\boldsymbol{x}_i) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)\big) + \frac{1}{2} - \frac{\alpha}{2} \,. \qquad (4.61)
\end{aligned}$$

从而可知

$$E(h) - E(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}) = \frac{\alpha}{d} \sum_{\boldsymbol{x}_i \in S} \mathbb{I}\left(h(\boldsymbol{x}_i) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}_i)\right) \,. \qquad (4.62)$$

令 h_Z 表示算法 $\mathfrak{L}$ 基于从分布 $\mathcal{D}_{\boldsymbol{\sigma}}$ 独立同分布采样得到的 Z 而输出的假设，$|Z|_{\boldsymbol{x}}$ 表示样本 $\boldsymbol{x}$ 在 Z 中出现的次数，$\mathcal{U}$ 为 $\{-1,+1\}^d$ 上的均匀分布，基于 (4.62) 计算可得:

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{\sigma} \sim \mathcal{U}, Z \sim \mathcal{D}^m_{\boldsymbol{\sigma}}} \left[\frac{1}{\alpha} \left(E(h_Z) - E(h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}) \right) \right] \\
& = \frac{1}{d} \sum_{\boldsymbol{x} \in S} \mathbb{E}_{\boldsymbol{\sigma} \sim \mathcal{U}, Z \sim \mathcal{D}^m_{\boldsymbol{\sigma}}} \left[\mathbb{I}\left(h_Z(\boldsymbol{x}) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}) \right) \right] \\
& = \frac{1}{d} \sum_{\boldsymbol{x} \in S} \mathbb{E}_{\boldsymbol{\sigma} \sim \mathcal{U}} \left[P_{Z \sim \mathcal{D}^m_{\boldsymbol{\sigma}}} \left(h_Z(\boldsymbol{x}) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}) \right) \right] \\
& = \frac{1}{d} \sum_{\boldsymbol{x} \in S} \sum_{n=0}^{m} \mathbb{E}_{\boldsymbol{\sigma} \sim \mathcal{U}} \left[P_{Z \sim \mathcal{D}^m_{\boldsymbol{\sigma}}} \left(h_Z(\boldsymbol{x}) \neq h^*_{\mathcal{D}_{\boldsymbol{\sigma}}}(\boldsymbol{x}) \middle| |Z|_{\boldsymbol{x}} = n \right) P(|Z|_{\boldsymbol{x}} = n) \right] \qquad \text{按 } |Z|_{\boldsymbol{x}} \text{ 取值进行展开.} \\
& \geqslant \frac{1}{d} \sum_{\boldsymbol{x} \in S} \sum_{n=0}^{m} \Phi(2\lceil n/2 \rceil, \alpha) P(|Z|_{\boldsymbol{x}} = n) \qquad \text{基于引理 4.2.} \\
& \geqslant \frac{1}{d} \sum_{\boldsymbol{x} \in S} \sum_{n=0}^{m} \Phi(n+1, \alpha) P(|Z|_{\boldsymbol{x}} = n) \qquad \text{基于 } \Phi(\cdot,\alpha) \text{ 是凸函数和 Jensen 不等式 (1.11)，} \sum_{n=0}^{m} \Phi(n+1,\alpha) P(|Z|_{\boldsymbol{x}} = n) \geqslant \Phi(m/d+1,\alpha), \text{ } m/d+1 \text{ 是期望.} \\
& \geqslant \frac{1}{d} \sum_{\boldsymbol{x} \in S} \Phi(m/d+1, \alpha) = \Phi(m/d+1, \alpha) \,. \qquad (4.63)
\end{aligned}$$

由于上述关于 $\boldsymbol{\sigma}$ 期望的下界被 $\Phi(m/d+1,\alpha)$ 限制住, 则必定存在 $\boldsymbol{\sigma}^* \in \{-1,+1\}^d$ 使得下式成立

$$\mathbb{E}_{Z\sim\mathcal{D}_{\boldsymbol{\sigma}^*}^m}\left[\frac{1}{\alpha}\left(E(h_Z)-E(h^*_{\mathcal{D}_{\boldsymbol{\sigma}^*}^m})\right)\right]\geqslant\Phi(m/d+1,\alpha)\,. \tag{4.64}$$

根据引理 4.3 可知, 对于 $\boldsymbol{\sigma}^*$以及任意 $\gamma\in[0,1)$ 有

$$P_{Z\sim\mathcal{D}_{\boldsymbol{\sigma}^*}^m}\left(\frac{1}{\alpha}\left(E(h_Z)-E(h^*_{\mathcal{D}_{\boldsymbol{\sigma}^*}^m})\right)>\gamma u\right)\geqslant(1-\gamma)u\,. \tag{4.65}$$

其中 $u=\Phi(m/d+1,\alpha)$. 令 δ 与 ϵ 满足条件 $\delta\leqslant(1-\gamma)u$ 以及 $\epsilon\leqslant\gamma\alpha u$, 则有

$$P_{Z\sim\mathcal{D}_{\boldsymbol{\sigma}^*}^m}\left(E\left(h_Z\right)-E\left(h^*_{\mathcal{D}_{\boldsymbol{\sigma}^*}^m}\right)>\epsilon\right)\geqslant\delta\,. \tag{4.66}$$

为了找到满足条件的 δ 与 ϵ, 令 $\gamma=1-8\delta$, 则

$\Longleftrightarrow$ 表示“等价于”.

$$\begin{aligned}
\delta\leqslant(1-\gamma)u &\Longleftrightarrow u\geqslant\frac{1}{8}\\
&\Longleftrightarrow\frac{1}{4}\left(1-\sqrt{1-\exp\left(-\frac{(m/d+1)\alpha^2}{1-\alpha^2}\right)}\right)\geqslant\frac{1}{8}\\
&\Longleftrightarrow\frac{(m/d+1)\alpha^2}{1-\alpha^2}\leqslant\ln\frac{4}{3}\\
&\Longleftrightarrow\frac{m}{d}\leqslant\left(\frac{1}{\alpha^2}-1\right)\ln\frac{4}{3}-1\,.
\end{aligned} \tag{4.67}$$

令 $\alpha=8\epsilon/(1-8\delta)$, 即 $\epsilon=\gamma\alpha/8$, 可将 (4.67) 转换为

$$\frac{m}{d}\leqslant\left(\frac{(1-8\delta)^2}{64\epsilon^2}-1\right)\ln\frac{4}{3}-1\,. \tag{4.68}$$

令 $\delta\leqslant1/64$, 可得

$$\left(\frac{(1-8\delta)^2}{64\epsilon^2}-1\right)\ln\frac{4}{3}-1\geqslant\left(\frac{7}{64}\right)^2\frac{1}{\epsilon^2}\ln\frac{4}{3}-\ln\frac{4}{3}-1\,. \tag{4.69}$$

(4.69) 右端为关于 $\frac{1}{\epsilon^2}$ 的函数 $f(\frac{1}{\epsilon^2})$, 可寻找 w 使得 $m/d\leqslant w/\epsilon^2$. 令 $\epsilon\leqslant1/64$, 由 $\frac{w}{(1/64)^2}=f\left(\frac{1}{(1/64)^2}\right)$ 可得

$$w=(7/64)^2\ln(4/3)-(1/64)^2(\ln(4/3)+1)\approx0.003127\geqslant1/320. \tag{4.70}$$

因此, 当 $\epsilon^2 \leqslant \frac{1}{320m/d}$ 时, 满足 $\delta \leqslant (1-\gamma)u$ 以及 $\epsilon \leqslant \gamma\alpha u$. 取 $\epsilon = \sqrt{\frac{d}{320m}}$ 和 $\delta = 1/64$, 定理得证. □

定理 4.7 表明对于任意学习算法 $\mathfrak{L}$, 在不可分情形下必存在一种 "坏" 分布 $\mathcal{D}_{\boldsymbol{\sigma}^\star}$, 使得 $\mathfrak{L}$ 输出的假设 h_Z 的泛化误差以常数概率为 $O\left(\sqrt{\frac{d}{m}}\right)$.

4.3 分析实例

本节将分析支持向量机的泛化误差界.

支持向量机考虑的假设空间是线性超平面, 定理 3.6 证明了 $\mathbb{R}^d$ 中线性超平面的 VC 维为 $d+1$, 再结合定理 4.3 可以得到支持向量机基于 VC 维的泛化误差界: 对于 $0 < \delta < 1$, 以至少 $1-\delta$ 的概率有

$$E(h) \leqslant \widehat{E}(h) + \sqrt{\frac{8(d+1)\ln\frac{2em}{d+1} + 8\ln\frac{4}{\delta}}{m}} \,. \tag{4.71}$$

当样本空间的维数相对于样本的数目很大时, (4.71) 没有给出具有实际意义的信息. 另外, 定理 3.8 给出了一种不依赖样本空间维数的 VC 维的估计方法, 但是需要限制样本的范数, 使得 $\|\boldsymbol{x}\| \leqslant r$. 因此当样本空间有界时, 基于定理 3.8 和定理 4.3 也可以得到支持向量机基于 VC 维的泛化误差界.

更多关于替代损失函数的内容参见 6.2 节.

然而在实际应用中, 支持向量机通常会使用替代损失函数, 例如 (1.72) 中提到的 hinge 损失函数. 下面我们就来讨论使用替代损失函数的支持向量机的泛化误差界. 考虑比 hinge 损失函数更具一般性的间隔损失函数:

定义 4.1 对于任意 $\rho > 0$, **ρ-间隔损失** 为定义在 $z, z' \in \mathbb{R}$ 上的损失函数 $\ell_\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$, $\ell_\rho(z, z') = \Phi_\rho(zz')$, 其中

$$\Phi_\rho(x) = \begin{cases} 0 & \rho \leqslant x \\ 1 - x/\rho & 0 \leqslant x \leqslant \rho \\ 1 & x \leqslant 0 \end{cases} \,. \tag{4.72}$$

对于集合 $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ 与假设 h, 经验间隔损失表示为

$$\widehat{E}_\rho(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho\left(y_i h(\boldsymbol{x}_i)\right) \,. \tag{4.73}$$

考虑到 $\Phi_\rho(y_i h(\boldsymbol{x}_i)) \leqslant \mathbb{I}_{y_i h(\boldsymbol{x}_i) \leqslant \rho}$, 对于经验间隔损失, 有

$$\widehat{E}_\rho(h) \leqslant \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}_{y_i h(\boldsymbol{x}_i) \leqslant \rho} \,. \tag{4.74}$$

Φ_ρ 的导数最大为 $\frac{1}{\rho}$.

由经验间隔损失 (4.72) 可知 Φ_ρ 最多是 $\frac{1}{\rho}$-Lipschitz. 引理 4.4 表明 Lipschitz 函数和假设空间 $\mathcal{H}$ 复合后的经验 Rademacher 复杂度可以基于假设空间 $\mathcal{H}$ 的经验 Rademacher 复杂度进行表示.

证明过程参阅 [Ledoux and Talagrand, 1991].

引理 4.4 若 $\Phi : \mathbb{R} \mapsto \mathbb{R}$ 为 l-Lipschitz 函数, 则对于任意实值假设空间 $\mathcal{H}$ 有下式成立:

$$\widehat{\mathfrak{R}}_D(\Phi \circ \mathcal{H}) \leqslant l \widehat{\mathfrak{R}}_D(\mathcal{H}) \,. \tag{4.75}$$

函数的复合 $f_2 \circ f_1(\boldsymbol{x}) = f_2(f_1(\boldsymbol{x}))$.

下面将给出基于间隔损失函数的二分类问题支持向量机的泛化误差界.

定理 4.8 令 $\mathcal{H}$ 为实值假设空间, 给定 $\rho > 0$, 对于 $0 < \delta < 1$ 和 $h \in \mathcal{H}$, 以至少 $1 - \delta$ 的概率有

$$E(h) \leqslant \widehat{E}_\rho(h) + \frac{2}{\rho} \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \,, \tag{4.76}$$

$$E(h) \leqslant \widehat{E}_\rho(h) + \frac{2}{\rho} \widehat{\mathfrak{R}}_D(\mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \,. \tag{4.77}$$

证明 构造 $\widetilde{\mathcal{H}} = \{\boldsymbol{z} = (\boldsymbol{x}, y) \mapsto y h(\boldsymbol{x}) : h \in \mathcal{H}\}$, 考虑值域为 $[0, 1]$ 的假设空间 $\mathcal{F} = \{\Phi_\rho \circ f : f \in \widetilde{\mathcal{H}}\}$, 根据 (4.25) 可知对于所有 $g \in \mathcal{F}$, 以至少 $1 - \delta$ 的概率有

$$\mathbb{E}[g(\boldsymbol{z})] \leqslant \frac{1}{m} \sum_{i=1}^{m} g(\boldsymbol{z}_i) + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \,. \tag{4.78}$$

因此, 对 $h \in \mathcal{H}$, 以至少 $1 - \delta$ 的概率有

$$\mathbb{E}[\Phi_\rho(y h(\boldsymbol{x}))] \leqslant \widehat{E}_\rho(h) + 2\mathfrak{R}_m(\Phi_\rho \circ \widetilde{\mathcal{H}}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \,. \tag{4.79}$$

因为 $\mathbb{I}_{u \leqslant 0} \leqslant \Phi_\rho(u)$ 对任意 $u \in \mathbb{R}$ 成立, 所以 $E(h) = \mathbb{E}[\mathbb{I}_{y h(\boldsymbol{x}) \leqslant 0}] \leqslant \mathbb{E}[\Phi_\rho(y h(\boldsymbol{x}))]$, 代入 (4.96) 可知

$$E(h) \leqslant \widehat{E}_\rho(h) + 2\mathfrak{R}_m(\Phi_\rho \circ \widetilde{\mathcal{H}}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{4.80}$$

以至少 $1-\delta$ 的概率成立. 由于 Φ_ρ 是 $\frac{1}{\rho}$-Lipschitz, 根据引理 4.4 可知

$$\mathfrak{R}_m(\Phi_\rho \circ \widetilde{\mathcal{H}}) \leqslant \frac{1}{\rho}\mathfrak{R}_m(\widetilde{\mathcal{H}}) \ . \tag{4.81}$$

Rademacher 复杂度 $\mathfrak{R}_m(\widetilde{\mathcal{H}})$ 参见定义 3.4.

考虑到 $\mathfrak{R}_m(\widetilde{\mathcal{H}})$ 可以重写为

$$\begin{aligned}\mathfrak{R}_m(\widetilde{\mathcal{H}}) &= \frac{1}{m}\mathbb{E}_{D,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i y_i h(\boldsymbol{x}_i)\right] \\ &= \frac{1}{m}\mathbb{E}_{D,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)\right] \\ &= \mathfrak{R}_m(\mathcal{H}) \ ,\end{aligned} \tag{4.82}$$

基于 (4.81) 可得

$$\mathfrak{R}_m(\Phi_\rho \circ \widetilde{\mathcal{H}}) \leqslant \frac{1}{\rho}\mathfrak{R}_m(\mathcal{H}) \ . \tag{4.83}$$

将其代入 (4.80), 可知

$$E(h) \leqslant \widehat{E}_\rho(h) + \frac{2}{\rho}\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \tag{4.84}$$

以至少 $1-\delta$ 的概率成立, 从而 (4.76) 得证.

基于 (4.26) 可知

$$\mathbb{E}[g(\boldsymbol{z})] \leqslant \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{z}_i) + 2\widehat{\mathfrak{R}}_D(\mathcal{F}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}} \tag{4.85}$$

以至少 $1-\delta$ 的概率成立. 通过类似于 (4.78)$\sim$(4.80) 的推导可知

$$E(h) \leqslant \widehat{E}_\rho(h) + 2\widehat{\mathfrak{R}}_D(\Phi_\rho \circ \widetilde{\mathcal{H}}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}} \tag{4.86}$$

以至少 $1-\delta$ 的概率成立.

对于经验 Rademacher 复杂度, 由于 Φ_ρ 是 $\frac{1}{\rho}$-Lipschitz, 类似 (4.81) 可得

$$\widehat{\mathfrak{R}}_D(\Phi_\rho \circ \widetilde{\mathcal{H}}) \leqslant \frac{1}{\rho}\widehat{\mathfrak{R}}_D(\widetilde{\mathcal{H}}) \ . \tag{4.87}$$

考虑到 $\widehat{\mathfrak{R}}_D(\widetilde{\mathcal{H}})$ 可以重写为

$$\begin{aligned}\widehat{\mathfrak{R}}_D(\widetilde{\mathcal{H}}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i y_i h(\boldsymbol{x}_i)\right] \\ &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)\right] \\ &= \widehat{\mathfrak{R}}_m(\mathcal{H})\,, \end{aligned} \tag{4.88}$$

结合 (4.86)~(4.88) 可知

$$E(h) \leqslant \widehat{E}_\rho(h) + \frac{2}{\rho}\widehat{\mathfrak{R}}_D(\mathcal{H}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}} \tag{4.89}$$

以至少 $1-\delta$ 的概率成立, 从而 (4.77) 得证. □

定理 4.8 中要求 ρ 是事先给定的, 下面给出的定理则可以对任意 $\rho\in(0,1)$ 均成立.

证明过程参阅 [Mohri et al., 2018].

定理 4.9 令 $\mathcal{H}$ 为实值假设空间, 对于 $0<\delta<1$, $h\in\mathcal{H}$ 以及任意 $\rho\in(0,1)$, 以至少 $1-\delta$ 的概率有

$$E(h) \leqslant \widehat{E}_\rho(h) + \frac{4}{\rho}\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\log_2\frac{2}{\rho}}{m}} + \sqrt{\frac{\ln\frac{2}{\delta}}{2m}}\,, \tag{4.90}$$

$$E(h) \leqslant \widehat{E}_\rho(h) + \frac{4}{\rho}\widehat{\mathfrak{R}}_D(\mathcal{H}) + \sqrt{\frac{\ln\log_2\frac{2}{\rho}}{m}} + 3\sqrt{\frac{\ln\frac{4}{\delta}}{2m}}\,. \tag{4.91}$$

由定理 3.7 可知 $\widehat{\mathfrak{R}}_D(\mathcal{H}) \leqslant \sqrt{\frac{r^2\Lambda^2}{m}}$, 对其两边取期望可得 $\mathfrak{R}_m(\mathcal{H}) \leqslant \sqrt{\frac{r^2\Lambda^2}{m}}$, 进一步结合定理 4.8 和定理 4.9 可得下面的两个推论.

推论 4.1 令 $\mathcal{H}=\{\boldsymbol{x}\mapsto\boldsymbol{w}\cdot\boldsymbol{x}:\|\boldsymbol{w}\|\leqslant\Lambda\}$ 且 $\|\boldsymbol{x}\|\leqslant r$, 对于 $0<\delta<1$, $h\in\mathcal{H}$ 和固定的 $\rho>0$, 以至少 $1-\delta$ 的概率有

$$E(h) \leqslant \widehat{E}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}\,. \tag{4.92}$$

推论 4.2 令 $\mathcal{H}=\{\boldsymbol{x}\mapsto\boldsymbol{w}\cdot\boldsymbol{x}:\|\boldsymbol{w}\|\leqslant\Lambda\}$ 且 $\|\boldsymbol{x}\|\leqslant r$, 对于 $0<\delta<1$,

$h \in \mathcal{H}$ 和任意 $\rho \in (0,1)$, 以至少 $1-\delta$ 的概率有

$$E(h) \leqslant \widehat{E}_\rho(h) + 4\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\ln \log_2 \frac{2}{\rho}}{m}} + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}} . \tag{4.93}$$

由本章的内容可以发现, 泛化界主要讨论的是学习算法 $\mathfrak{L}$ 输出假设 h 的泛化误差与经验误差之间的关系, 而第 2 章介绍的 PAC 学习理论要求的是找到假设空间中具有最小泛化误差假设的 ϵ 近似. 若要实现这一目标, 则需要引入 **经验风险最小化** (Empirical Risk Minimization) :

更多关于经验风险最小化的内容参见 5.2.2 节.

如果学习算法 $\mathfrak{L}$ 输出 $\mathcal{H}$ 中具有最小经验误差的假设 h, 即 $\widehat{E}(h) = \min_{h' \in \mathcal{H}} \widehat{E}(h')$, 则称 $\mathfrak{L}$ 为满足经验风险最小化原则的算法.

接下来我们讨论是否能够基于经验风险最小化原则找到假设空间中具有最小泛化误差假设的 ϵ 近似. 假设 $\mathfrak{L}$ 为满足经验风险最小化原则的算法, 令 g 表示 $\mathcal{H}$ 中具有最小泛化误差的假设, 即 $E(g) = \min_{h \in \mathcal{H}} E(h)$, 对于 $0 < \epsilon, \delta < 1$, 由引理 2.1 可知

$$P\left(\left|\widehat{E}(g) - E(g)\right| \geqslant \frac{\epsilon}{2}\right) \leqslant 2\exp\left(-\frac{m\epsilon^2}{2}\right) . \tag{4.94}$$

令 $\delta' = \frac{\delta}{2}$, $\sqrt{\frac{(\ln 2/\delta')}{2m}} \leqslant \frac{\epsilon}{2}$, 由 (4.94) 可知

$$\widehat{E}(g) - \frac{\epsilon}{2} \leqslant E(g) \leqslant \widehat{E}(g) + \frac{\epsilon}{2} \tag{4.95}$$

以至少$1-\delta/2$的概率成立. 令

$$\sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta'}}{m}} \leqslant \frac{\epsilon}{2} , \tag{4.96}$$

这里以基于定理 4.3 的分析作为例子, 其他关于泛化误差的定理 (如定理 4.2、定理 4.5 等) 也可以有类似分析.

由定理 4.3 可知

$$\left|E(h) - \widehat{E}(h)\right| \leqslant \frac{\epsilon}{2} \tag{4.97}$$

以至少$1-\delta/2$的概率成立. 由 (4.95)、(4.97) 和联合界不等式 (1.19) 可知

$$\begin{aligned} E(h) - E(g) &\leqslant \widehat{E}(h) + \frac{\epsilon}{2} - \left(\widehat{E}(g) - \frac{\epsilon}{2}\right) \\ &= \widehat{E}(h) - \widehat{E}(g) + \epsilon \\ &\leqslant \epsilon \end{aligned} \tag{4.98}$$

根据经验风险最小化原则, $\widehat{E}(h) \leqslant \widehat{E}(g)$.

以至少 $1-\delta$ 的概率成立. 因此, 若学习算法 $\mathfrak{L}$ 输出 $\mathcal{H}$ 中具有最小经验误差的假设 h, 其泛化误差 $E(h)$ 以至少 $1-\delta$ 的概率不大于最小泛化误差 $E(g)+\epsilon$.

ϵ 与 m 有关, 具体取值可由 $\sqrt{\frac{(\ln 2/\delta')}{2m}} \leqslant \frac{\epsilon}{2}$ 和 (4.96) 求解, 有兴趣的读者可以作为习题练习.

4.4 阅读材料

本章主要讨论了二分类问题的泛化误差界, 对于多分类问题 [Natarajan, 1989]、回归问题 [Cherkassky et al., 1999] 等也有相应的泛化误差界分析.

基于 VC 维的泛化误差界是分布无关且数据独立的, 仅与 VC 维和训练集大小有关, 而基于 Rademacher 复杂度的泛化误差界与分布 $\mathcal{D}$ 有关 ($\mathfrak{R}_m(\mathcal{H})$ 项) 或与数据 D 有关 ($\widehat{\mathfrak{R}}_D(\mathcal{H})$ 项). 换言之, 基于 Rademacher 复杂度的泛化误差界依赖于具体学习问题上的数据及其分布, 是为具体的学习问题量身定制, 因此通常比基于 VC 维的泛化误差界更紧一些. 更多关于二者之间区别的讨论可参阅 [Bartlett and Mendelson, 2002].

本章的泛化误差界分析是基于第 2 章的 PAC 学习理论, McAllester [1999] 等从贝叶斯 (Bayes) 的角度提出了 PAC-Bayesian 理论, 后续相继出现了 PAC-Bayes 泛化误差界分析, 例如 Seeger [2002] 给出了高斯过程的 PAC-Bayes 泛化误差界, Morvant et al. [2012] 给出了多分类问题的 PAC-Bayes 泛化误差界等. 一直以来, Rademacher 复杂度与 PAC-Bayes 理论并行发展, 近来 Yang et al. [2019] 尝试建立 Rademacher 复杂度和 PAC-Bayes 理论之间的联系. 目前, 大部分泛化误差界分析都基于样本独立同分布这一重要假设, 一些考虑放松独立同分布假设的泛化误差分析工作也逐步出现 [Kuznetsov and Mohri, 2017].

本章在分析支持向量机的泛化误差界时用到了 **间隔理论**, 引入间隔后使得泛化误差界与数据分布相关. 间隔理论还在 AdaBoost 理论分析上发挥了重要作用 [周志华, 2020]. 许多学习算法已有泛化误差分析结果, 如决策树 [Mansour and Mcallester, 2000]、对率回归 [Krishnapuram et al., 2005] 等.

由于深度学习的兴起, 关于深度神经网络的泛化误差界分析也引起了关注. Zhang et al. [2017] 通过实验结果指出传统基于 VC 维和 Rademacher 复杂度的泛化误差分析无法解释深度神经网络中参数数目远远大于样本数目却仍具有良好泛化性这一现象. Arpit et al. [2017] 进一步指出分析深度神经网络的泛化误差界不能简单考虑深度神经网络理论上所能表达的假设空间复杂度, 而需结合训练深度神经网络采用的优化算法及训练数据, 考虑分析深度神经网络所能优化假设构成的假设空间的复杂度. 深度神经网络的泛化误差分析目前仅有一些初步探索工作 [Arora et al., 2018].

习题

4.1 试给出轴平行矩形假设空间基于 VC 维的泛化误差界, 并与 (2.23) 进行比较.

4.2 若假设空间 $\mathcal{H}$ 的 VC 维为 d,

(1) 试证明对任一大小为 m 的集合 D, $\widehat{\mathfrak{R}}_D(\mathcal{H}) \leqslant \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$; 进一步, 对任一分布 $\mathcal{D}$, $\mathfrak{R}_m(\mathcal{H}) \leqslant \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$.

(2) 试利用基于 Rademacher 复杂度的泛化误差界 (定理 4.5) 和 (1) 中的结果推导基于 VC 维的泛化误差界, 并与定理 4.3 进行比较.

4.3 若假设空间 $\mathcal{H}$ 满足 $|\mathcal{H}| \geqslant 3$, 试证明对于任意学习算法 $\mathfrak{L}$ 存在分布 $\mathcal{D}$ 和目标概念 $c \in \mathcal{H}$, 使得至少需要 $\Omega(\frac{1}{\epsilon}\ln\frac{1}{\delta})$ 个样本才有

$$P\left(E_{\mathcal{D}}(h_D, c) \leqslant \epsilon\right) \geqslant 1-\delta \ ,$$

其中 h_D 为 $\mathfrak{L}$ 基于从 $\mathcal{D}$ 独立同分布采样得到的训练集 D 输出的假设.

4.4 4.3 节分析实例中给出了二分类问题中支持向量机的泛化误差界. 对于多分类问题, 可以定义打分函数 $h(\boldsymbol{x}, y) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ 实现分类结果 $\arg\max_{y\in\mathcal{Y}} h(\boldsymbol{x}, y)$, 其中 $\mathcal{Y} = \{0, \ldots, K-1\}$.

(1) 定义打分函数 h 在点 $(\boldsymbol{x}, y)$ 处的间隔为 $\tau_h(\boldsymbol{x}, y) = h(\boldsymbol{x}, y) - \max_{y'\neq y} h(\boldsymbol{x}, y')$, h 在 $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ 上的经验间隔损失为 $\widehat{E}_{D,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho\left(\tau_h(\boldsymbol{x}_i, y_i)\right)$, 试证明:

$\Phi_\rho(\cdot)$ 参见 (4.72).

$$\widehat{E}_{D,\rho}(h) \leqslant \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(\tau_h(\boldsymbol{x}_i, y_i) \leqslant \rho\right) \ .$$

(2) 定义 $\tau_{\theta,h}(\boldsymbol{x}, y) = \min_{y'\in\mathcal{Y}}\left(h(\boldsymbol{x}, y) - h(\boldsymbol{x}, y') + \theta\mathbb{I}(y' \neq y)\right)$, 其中 $\theta > 0$ 为任意常数, $(\boldsymbol{x}, y) \sim \mathcal{D}$, 试证明:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{I}\left(\tau_h(\boldsymbol{x}, y) \leqslant 0\right)\right] \leqslant \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}\left(\tau_{\theta,h}(\boldsymbol{x}, y) \leqslant 0\right)\right] \ .$$

(3) 令 $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ 表示函数 h 构成的集合, 固定 $\rho > 0$, 考虑假设空间 $\widetilde{\mathcal{H}} = \{(\boldsymbol{x}, y) \mapsto \tau_{\theta,h}(\boldsymbol{x}, y) : h \in \mathcal{H}\}$, 其中 $\theta = 2\rho$, h 的泛化误差表示为 $E(h) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}\left(\tau_h(\boldsymbol{x}, y) \leqslant 0\right)\right]$, 试证明对 $0 < \delta < 1$,

$h \in \mathcal{H}$, 以至少 $1-\delta$ 的概率有

$$E(h) \leqslant \widehat{E}_{D,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_m(\widetilde{\mathcal{H}}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} .$$

参考文献

周志华. (2020). "Boosting 学习理论的探索." 中国计算机学会通讯, 16(4): 36–42.

Anthony, M. and P. L. Bartlett. (2009). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK.

Arora, S., R. Ge, B. Neyshabur, and Y. Zhang. (2018), "Stronger generalization bounds for deep nets via a compression approach." In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 254–263, Stockholm, Sweden.

Arpit, D., S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, et al. (2017), "A closer look at memorization in deep networks." In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 233–242, Sydney, Australia.

Bartlett, P. L. and S. Mendelson. (2002). "Rademacher and Gaussian complexities: Risk bounds and structural results." *Journal of Machine Learning Research*, 3:463–482.

Cherkassky, V., X. Shao, F. M. Mulier, and V. N. Vapnik. (1999). "Model complexity control for regression using VC generalization bounds." *IEEE Transactions on Neural Networks*, 10(5):1075–1089.

Ehrenfeucht, A., D. Haussler, M. J. Kearns, and L. G. Valiant. (1988), "A general lower bound on the number of examples needed for learning." In *Proceedings of the 1st Annual Conference on Computational Learning Theory (COLT)*, pp. 139–154, Cambridge, MA.

Krishnapuram, B., L. Carin, et al. (2005). "Sparse multinomial logistic regression: Fast algorithms and generalization bounds." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968.

Kuznetsov, V. and M. Mohri. (2017). "Generalization bounds for non-stationary mixing processes." *Machine Learning*, 106(1):93–117.

Ledoux, M. and M. Talagrand. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, Germany.

Mansour, Y. and D. A. Mcallester. (2000), "Generalization bounds for decision trees." In *Proceedings of the 13th Annual Conference on Computational*

Learning Theory (COLT), pp. 69–74, Palo Alto, CA.

McAllester, D. A. (1999). "Some PAC-Bayesian theorems." *Machine Learning*, 37(3):355–363.

Mohri, M., A. Rostamizadeh, and A. Talwalkar. (2018). *Foundations of Machine Learning*, 2nd edition. MIT Press, Cambridge, MA.

Morvant, E. et al. (2012), "PAC-Bayesian generalization bound on confusion matrix for multi-class classification." In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1211–1218, Edinburgh, UK.

Natarajan, B. K. (1989). "On learning sets and functions." *Machine Learning*, 4(1):67–97.

Seeger, M. W. (2002). "PAC-Bayesian generalisation error bounds for Gaussian process classification." *Journal of Machine Learning Research*, 3:233–269.

Slud, E. V. (1977). "Distribution inequalities for the binomial law." *Annals of Probability*, 5(3):404–412.

Stein, E. M. and R. Shakarchi. (2009). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces.* Princeton University Press, Princeton, NJ.

Vapnik, V. N. and A. Chervonenkis. (1971). "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability and Its Applications*, 16(2):264–280.

Yang, J., S. Sun, and D. M. Roy. (2019), "Fast-rate PAC-Bayes generalization bounds via shifted rademacher processes." In *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 10802–10812, Curran Associates Inc., Red Hook, NY.

Zhang, C., S. Bengio, M. Hardt, B. Recht, et al. (2017), "Understanding deep learning requires rethinking generalization." In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France.