学习理论初步

张腾

2025 年 9 月 7 日

一些符号说明：

1. 输入空间 $\mathcal{X} \subseteq \mathbb{R}^n$，类别标记集合 $\mathcal{Y} = \{1, -1\}$
2. 未知分布 P 定义在 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 上
3. 数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i \in [m]} \in \mathcal{Z}^m$，其中对 $\forall i : (x_i, y_i) \overset{\text{IID}}{\sim} P$
4. 设学习算法 A 考虑的假设空间为 $\mathcal{H}$，则 $A : \cup_{m=1}^{\infty} \mathcal{Z}^m \mapsto \mathcal{H}$ 是任意数据集到 $\mathcal{H}$ 的映射
5. 对任意假设 h，其泛化错误率和在数据集 $\mathcal{D}$ 上的经验错误率为
$$\operatorname{er}(h) = P\{(x, y) \in \mathcal{Z} \mid h(x) \neq y\} = \mathbb{E}_{(x,y)\sim P}[\mathbb{I}(h(x) \neq y)], \quad \operatorname{er}_{\mathcal{D}}(h) = \frac{1}{m} \sum_{i \in [m]} \mathbb{I}(h(x_i) \neq y_i)$$
6. $\mathcal{H}$ 中泛化错误率最小的假设 $h^\star = \operatorname{argmin}_{h \in \mathcal{H}} \operatorname{er}(h)$

1 学习的形式化定义

定义 1 (可学习). 设 $\mathcal{H}$ 是 $\mathcal{X} \mapsto \{1, -1\}$ 的函数类，若对任意给定的

1. 准确率参数 $\epsilon \in (0, 1)$
2. 置信度参数 $\delta \in (0, 1)$

存在 $m_0(\epsilon, \delta)$ 使得对任意 $m \geq m_0(\epsilon, \delta)$ 和任意分布 P，数据集 $\mathcal{D} \sim P^m$，学习算法 A 以至少 $1 - \delta$ 的概率输出一个 ϵ-好的假设，即
$$P^m\{\mathcal{D} \in \mathcal{Z}^m \mid \operatorname{er}(A(\mathcal{D})) < \operatorname{er}(h^\star) + \epsilon\} \geq 1 - \delta$$
则称 $\mathcal{H}$ 对于 A 是可学习的。

注. 分布 P^m 定义在 $\mathcal{Z}^m$ 上，其输入是 $\mathcal{Z}^m$ 的子集 (包含 m 个样本的数据集的集合)，输出是其测度。为保持符号系统的简洁，后文省略 $\mathcal{D} \in \mathcal{Z}^m$。

由于泛化错误率依赖未知分布 P，不可计算，考虑经验错误率最小化 (empirical risk minimization, ERM) 算法，其输出 $h_{\mathcal{D}}^{\text{ERM}} = \operatorname{argmin}_{h \in \mathcal{H}} \operatorname{er}_{\mathcal{D}}(h)$，于是
$$\begin{aligned}
\operatorname{er}(h_{\mathcal{D}}^{\text{ERM}}) - \operatorname{er}(h^\star) &= \operatorname{er}(h_{\mathcal{D}}^{\text{ERM}}) - \operatorname{er}_{\mathcal{D}}(h_{\mathcal{D}}^{\text{ERM}}) + \operatorname{er}_{\mathcal{D}}(h_{\mathcal{D}}^{\text{ERM}}) - \operatorname{er}(h^\star) \\
&\leq \operatorname{er}(h_{\mathcal{D}}^{\text{ERM}}) - \operatorname{er}_{\mathcal{D}}(h_{\mathcal{D}}^{\text{ERM}}) + \operatorname{er}_{\mathcal{D}}(h^\star) - \operatorname{er}(h^\star) \\
&\leq |\operatorname{er}_{\mathcal{D}}(h_{\mathcal{D}}^{\text{ERM}}) - \operatorname{er}(h_{\mathcal{D}}^{\text{ERM}})| + |\operatorname{er}_{\mathcal{D}}(h^\star) - \operatorname{er}(h^\star)|
\end{aligned} \tag{1}$$

因此我们需要一个刻画 |经验错误率 − 泛化错误率| 上界的工具。注意给定 h，经验错误率

$$\mathrm{er}_{\mathcal{D}}(h) = \frac{1}{m}\sum_{i\in[m]} \mathbb{I}(h(x_i) \neq y_i) \triangleq \frac{1}{m}\sum_{i\in[m]} X_i$$

其中 $X_i = \mathbb{I}(h(x_i) \neq y_i) \overset{\text{IID}}{\sim} \mathrm{Bern}(\mathrm{er}(h))$，因此其期望就等于 $\mathrm{er}(h)$，故刻画 rv 偏离其均值程度的集中不等式可以为我们所用：

1. 注意 $m \cdot \mathrm{er}_{\mathcal{D}}(h) = \sum_{i\in[m]} X_i \sim \mathrm{Bin}(\mathrm{er}(h), m)$，因此根据 Chebyshev's 不等式有

$$\begin{aligned} P^m\{|\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} &= P^m\{|m\cdot \mathrm{er}_{\mathcal{D}}(h) - m\cdot \mathrm{er}(h)|^2 \geq m^2\epsilon^2\} \\ &\leq \frac{m\cdot \mathrm{er}(h)(1-\mathrm{er}(h))}{m^2\epsilon^2} = \frac{\mathrm{er}(h)(1-\mathrm{er}(h))}{m\epsilon^2} \leq \frac{1}{4m\epsilon^2} \end{aligned}$$

2. 注意 $X_i \in [0,1]$，因此根据 Hoeffding's 不等式有

$$\begin{aligned} P^m\{\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h) \geq \epsilon\} &= P^m\{m\cdot \mathrm{er}_{\mathcal{D}}(h) - m\cdot \mathrm{er}(h) \geq m\epsilon\} \\ &\leq \exp\left(\frac{-2m^2\epsilon^2}{\sum_{i\in[m]}(1-0)^2}\right) = \exp(-2m\epsilon^2) \end{aligned}$$

   对称的有 $P^m\{\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h) \leq -\epsilon\} \leq \exp(-2m\epsilon^2)$，结合 union bound 有

$$P^m\{|\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} \leq 2\exp(-2m\epsilon^2)$$

当 $m\epsilon^2 \geq 2$ 时，Hoeffding's 不等式更紧。令 $2\exp(-2m\epsilon^2) = \delta$ 可得 $\epsilon = \sqrt{1/2m \ln 2/\delta}$，于是

$$\begin{aligned} P^m\left\{|\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\} \leq \delta \\ P^m\left\{|\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| < \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\} \geq 1-\delta \end{aligned} \tag{2}$$

注. 式(2)可以这样理解，在所有包含 m 个样本的数据集构成的空间中，给定 h，数据集有“好”有“坏”。好数据集上，h 的经验错误率可以近似泛化错误率，差别不超过 $\sqrt{1/2m \ln 2/\delta}$，此时 ERM 算法是靠谱的，而这样的好数据集在整个空间的占比至少为 $1-\delta$；在坏数据集上，h 的经验错误率不是泛化错误率的好近似，最小化它没啥意义，ERM 算法会学习失败，失败的概率即坏数据集的占比，不超过 δ。此外，δ 越小，m 越大，换言之，随着样本数的增加，好数据集越来越多，坏数据集越来越少。

式(2)只能用于 $h^\star$、不能用于 $h_{\mathcal{D}}^{\mathrm{ERM}}$，因为式(2)的推导用到了 Hoeffding's 不等式，其前提“经验错误率的期望等于泛化错误率”只对固定的 h 成立，而 $h_{\mathcal{D}}^{\mathrm{ERM}}$ 是依 $\mathcal{D}$ 而变化的。对此我们釜底抽薪，注意 $h_{\mathcal{D}}^{\mathrm{ERM}}$ 来自 $\mathcal{H}$，故要求 $\mathcal{H}$ 中的任意假设的坏数据集的占比都不超过 δ，一个更强的要求是所有假设的坏数据集占比的和不超过 δ，即 union bound，这就要求 $\mathcal{H}$ 是有限的，于是

$$P^m\{\exists h\in\mathcal{H}: |\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} \leq \sum_{h\in\mathcal{H}} P^m\{|\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} \leq |\mathcal{H}| 2\exp(-2m\epsilon^2) \tag{3}$$

令 $|\mathcal{H}|2\exp(-2m\epsilon^2)=\delta$ 可得 $\epsilon=\sqrt{1/2m\ln 2|\mathcal{H}|/\delta}$，于是

$$P^m\left\{\forall h\in\mathcal{H}:|\mathrm{er}_{\mathcal{D}}(h)-\mathrm{er}(h)|<\sqrt{\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}}\right\}\geq 1-\delta$$

显然上式等价于

$$P^m\left\{\max_{h\in\mathcal{H}}|\mathrm{er}_{\mathcal{D}}(h)-\mathrm{er}(h)|<\sqrt{\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}}\right\}\geq 1-\delta \tag{4}$$

由式(4)不难看出随着 $m\to\infty$，$\mathcal{H}$ 中所有假设的经验错误率 $\to$ 泛化错误率，这称为一致收敛。回代入式(1)可得

$$\begin{aligned}\mathrm{er}(h_{\mathcal{D}}^{\mathrm{ERM}})-\mathrm{er}(h^\star)&\leq|\mathrm{er}_{\mathcal{D}}(h_{\mathcal{D}}^{\mathrm{ERM}})-\mathrm{er}(h_{\mathcal{D}}^{\mathrm{ERM}})|+|\mathrm{er}_{\mathcal{D}}(h^\star)-\mathrm{er}(h^\star)|\\&\leq 2\max_{h\in\mathcal{H}}|\mathrm{er}_{\mathcal{D}}(h)-\mathrm{er}(h)|\\&<\sqrt{\frac{2}{m}\ln\frac{2|\mathcal{H}|}{\delta}}\ \text{with probability at least } 1-\delta\end{aligned}$$

上式称为估计误差界 (estimation error bound)。

定理 2. 对 $\forall\epsilon,\delta\in(0,1)$ 和 $\forall m\geq 2/\epsilon^2\ln 2|\mathcal{H}|/\delta$，任何有限假设空间 $\mathcal{H}$ 对 ERM 算法都是可学习的。

1.1 目标函数存在的情形

前面假设 P 是定义在 $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ 上的任意分布，因此存在同样的 x 对应不同 y 的可能，此时不存在目标函数 $t\in\mathcal{H}$ 使得 $y=t(x)$，这称为不可知 (agnostic) 学习。现对其做些简化，假设存在目标函数 $t\in\mathcal{H}$ 和定义在 $\mathcal{X}$ 上的分布 μ，使得对 $\mathcal{X}$ 的任意可测子集 $\mathcal{A}$ 有

$$P\{(x,t(x))\mid x\in\mathcal{A}\}=\mu(\mathcal{A}),\quad P\{(x,y)\mid x\in\mathcal{A},\ y\neq t(x)\}=0$$

换言之在此设定下，每个 x 有唯一的类别标记 $t(x)$，训练数据集 $\mathcal{D}=\{(x_i,t(x_i))\}_{i\in[m]}$，假设 h 的泛化错误率为 $\mathrm{er}(h,t)=\mu\{x\in\mathcal{X}\mid h(x)\neq t(x)\}$。

由于 $t\in\mathcal{H}$，因此 ERM 算法必然找到一个经验错误率为零的假设 h，若 $\mathrm{er}(h,t)\geq\epsilon$，则 iid 采样出 m 个 t、h 预测完全一致的样本的概率 $\leq(1-\epsilon)^m\leq\exp(-\epsilon m)$，故

$$P^m\{\exists h\in\mathcal{H}:\mathrm{er}(h,t)\geq\epsilon\}\leq|\mathcal{H}|\exp(-\epsilon m)$$

令 $|\mathcal{H}|\exp(-\epsilon m)=\delta$ 可得 $\epsilon=1/m\ln|\mathcal{H}|/\delta$，于是

$$P^m\left\{\max_{h\in\mathcal{H}}\mathrm{er}(h,t)<\frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta}\right\}\geq 1-\delta$$

对 $\forall m\geq 1/\epsilon\ln|\mathcal{H}|/\delta$，任何有限假设空间 $\mathcal{H}$ 对 ERM 算法都是可学习的。

注. 跟不可知学习的结果对比可以发现，此时对样本数的要求从 $O(1/\epsilon^2)$ 降到了 $O(1/\epsilon)$，这也表明该情形比不可知学习要简单。

2 增长函数

通常假设空间是无限的，此时式(3)中的 union bound 不能直接用，因为所有假设的坏数据集占比的和是无穷大。我们需要一种将无限归约到有限的工具，将 $|\mathcal{H}|$ 替换掉。注意数据集是有限的，因此定义增长函数 (growth function) 如下：

定义 3. 对任意假设空间 $\mathcal{H}$，增长函数 $\Pi_{\mathcal{H}}(m)$ 是其对 m 个样本的最大不同预测结果数

$$\forall m \in \mathbb{Z}^+ : \Pi_{\mathcal{H}}(m) = \max_{\mathcal{D} \in \mathcal{Z}^m} |\{[h(x_1), \dots, h(x_m)] : h \in \mathcal{H}\}|$$

注. 增长函数是个有限值，最大为 2^m。

通过考虑不同预测结果数，增长函数将无限的假设空间划分成了有限个等价类，每类中的假设对 m 个样本的预测完全一致。1971 年，Vapnik 证明了对任意假设空间 $\mathcal{H}$ 有

$$P^m\{\exists h \in \mathcal{H} : |\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} \leq 4\Pi_{\mathcal{H}}(2m) \exp(-m\epsilon^2/8) \tag{5}$$

式(5)的证明比较繁琐，依赖如下三个引理：

引理 4. 定义 $\mathcal{Z}^m$ 和 $\mathcal{Z}^{2m}$ 的子集

$$\mathcal{Q} = \{\mathcal{D} \mid \exists h \in \mathcal{H} : |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}}(h)| \geq \epsilon\}, \quad \mathcal{R} = \{(\mathcal{D}_1, \mathcal{D}_2) \mid \exists h \in \mathcal{H} : |\mathrm{er}_{\mathcal{D}_1}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \geq \epsilon/2\}$$

若 $m\epsilon^2 \geq 2$，则 $P^m(\mathcal{Q}) \leq 2P^{2m}(\mathcal{R})$。

证明. 根据三角不等式有

$$\left.\begin{aligned} |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_1}(h)| \geq \epsilon \\ |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \leq \epsilon/2 \end{aligned}\right\} \Longrightarrow |\mathrm{er}_{\mathcal{D}_1}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \geq \epsilon/2$$

于是

$$\begin{aligned} P^{2m}(\mathcal{R}) &\geq P^{2m}\{(\mathcal{D}_1, \mathcal{D}_2) \mid \exists h \in \mathcal{H} : |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_1}(h)| \geq \epsilon \wedge |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \leq \epsilon/2\} \\ &= \int_{\mathcal{Q}} P^m\{\mathcal{D}_2 \mid \exists h \in \mathcal{H} : |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_1}(h)| \geq \epsilon \wedge |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \leq \epsilon/2\} \,\mathrm{d}P^m(\mathcal{D}_1) \\ &\geq \int_{\mathcal{Q}} \frac{1}{2} \mathrm{d}P^m(\mathcal{D}_1) = \frac{1}{2} P^m(\mathcal{Q}) \end{aligned}$$

其中第二个不等号是因为对 $\forall \mathcal{D}_1 \in \mathcal{Q}(\exists h \in \mathcal{H} : |\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_1}(h)| \geq \epsilon)$，对这样的 h 由 Chebyshev's 不等式有

$$\begin{aligned} P^m\{|\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \geq \epsilon/2\} &= P^m\{|m \cdot \mathrm{er}(h) - m \cdot \mathrm{er}_{\mathcal{D}_2}(h)|^2 \geq (m\epsilon/2)^2\} \\ &\leq \frac{m \cdot \mathrm{er}(h)(1 - \mathrm{er}(h))}{(m\epsilon/2)^2} = \frac{4\mathrm{er}(h)(1 - \mathrm{er}(h))}{m\epsilon^2} \leq \frac{1}{m\epsilon^2} \leq \frac{1}{2} \end{aligned}$$

♣

记 Γ_m 是集合 $[2m]$ 上的一类置换，对 $\forall \sigma \in \Gamma_m$ 和 $\forall i \in [m]$，以下两种情形恰发生一种

1. 不变：$\sigma(i) = i$、$\sigma(m+i) = m+i$
2. 对换：$\sigma(i) = m+i$、$\sigma(m+i) = i$

对 $\forall \mathcal{D} \in \mathcal{Z}^{2m}$，假设其中元素是有顺序的，$\sigma\mathcal{D}$ 为对前半元素和后半元素的置换，例如 $\sigma = (25)(36)$ 作用到 $\mathcal{D} = \{z_1, z_2, z_3, z_4, z_5, z_6\}$ 上为 $\sigma\mathcal{D} = \{z_1, z_5, z_6, z_4, z_2, z_3\}$。

引理 5. 对 $\forall \mathcal{R} \subseteq \mathcal{Z}^{2m}$ 有 $P^{2m}(\mathcal{R}) = \mathbb{E}_{\mathcal{D}\sim P^{2m}}[\mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R})] \le \max_{\mathcal{D}\in\mathcal{Z}^{2m}} \mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R})$，其中 $\mathbb{P}_\sigma$ 表示从 Γ_m 中等概率挑选 σ。

证明. 置换是双射，因此对 $\forall \sigma \in \Gamma_m$ 有 $P^{2m}(\mathcal{R}) = P^{2m}\{\mathcal{D} \mid \sigma\mathcal{D} \in \mathcal{R}\}$，于是

$$
\begin{aligned}
P^{2m}(\mathcal{R}) &= \frac{1}{|\Gamma_m|} \sum_{\sigma\in\Gamma_m} P^{2m}\{\mathcal{D} \mid \sigma\mathcal{D} \in \mathcal{R}\} = \frac{1}{|\Gamma_m|} \sum_{\sigma\in\Gamma_m} \int_{\mathcal{Z}^{2m}} \mathbb{I}(\sigma\mathcal{D} \in \mathcal{R}) \mathrm{d}P^{2m}(\mathcal{D}) \\
&= \int_{\mathcal{Z}^{2m}} \frac{1}{|\Gamma_m|} \sum_{\sigma\in\Gamma_m} \mathbb{I}(\sigma\mathcal{D} \in \mathcal{R}) \mathrm{d}P^{2m}(\mathcal{D}) = \int_{\mathcal{Z}^{2m}} \mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R}) \mathrm{d}P^{2m}(\mathcal{D}) \le \max_{\mathcal{D}\in\mathcal{Z}^{2m}} \mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R})
\end{aligned}
$$

♣

引理 6. 对引理4中定义的集合 $\mathcal{R} = \{(\mathcal{D}_1, \mathcal{D}_2) \mid \exists h \in \mathcal{H} : |\mathrm{er}_{\mathcal{D}_1}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \ge \epsilon/2\} \subseteq \mathcal{Z}^{2m}$ 和从 Γ_m 中等概率挑选的 σ 有

$$
\max_{\mathcal{D}\in\mathcal{Z}^{2m}} \mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R}) \le 2\Pi_{\mathcal{H}}(2m) \exp(-m\epsilon^2/8)
$$

证明. 设 $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{2m}, y_{2m})\}$、$\mathcal{S} = \{x_1, x_2, \ldots, x_{2m}\}$，$\mathcal{H}$ 在 $\mathcal{S}$ 上的不同预测结果数为 $t \le \Pi_{\mathcal{H}}(2m)$，不妨设 $h_1, h_2, \ldots, h_t$ 就是 t 个预测不同的假设，易知 $\sigma\mathcal{D} \in \mathcal{R}$ 等价于

$$
\exists j \in [t] : \left| \frac{1}{m} \sum_{i\in[m]} \mathbb{I}(h_j(x_{\sigma(i)}) \ne y_{\sigma(i)}) - \frac{1}{m} \sum_{i\in[m]} \mathbb{I}(h_j(x_{\sigma(m+i)}) \ne y_{\sigma(m+i)}) \right| \ge \frac{\epsilon}{2}
$$

于是根据 union bound 有

$$
\begin{aligned}
\mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R}) &= \mathbb{P}_\sigma\left(\exists j \in [t] : \left| \frac{1}{m} \sum_{i\in[m]} (\mathbb{I}(h_j(x_{\sigma(i)}) \ne y_{\sigma(i)}) - \mathbb{I}(h_j(x_{\sigma(m+i)}) \ne y_{\sigma(m+i)})) \right| \ge \frac{\epsilon}{2} \right) \\
&\le \sum_{j\in[t]} \mathbb{P}_\sigma\left(\left| \frac{1}{m} \sum_{i\in[m]} (\mathbb{I}(h_j(x_{\sigma(i)}) \ne y_{\sigma(i)}) - \mathbb{I}(h_j(x_{\sigma(m+i)}) \ne y_{\sigma(m+i)})) \right| \ge \frac{\epsilon}{2} \right) \\
&\le t \max_{j\in[t]} \mathbb{P}_\sigma\left(\left| \frac{1}{m} \sum_{i\in[m]} (\mathbb{I}(h_j(x_{\sigma(i)}) \ne y_{\sigma(i)}) - \mathbb{I}(h_j(x_{\sigma(m+i)}) \ne y_{\sigma(m+i)})) \right| \ge \frac{\epsilon}{2} \right) \\
&\le \Pi_{\mathcal{H}}(2m) \max_{j\in[t]} \mathbb{P}_\sigma\left(\left| \frac{1}{m} \sum_{i\in[m]} (\mathbb{I}(h_j(x_{\sigma(i)}) \ne y_{\sigma(i)}) - \mathbb{I}(h_j(x_{\sigma(m+i)}) \ne y_{\sigma(m+i)})) \right| \ge \frac{\epsilon}{2} \right)
\end{aligned}
$$

注意求和中的 $\mathbb{I}(h_j(x_{\sigma(i)}) \ne y_{\sigma(i)}) - \mathbb{I}(h_j(x_{\sigma(m+i)}) \ne y_{\sigma(m+i)})$ 是个二值 rv 且

$$
= \begin{cases} \mathbb{I}(h_j(x_i) \ne y_i) - \mathbb{I}(h_j(x_{m+i}) \ne y_{m+i}), & \text{with probability 0.5} \\ \mathbb{I}(h_j(x_{m+i}) \ne y_{m+i}) - \mathbb{I}(h_j(x_i) \ne y_i), & \text{with probability 0.5} \end{cases}
$$

因此其均值为零，取值 $\in [-1,1]$，由 Hoeffding's 不等式有

$$\mathbb{P}_\sigma \left(\left| \frac{1}{m} \sum_{i\in[m]} (\mathbb{I}(h_j(x_{\sigma(i)}) \neq y_{\sigma(i)}) - \mathbb{I}(h_j(x_{\sigma(m+i)}) \neq y_{\sigma(m+i)})) \right| \geq \frac{\epsilon}{2} \right) \leq 2\exp\left(\frac{-2(m\epsilon/2)^2}{4m} \right) = 2\exp(-m\epsilon^2/8)$$

回代可得 $\mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R}) \leq \Pi_{\mathcal{H}}(2m) \max_{j\in[t]} 2\exp(-m\epsilon^2/8) = 2\Pi_{\mathcal{H}}(2m)\exp(-m\epsilon^2/8)$，由 $\mathcal{D}$ 的任意性知结论成立。 ♣

式(5)的证明. 结合引理4、引理5、引理6的结论易知当 $m\epsilon^2 \geq 2$ 时有

$$P^m\{\exists h \in \mathcal{H} : |\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} = P^m(\mathcal{Q}) \leq 2P^{2m}(\mathcal{R}) \leq 2 \max_{\mathcal{D}\in\mathcal{Z}^{2m}} \mathbb{P}_\sigma(\sigma\mathcal{D} \in \mathcal{R}) \leq 4\Pi_{\mathcal{H}}(2m)\exp(-m\epsilon^2/8)$$

当 $m\epsilon^2 < 2$ 时，式(5)右边大于 1，结论显然成立。 ♣

注. 引理4证明的最后若用 Hoeffding's 不等式则有 $P^m\{|\mathrm{er}(h) - \mathrm{er}_{\mathcal{D}_2}(h)| \geq \epsilon/2\} \leq 2\exp(-m\epsilon^2/2)$，该式在 $m\epsilon^2 > 2\ln 2$ 时有意义，在 $m\epsilon^2 \geq 5$ 时比 Chebyshev's 不等式的 $1/m\epsilon^2$ 更紧，式(5)可加强为

$$P^m\{\exists h \in \mathcal{H} : |\mathrm{er}_{\mathcal{D}}(h) - \mathrm{er}(h)| \geq \epsilon\} \leq \frac{2}{1 - 2\exp(-m\epsilon^2/2)} \Pi_{\mathcal{H}}(2m)\exp(-m\epsilon^2/8)$$

但不管 $m\epsilon^2$ 如何增大都有 $2/1-2\exp(-m\epsilon^2/2) > 2$，因此相对于 4，只是很小的常数倍改进。

3 VC 维

增长函数 $\Pi_{\mathcal{H}}(m)$ 与样本数 m 有关，计算起来不太方便，于是 Vapnik 和 Chervonenkis 提出了 VC 维，它是直接刻画 $\mathcal{H}$ 复杂度的标量，与 m 无关，比增长函数容易计算。

定义 7 (VC 维). 对于包含 m 个样本的数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i\in[m]}$，如果 $\{y_1, \ldots, y_m\}$ 不管如何取值，都 $\exists h \in \mathcal{H} : \mathrm{er}_{\mathcal{D}}(h) = 0$，则称 $\mathcal{D}$ 可以被 $\mathcal{H}$ 打散 (shattering)，即此时 $\Pi_{\mathcal{H}}(m) = 2^m$。

假设空间 $\mathcal{H}$ 的 VC 维是它能打散的最大数据集中的样本数：

$$\mathcal{V}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

注. $\mathcal{V}(\mathcal{H}) = m$ 表示存在 (不是任意) 一个 m 个样本的数据集可以被 $\mathcal{H}$ 打散，因此要证明某假设空间的 VC 维为 d 需要做两件事，一是构造一个可以被打散的 d 个样本的数据集，二是证明对于任意 $d+1$ 个样本的数据集都不可能被 $\mathcal{H}$ 打散。

例 8 ($\mathbb{R}^n$ 中的超平面集合的 VC 维为 $n+1$). 先构造可被打散的 $n+1$ 个样本构成的数据集

$$\{(\boldsymbol{x}_0 = \boldsymbol{0}, y_0), (\boldsymbol{x}_1 = \boldsymbol{e}_1, y_1), \ldots (\boldsymbol{x}_n = \boldsymbol{e}_n, y_n)\}$$

记 $\boldsymbol{w}^\top = [y_1, \ldots, y_n]$，则 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + y_0/2 = 0$ 可打散该数据集。

对于任意 $n+2$ 个样本，由 Radon's 定理知其必然可以分为两个子集，其凸包是相交的。注意若两个子集可以被超平面分开，那它们的凸包必然不相交，故不存在超平面可以将这 $n+2$ 个样本分开。

1972 年，Sauer 用 VC 维给出了增长函数的上界。

引理 9 (Sauer's 引理). 若 $\mathcal{V}(\mathcal{H}) = d$，则对 $\forall m \in \mathbb{Z}^+$ 有

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

证明. 若 $m \le d$，则存在 m 个样本的数据集被 $\mathcal{H}$ 打散，故

$$\Pi_{\mathcal{H}}(m) = 2^m = \sum_{i=0}^{m} \binom{m}{i} \le \sum_{i=0}^{d} \binom{m}{i}$$

下面考虑 $m > d$ 的情况，首先若 $d = 0$，即对任意将本，$\mathcal{H}$ 都只有一种预测，则有

$$\Pi_{\mathcal{H}}(m) = 1 = \binom{m}{0}$$

故引理对 $(\forall m \ge 1, d = 0)$ 都成立。若能用数学归纳法证明

$$\left.\begin{array}{r} (m-1, 0) \\ \vdots \\ (m-1, d-1) \\ (m-1, d) \end{array}\right\} \Longrightarrow (m, d)$$

则由引理对 $(\forall m \ge 1, d = 0)$ 和 $(m = 1, d = 1)$ 成立，可推得引理对 $(\forall m \ge 2, d = 1)$ 都成立；同理，由引理对 $(\forall m \ge 2, d = 0, 1)$ 和 $(m = 2, d = 2)$ 成立，可推得引理对 $(\forall m \ge 3, d = 2)$ 都成立，以此类推，可知引理对任意的 m 和 d 都成立。

假设引理对 $(m-1, 0), \ldots, (m-1, d)$ 成立。设 $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ 并且 $\mathcal{H}$ 对其不同预测结果数达到最大，即 $\Pi_{\mathcal{H}}(m)$。根据对 $\mathcal{S}$ 的预测结果不同，$\mathcal{H}$ 分成 $\Pi_{\mathcal{H}}(m)$ 个等价类，记该等价类集合为 $\mathcal{H}_{\mathcal{S}}$，从每一个等价类中任选一个假设构成集合 $\mathcal{G}$，显然 $|\mathcal{H}_{\mathcal{S}}| = |\mathcal{G}| = \Pi_{\mathcal{H}}(m)$。

设 $\mathcal{S}' = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m-1}\}$，同样根据对 $\mathcal{S}'$ 的预测结果不同可以得到 $\mathcal{H}$ 在 $\mathcal{S}'$ 上的等价类集合 $\mathcal{H}_{\mathcal{S}'}$，从每一个等价类中任选一个假设构成集合 $\mathcal{G}'$。对于 $\mathcal{H}_{\mathcal{S}'}$ 中每一个等价类，若其中所有假设对 $\boldsymbol{x}_m$ 的预测一致，则它们还会作为一个等价类出现在 $\mathcal{H}_{\mathcal{S}}$ 中；否则则会一分为二作为两个等价类出现在 $\mathcal{H}_{\mathcal{S}}$ 中，设 $\mathcal{H}_{\mathcal{S}'}$ 中被一分为二的等价类集合为 $\mathcal{H}_{\mathcal{S}''}$，从其每一个等价类中任选一个假设构成集合 $\mathcal{G}''$，于是有

$$|\mathcal{G}| = |\mathcal{G}'| + |\mathcal{G}''|$$

显然 $\mathcal{V}(\mathcal{G}') \le d$，于是由增长函数的定义和归纳假设有

$$|\mathcal{G}'| \le \Pi_{\mathcal{G}'}(m-1) \le \sum_{i=0}^{\mathcal{V}(\mathcal{G}')} \binom{m-1}{i} \le \sum_{i=0}^{d} \binom{m-1}{i}$$

若 $\forall \mathcal{R} \subseteq \mathcal{S}'$ 可被 $\mathcal{G}''$ 打散，则 $\mathcal{R} \cup \{\boldsymbol{x}_m\}$ 可被 $\mathcal{G}$ 打散，于是 $\mathcal{V}(\mathcal{G}'') \le d-1$，否则 $\mathcal{V}(\mathcal{G}) > d$，由增长函数的定义和归纳假设有

$$|\mathcal{G}''| \le \Pi_{\mathcal{G}''}(m-1) \le \sum_{i=0}^{\mathcal{V}(\mathcal{G}'')} \binom{m-1}{i} \le \sum_{i=0}^{d-1} \binom{m-1}{i}$$

回代有

$$|\mathcal{G}| \leq \sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{d-1}\binom{m-1}{i}=1+\sum_{i=1}^{d}\binom{m-1}{i}+\sum_{i=1}^{d}\binom{m-1}{i-1}=1+\sum_{i=1}^{d}\binom{m}{i}=\sum_{i=0}^{d}\binom{m}{i}$$

故引理对 (m,d) 成立。 ♣

根据 Sauer's 引理可得

推论 10. 若 $\mathcal{V}(\mathcal{H})=d$，则对任意正整数 $m \geq d$ 有

$$\Pi_{\mathcal{H}}(m) \leq (^{em}/_{d})^{d} = O(m^d)$$

证明. 由 Sauer's 引理有

$$\begin{aligned}\Pi_{\mathcal{H}}(m) &\leq \sum_{i=0}^{d}\binom{m}{i} \leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \\ &= \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} = \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m} \leq \left(\frac{m}{d}\right)^{d} e^{d} = \left(\frac{em}{d}\right)^{d} = O(m^d)\end{aligned}$$

♣

回代入式(5)可得

$$P^m\{\exists h \in \mathcal{H} : |\mathrm{er}_{\mathcal{D}}(h)-\mathrm{er}(h)| \geq \epsilon\} \leq 4\Pi_{\mathcal{H}}(2m)\exp(^{-m\epsilon^2}/_8) \leq 4(^{2em}/_d)^d \exp(^{-m\epsilon^2}/_8)$$

令 $4(^{2em}/_d)^d \exp(^{-m\epsilon^2}/_8)=\delta$ 可得 $\epsilon=\sqrt{^8/_m(d\ln {^{2em}/_d}+\ln {^4/_\delta})}$，于是有基于 VC 维的一致收敛界:

$$P^m\left\{\sup_{h\in\mathcal{H}} |\mathrm{er}_{\mathcal{D}}(h)-\mathrm{er}(h)| < \sqrt{\frac{8}{m}\left(d\ln\frac{2em}{d}+\ln\frac{4}{\delta}\right)}\right\} \geq 1-\delta$$

和估计误差界:

$$P^m\left\{\sup_{h\in\mathcal{H}} |\mathrm{er}(h_{\mathcal{D}}^{\mathrm{ERM}})-\mathrm{er}(h^\star)| < \sqrt{\frac{32}{m}\left(d\ln\frac{2em}{d}+\ln\frac{4}{\delta}\right)}\right\} \geq 1-\delta$$

定理 11. 对 $\forall\epsilon,\delta \in (0,1)$ 和 $\forall m \geq {^{32}/_{\epsilon^2}}(d\ln {^{2em}/_d}+\ln {^4/_\delta})$，任何 VC 维为 d 的假设空间 $\mathcal{H}$ 对 ERM 算法都是可学习的。

4 附录

定理 12 (Markov's 不等式). 设非负随机变量 X 具有有限均值 $\mu=\mathbb{E}[X]$，则对任意实数 $t>0$ 有

$$\mathbb{P}(X \geq t) \leq \frac{\mu}{t}$$

证明. 由于

$$\mu=\int_0^\infty xp(x)\mathrm{d}x=\int_0^t xp(x)\mathrm{d}x+\int_t^\infty xp(x)\mathrm{d}x \geq 0+t\int_t^\infty p(x)\mathrm{d}x = t\mathbb{P}(X\geq t)$$

两边除以 t 后命题得证。 ♣

定理 13 (Chebyshev's 不等式). 设随机变量 X 具有有限均值 μ 和方差 σ^2，则对任意实数 $t > 0$ 有

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

证明. 由于 $|X-\mu|^2$ 是非负随机变量，由 Markov's 不等式可得

$$\mathbb{P}(|X-\mu| \geq t) = \mathbb{P}(|X-\mu|^2 \geq t^2) \leq \frac{\mathbb{E}[|X-\mu|^2]}{t^2} = \frac{\sigma^2}{t^2}$$

故命题得证。 ♣

引理 14 (Hoeffding's 法则). 设 X 是随机变量且 $\mathbb{E}[X] = 0$，$X \in [a, b]$，则对任意实数 $t > 0$ 有

$$\mathbb{E}[\exp(tX)] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

证明. 由于 $\exp(tX)$ 是凸函数，对于任意 $x \in [a, b]$，由 Jensen's 不等式得

$$\exp(tX) = \exp\left(\frac{b-x}{b-a}ta + \frac{x-a}{b-a}tb\right) \leq \frac{b-x}{b-a}\exp(ta) + \frac{x-a}{b-a}\exp(tb)$$

注意 $\mathbb{E}[X] = 0$，于是

$$\mathbb{E}[\exp(tX)] \leq \mathbb{E}\left[\frac{b-X}{b-a}\exp(ta) + \frac{X-a}{b-a}\exp(tb)\right] = \frac{b}{b-a}\exp(ta) + \frac{-a}{b-a}\exp(tb) = \exp(\phi(t))$$

其中

$$\phi(t) = \ln\left(\frac{b}{b-a}\exp(ta) + \frac{-a}{b-a}\exp(tb)\right) = ta + \ln\left(\frac{b}{b-a} + \frac{-a}{b-a}\exp(t(b-a))\right)$$

对于任意 $t > 0$ 易知有

$$\begin{aligned}
\phi'(t) &= a - a\exp(t(b-a)) \Big/ \left(\frac{b}{b-a} + \frac{-a}{b-a}\exp(t(b-a))\right) \\
&= a - a \Big/ \left(\frac{b}{b-a}\exp(-t(b-a)) - \frac{a}{b-a}\right) \\
\phi''(t) &= -ab\exp(-t(b-a)) \Big/ \left(\frac{b}{b-a}\exp(-t(b-a)) - \frac{a}{b-a}\right)^2 \\
&= \frac{\alpha(1-\alpha)\exp(-t(b-a))(b-a)^2}{((1-\alpha)\exp(-t(b-a)) + \alpha)^2} \\
&= \frac{(1-\alpha)\exp(-t(b-a))}{(1-\alpha)\exp(-t(b-a)) + \alpha}\frac{\alpha}{(1-\alpha)\exp(-t(b-a)) + \alpha}(b-a)^2 \\
&\leq \frac{1}{4}(b-a)^2
\end{aligned}$$

其中 $\alpha = {}^{-a}/_{b-a}$、$1-\alpha = {}^{b}/_{b-a}$，注意 $\phi(0) = \phi'(0) = 0$，于是存在 $\theta \in [0, t]$ 满足

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta) \leq \frac{t^2(b-a)^2}{8}$$

♣

定理 15 (Hoeffding's 不等式). 设 $X_1,\ldots,X_m$ 为相互独立的随机变量，且 $X_i \in [a_i, b_i]$，$i \in [m]$。记 $S_m = \sum_{i\in[m]} X_i$，对任意实数 $\epsilon > 0$ 有

$$\mathbb{P}(S_m - \mathbb{E}[S_m] \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i\in[m]}(b_i - a_i)^2}\right), \quad \mathbb{P}(S_m - \mathbb{E}[S_m] \leq -\epsilon) \leq \left(\frac{-2\epsilon^2}{\sum_{i\in[m]}(b_i - a_i)^2}\right)$$

证明. 对任意实数 $t > 0$，由 Markov's 不等式和 Hoeffding's 法则可得

$$\begin{aligned}
\mathbb{P}(S_m - \mathbb{E}[S_m] \geq \epsilon) &= \mathbb{P}(\exp(t(S_m - \mathbb{E}[S_m])) \geq \exp(t\epsilon)) \\
&\leq \frac{\mathbb{E}[\exp(t(S_m - \mathbb{E}[S_m])]}{\exp(t\epsilon)} \\
&= \frac{\mathbb{E}[\prod_{i\in[m]} \exp(t(X_i - \mathbb{E}[X_i]))]}{\exp(t\epsilon)} = \frac{\prod_{i\in[m]} \mathbb{E}[\exp(t(X_i - \mathbb{E}[X_i]))]}{\exp(t\epsilon)} \\
&\leq \frac{\prod_{i=1}^m \exp\left({}^{t^2(b_i - a_i)^2}/_8\right)}{\exp(t\epsilon)} = \exp\left(\frac{t^2\sum_{i\in[m]}(b_i - a_i)^2}{8} - t\epsilon\right)
\end{aligned}$$

令 $t = {}^{4\epsilon}/_{\sum_{i\in[m]}(b_i - a_i)^2}$ 即可。 ♣

定理 16 (Radon's 定理). $\mathbb{R}^n$ 中任意 $n+2$ 个点构成的集合 $\mathcal{D}$ 都可以划分成两个子集 $\mathcal{D}_1$ 和 $\mathcal{D}_2$ 使其凸包相交。

证明. 设 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n+2}\} \subseteq \mathbb{R}^n$，考虑

$$\sum_{i\in[n+2]} \alpha_i \boldsymbol{x}_i = \mathbf{0}, \quad \sum_{i\in[n+2]} \alpha_i = 0$$

这是包含 $n+2$ 个变量、由 $n+1$ 个方程组成的线性方程组，因此必然有非零解，设 $\beta_1, \ldots, \beta_{n+2}$ 是一组非零解，记

$$\mathcal{I}_1 = \{i \mid \beta_i > 0\}, \quad \mathcal{I}_2 = \{i \mid \beta_i \leq 0\}, \quad \beta = \sum_{i\in\mathcal{I}_1} \beta_i, \quad \mathcal{D}_1 = \{\boldsymbol{x}_i \mid i \in \mathcal{I}_1\}, \quad \mathcal{D}_2 = \{\boldsymbol{x}_i \mid i \in \mathcal{I}_2\}$$

显然 $\mathcal{D}_1$ 与 $\mathcal{D}_2$ 是对 $\mathcal{D}$ 的一个划分，易知

$$\sum_{i\in\mathcal{I}_1} \beta_i \boldsymbol{x}_i + \sum_{i\in\mathcal{I}_2} \beta_i \boldsymbol{x}_i = \mathbf{0} \Longrightarrow \sum_{i\in\mathcal{I}_1} \beta_i \boldsymbol{x}_i = -\sum_{i\in\mathcal{I}_2} \beta_i \boldsymbol{x}_i \Longrightarrow \sum_{i\in\mathcal{I}_1} \frac{\beta_i}{\beta} \boldsymbol{x}_i = \sum_{i\in\mathcal{I}_2} \frac{-\beta_i}{\beta} \boldsymbol{x}_i$$

注意 $\sum_{i\in\mathcal{I}_1} {}^{\beta_i}/_{\beta} = \sum_{i\in\mathcal{I}_2} {}^{-\beta_i}/_{\beta} = 1$，这表明存在一个点既属于 $\mathcal{D}_1$ 的凸包、也属于 $\mathcal{D}_2$ 的凸包。 ♣

注. $n+2$ 个点是必须的，如果只有 $n+1$ 个点，线性方程组未必有非零解。