Learning Theory: generalization bound

Apple Zhang

2023 年 6 月 13 日

1 集中不等式 (Concentration inequality)

定理 1 (Chernoff bounding). 对于任意随机变量 X 以及 $\epsilon > 0$、$t > 0$，由 *Markov* 不等式可得

$$\Pr(X \geq \epsilon) = \Pr(e^{tX} \geq e^{t\epsilon}) \leq e^{-t\epsilon}\mathbb{E}[e^{tX}] \tag{1}$$

定理 2 (Hoeffding inequality). 设 $\{X_i\}_{i=1}^n$ 为一组独立随机变量且 $a_i \leq X_i \leq b_i$，且

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mu = \mathbb{E}(\bar{X}). \tag{2}$$

则对于任意$\epsilon > 0$，有以下不等式成立

$$\Pr(\bar{X} - \mu \geq \epsilon) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{3}$$

$$\Pr(\bar{X} - \mu \leq -\epsilon) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{4}$$

或等价的，若令 $S_n = \sum_{i=1}^n X_i$，则

$$\Pr(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{5}$$

$$\Pr(S_n - \mathbb{E}[S_n] \leq -\epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{6}$$

定义 1 (鞅). 随机过程序列 $Z_1, Z_2, \cdots$ 对一切 n 有 $\mathbb{E}[Z] < \infty$，且

$$\mathbb{E}[Z_{i+1}|Z_1, \cdots, Z_i] = Z_i, \tag{7}$$

则称 $Z_1, Z_2, \cdots$ 为鞅 *(Martingale)*. 定义 $Y_i = Z_i - Z_{i-1}$，则

$$\mathbb{E}[Y_{i+1}|Z_1, \cdots, Z_i] = 0, \tag{8}$$

则序列 $Y_1, Y_2, \cdots$ 称为鞅差 *(Martingale difference)*.

基于鞅的定义，可以得到 Azuma inequality：

定理 3 (Azuma inequality). 假设 $X_1, X_2, \cdots$ 是鞅，令其鞅差为 $Y_i = X_i - X_{i-1}$，若对任意 i 都存在 $c_i > 0$ 使得 $|Y_i| \leq c$ 成立，则

$$\Pr[X_n - X_0 \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right), \tag{9}$$

$$\Pr[X_n - X_0 \leq -\epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \tag{10}$$

在 Azuma inequality 的基础上可以进一步得到：

定理 4 (McDiarmid inequality)**.** 设一组独立随机变量 $\{X_i\}_{i=1}^n \in \mathcal{X}^n$，且存在 $c_1, \cdots, c_n > 0$ 使得对于函数 $f : \mathcal{X} \mapsto \mathbb{R}$，

$$|f(x_1, \cdots, x_t, \cdots, x_n) - f(x_1, \cdots, x_t', \cdots, x_n)| \leq c_t. \tag{11}$$

对任意 $t \in [n]$ 和任意 $x_1, \cdots, x_n, x_t' \in \mathcal{X}$ 成立. 设 $f(S) = f(X_1, \cdots, X_n)$，则对于任意 $\epsilon > 0$ 有

$$\Pr[f(S) - \mathbb{E}[f(S)] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right), \tag{12}$$

$$\Pr[f(S) - \mathbb{E}[f(S)] \leq -\epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \tag{13}$$

可以看到 Hoeffding inequality 就是当 $f(S) = \sum_{i=1}^n X_i$ 时的一个特殊形式.

2 Rademacher 复杂度

定义 2 (Rademacher 复杂度)**.** 设从分布 $\mathcal{D}$ 中抽样出的样本集合 $S = \{z_i\}_{i=1}^n$，且 $\mathcal{G}$ 表示将样本空间映射到区间 $[a, b]$ 的函数族，则经验 *Rademacher* 复杂度定义为

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)\right]. \tag{14}$$

Rademacher 复杂度定义为经验 *Rademacher* 复杂度的期望，即

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^n}[\hat{\mathfrak{R}}_S(\mathcal{G})] = \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)\right] \tag{15}$$

Rademacher 复杂度具有以下运算

$$\mathfrak{R}_n(a\mathcal{G} + b) = |a|\mathfrak{R}_n(\mathcal{G}), \quad \mathfrak{R}_n(\mathbf{conv}(\mathcal{G})) = \mathfrak{R}_n(\mathcal{G}). \tag{16}$$

其中 $\mathbf{conv}(\cdot)$ 为取凸包运算. 事实上，对于 $\mathbf{conv}(\mathcal{G}) = \{\sum_{j=1}^k \theta_i g_i : g_i \in \mathcal{G}, \quad \theta_j \geq 0, \sum_{j=1}^k \theta_i = 1, k \in \mathbb{N}_+\}$.

$$\begin{aligned}
\mathfrak{R}_n(\mathbf{conv}(\mathcal{G})) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_j\in\mathcal{G}, \mathbf{1}^T\boldsymbol{\theta}=1, \theta_i\geq 0, k\in\mathbb{N}_+} \frac{1}{n}\sum_{i=1}^n \sigma_i \sum_{j=1}^k \theta_j g_j(x_i)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_j\in\mathcal{G}} \sup_{\mathbf{1}^T\boldsymbol{\theta}=1, \theta_i\geq 0, k\in\mathbb{N}_+} \frac{1}{n}\sum_{j=1}^k \theta_j \left(\sum_{i=1}^n \sigma_i g_j(x_i)\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_j\in\mathcal{G}} \max_j \frac{1}{n}\sum_{i=1}^n \sigma_j g_j(x_i)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\right] = \mathfrak{R}_n(\mathcal{G}).
\end{aligned}$$

接下来的不等式则是基于 Rademacher 复杂度的泛化界

定理 5. 假设样本 $Z_1, \cdots, Z_n$ 为独立且采样自分布 $\mathcal{D}$ 的随机变量，对于函数族 $\mathcal{G}$，其中任意 $f \in \mathcal{G}$ 的值域为 $[a, b]$，则存在 $\delta > 0$，使得至少有 $1-\delta$ 的概率使得对于样本集 S，有

$$\mathbb{E}_Z[g(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} g(Z_i) + 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}, \quad \forall g \in \mathcal{G}. \tag{17}$$

$$\mathbb{E}_Z[g(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} g(Z_i) + 2\hat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}, \quad \forall g \in \mathcal{G}. \tag{18}$$

证明. 可以采用 McDiarmid inequality 进行证明. 令

$$T(g) = \mathbb{E}_Z(g(Z)), \quad \hat{T}_S(g) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i). \tag{19}$$

定义样本集上的函数

$$\phi(S) = \sup_{g\in\mathcal{G}}\left(T(g) - \hat{T}_S(g)\right) \tag{20}$$

同时考虑以下样本集

$$S = \{Z_1, \cdots, Z_t, \cdots, Z_n\}, \quad S' = \{Z_1, \cdots, Z_t', \cdots, Z_n\}. \tag{21}$$

则有

$$\begin{aligned}
\phi(S) - \phi(S') &= \sup_{g\in\mathcal{G}}\left(T(g) - \hat{T}_S(g)\right) - \sup_{g\in\mathcal{G}}\left(T(g) - \hat{T}_{S'}(g)\right) \\
&\leq \sup_{g\in\mathcal{G}}\left(T(g) - \hat{T}_S(g) - T(g) + \hat{T}_{S'}(g)\right) && \text{use } \sup(U+V) \leq \sup(U) + \sup(V). \\
&= \sup_{g\in\mathcal{G}}\left(\hat{T}_{S'}(g) - \hat{T}_S(g)\right) \\
&= \frac{1}{n}\sup_{g\in\mathcal{G}}(g(Z_t') - g(Z_t)) && . \\
&\leq \frac{b-a}{n}
\end{aligned} \tag{22}$$

因此，根据 McDiarmid inequality, 可以得到

$$\Pr[\phi(S) - \mathbb{E}_S[\phi(S)] \geq \epsilon] \leq \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right) \tag{23}$$

或者，至少有 $1-\delta$ 的概率，

$$\phi(S) \leq \mathbb{E}_S[\phi(S)] + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{24}$$

接下来计算 $\mathbb{E}_S[\phi(S)]$ 的上界，我们有

$$\begin{aligned}
\mathbb{E}_S[\phi(S)] &= \mathbb{E}_S[\sup_{g\in\mathcal{G}} T(g) - \hat{T}_S(g)] \\
&= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right] \\
&= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \mathbb{E}_{\tilde{S}}\left[\frac{1}{n}\sum_{i=1}^n g(\tilde{Z}_i)\right] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right] \\
&= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \mathbb{E}_{\tilde{S}}\left[\frac{1}{n}\sum_{i=1}^n g(\tilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i)\right]\right] \\
&\le \mathbb{E}_{S,\tilde{S}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left(g(\tilde{Z}_i) - g(Z_i)\right)\right], && \text{use } \sup\mathbb{E}[x] \le \mathbb{E}[\sup x]. \\
&= \mathbb{E}_{S,\tilde{S},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i\left(g(\tilde{Z}_i) - g(Z_i)\right)\right] && \Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = \frac{1}{2} \\
&\le \mathbb{E}_{\tilde{S},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(\tilde{Z}_i)\right] + \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n -\sigma_i g(Z_i)\right] \\
&= 2\mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(\tilde{Z}_i)\right] = 2\mathfrak{R}_n(\mathcal{G}). && (25)
\end{aligned}$$

由此以及 (24) 可以推出 (17).

同理，令 $\psi(S) = \hat{\mathfrak{R}}_S(\mathcal{G})$ 容易知道

$$\psi(S) - \psi(S') = \hat{\mathfrak{R}}_S(\mathcal{G}) - \hat{\mathfrak{R}}_{S'}(\mathcal{G}) \le \frac{b-a}{n}. \tag{26}$$

再次根据 McDiarmid inequality, 得到

$$\Pr\left[\mathfrak{R}_n(\mathcal{G}) \ge \hat{\mathfrak{R}}_S(\mathcal{G}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\right] \le \delta. \tag{27}$$

由于 δ 的任意性，所以

$$\Pr\left[\mathfrak{R}_n(\mathcal{G}) \ge \hat{\mathfrak{R}}_S(\mathcal{G}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right] \le \frac{\delta}{2}. \tag{28}$$

由于不等式两端同乘以 2 不改变概率，因此

$$\Pr\left[2\mathfrak{R}_n(\mathcal{G}) \ge 2\hat{\mathfrak{R}}_S(\mathcal{G}) + 2(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right] \le \frac{\delta}{2}. \tag{29}$$

同时根据 (17) 推出

$$\Pr\left[\mathbb{E}_Z[g(Z)] \ge \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right] \le \frac{\delta}{2}. \tag{30}$$

依据 union bound, 则有

$$\Pr\left[\mathbb{E}_Z[g(Z)] \ge \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2\hat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right] \le \frac{\delta}{2} + \frac{\delta}{2} = \delta. \tag{31}$$

至此 (18) 证毕. □

在二分类情形下，可以进一步得到假设集 $\mathcal{H}$ 的泛化界. 首先介绍下面的引理

引理 1. 假设 $\mathcal{H}$ 是取值为 $\{+1,-1\}$ 的函数族，$\mathcal{G}$ 为关于 $\mathcal{H}$ 的分类损失函数族，即 $\mathcal{G}=\{(x,y)\mapsto 1_{h(x)\neq y}\mid h\in\mathcal{H}\}$. 记样本集合为 $S=((x_1,y_1),\cdots,(x_n,y_n))$，且 $S_{\mathcal{X}}=(x_1,\cdots,x_n)$. 则有

$$\hat{\mathfrak{R}}_S(\mathcal{G})=\frac{1}{2}\hat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H}). \tag{32}$$

证明. 根据定义代入:

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{G}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n\sigma_i g(x_i,y_i)\right]\\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\sigma_i 1_{h(x_i)\neq y_i}\right]\\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\sigma_i \frac{1-y_ih(x_i)}{2}\right]\\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\sigma_i y_i h(x_i)\right], \qquad \text{use } \mathbb{E}_\sigma[\sigma]=0\\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\sigma_i h(x_i)\right]=\frac{1}{2}\mathfrak{R}_{S_{\mathcal{X}}}(\mathcal{H}).
\end{aligned}$$

证毕. □

因此对于二分类问题，可以直接得出其基于 Rademacher 复杂度的泛化界:

推论 1. 假设 $\mathcal{H}$ 是取值为 $\{+1,-1\}$ 的假设集，样本集合 $S\subset\mathcal{X}\times\{-1,1\}^n$ 对应的数据分布为 $\mathcal{D}$. 则至少有 $1-\delta$ 的概率，对任意 $h\in\mathcal{H}$ 有

$$R(h)\le\hat{R}_S(h)+\mathfrak{R}_n(\mathcal{H})+\sqrt{\frac{\log(1/\delta)}{2n}},\quad \forall g\in\mathcal{G}. \tag{33}$$

$$R(h)\le\hat{R}_S(h)+\hat{\mathfrak{R}}_S(\mathcal{H})+3\sqrt{\frac{\log(2/\delta)}{2n}},\quad \forall g\in\mathcal{G}. \tag{34}$$

其中，$R(h)$ 和 $\hat{R}_S(h)$ 分别为泛化误差和经验误差

$$R(h)=\Pr_{(x,y)\sim\mathcal{D}}[h(x)\neq y]=\mathbb{E}_{(x,y)\sim\mathcal{D}}[1_{h(x)\neq y}],\quad \hat{R}_S(h)=\frac{1}{n}\sum_{i=1}^n 1_{h(x)\neq y}. \tag{35}$$

2.1 Rademacher 复杂度的常用结论

引理 2 (Massart Lemma)**.** 假设 $\mathcal{A}\subseteq\mathbb{R}^n$ 为有限集，记 $r=\sup_{\mathbf{x}\in\mathcal{A}}\|\mathbf{x}\|_2$，则

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{x}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\sigma_i x_i\right]\le\frac{r\sqrt{2\log|\mathcal{A}|}}{n}. \tag{36}$$

引理 3 (Talagrand Lemma)**.** 设 $\phi:\mathbb{R}\mapsto\mathbb{R}$ 是一个 *L-Lipschitz* 函数，则

$$\mathfrak{R}_S(\phi\circ\mathcal{H})\le L\mathfrak{R}_S(\mathcal{H}). \tag{37}$$

其中 $\phi\circ\mathcal{H}=\{z\mapsto\phi(h(z)):h\in\mathcal{H}\}$ 表示函数集的复合.

未完待续...

2.2 基于 Covering Number 的复杂度上界

2.2.1 Covering 与 Packing

定义 3 (ϵ-covering). *在度量空间的子集 $\mathcal{V}$ 上定义度量 ρ，其 ϵ-covering 表示有限集 $\{v_1, \cdots, v_N\}$，使得任意 v_i, v_j $(i \neq j)$，有 $\rho(v_i, v_j) > \epsilon$. 同时，$\epsilon$-packing number 定义为最大的 ϵ-packin 的集合大小，记为*

$$N(\epsilon, \mathcal{V}, \rho) = \min\{|\boldsymbol{\Theta}| : \boldsymbol{\Theta} \subset \mathcal{V},\ \forall v \in \mathcal{V},\ \exists v_i \in \boldsymbol{\Theta}, \rho(v, v_i) \leq \epsilon\}. \tag{38}$$

定义 4 (ϵ-packing). *在度量空间的子集 $\mathcal{V}$ 上定义度量 ρ，其 ϵ-packing 表示有限集 $\{v_1, \cdots, v_N\}$，使得对于任意 $v \in \mathcal{V}$，总存在 $i \in [N]$ 使得 $\rho(v_i, v) \leq \epsilon$. 同时，$\epsilon$-covering number 定义为最小的 ϵ-covering 的集合大小，记为*

$$M(\epsilon, \mathcal{V}, \rho) = \max\{|\boldsymbol{\Theta}| : \boldsymbol{\Theta} \subset \mathcal{V},\ \forall v_i, v_j \in \boldsymbol{\Theta},\ v_i \neq v_j, \rho(v_i, v_j) > \epsilon\}. \tag{39}$$

引理 4. *对任意 $\epsilon > 0$，有以下不等式成立*

$$M(2\epsilon, \mathcal{V}, \rho) \leq N(\epsilon, \mathcal{V}, \rho) \leq M(\epsilon, \mathcal{V}, \rho). \tag{40}$$

这个引理的证明很有趣，可以作为小小的思维锻炼.

证明. 设 $\boldsymbol{\Theta}^*$ 是最大的 ϵ-packing，则对任意 $v \in \mathcal{V} \backslash \boldsymbol{\Theta}^*$，总存在 $\theta \in \boldsymbol{\Theta}^*$，使得 $\rho(v, \theta) \leq \epsilon$，否则，就说明任意 $\theta \in \boldsymbol{\Theta}^*$ 都有 $\rho(v, \theta_i) > \epsilon$，所以 $\boldsymbol{\Theta}^* \cup \{v\}$ 将形成一个更大的 ϵ-packing，与 $\boldsymbol{\Theta}^*$ 是最大的 ϵ-packing 矛盾. 这说明 $\boldsymbol{\Theta}^*$ 是一个 ϵ-covering. 则根据定义 $M(\epsilon, \mathcal{V}, \rho) = |\boldsymbol{\Theta}| \geq N(\epsilon, \mathcal{V}, \rho)$.

另一方面，任取 $\boldsymbol{\Theta} = \{\theta_1, \cdots, \theta_m\}$ 为 2ϵ-packing，以及 $\boldsymbol{\Psi} = \{\psi_1, \cdots, \psi_n\}$ 为 ϵ-covering，如果 $m > n$，则根据鸽巢原理，必将存在 $\theta_i, \theta_j \in \boldsymbol{\Theta}$ 以及一个 $\psi_k \in \boldsymbol{\Psi}$，使得 $\rho(\theta_i, \psi_k) \leq \epsilon$ 且 $\rho(\theta_j, \psi_k) \leq \epsilon$. 根据三角不等式，有 $\rho(\theta_i, \theta_j) \leq \rho(\theta_i, \psi_k) + \rho(\theta_j, \psi_k) \leq 2\epsilon$，而这与 $\rho(\theta_i, \theta_j) > 2\epsilon$ 矛盾. 因此，$m \leq n$，于是说明 $M(2\epsilon, \mathcal{V}, \rho) \leq N(\epsilon, \mathcal{V}, \rho)$. □

于是我们就有以下定理来得到 covering number 的上界.

定理 6. *对于任意 ϵ，有以下不等式成立*

$$N(\epsilon, \mathcal{V}, \rho) \leq M(\epsilon, \mathcal{V}, \rho) \leq \frac{\mathrm{vol}(\mathcal{V} + B(\epsilon/2))}{\mathrm{vol}(B(\epsilon/2))}. \tag{41}$$

其中的“+”表示 Minkovski 加法，$B(\epsilon)$ 表示圆心在原点，半径为 ϵ 的圆.

证明. 设 $\boldsymbol{\Theta} = \{\theta_1, \cdots, \theta_m\}$ 是一个 ϵ-packing，则 $B(\theta_i, \epsilon/2)$ 都是彼此不相交的. 并且

$$\bigcup_{i=1}^{m} B(\theta_i, \epsilon/2) \subset \boldsymbol{\Theta} + B(\epsilon/2) \tag{42}$$

则有

$$\mathrm{vol}(\boldsymbol{\Theta} + B(\epsilon/2)) \geq \mathrm{vol}\left(\bigcup_{i=1}^{m} B(\theta_i, \epsilon/2)\right) = m\mathrm{vol}(B(\epsilon/2)). \tag{43}$$

说明对任意的 ϵ-packing，其集合大小满足

$$m \leq \frac{\mathrm{vol}(\mathcal{V} + B(\epsilon/2))}{\mathrm{vol}(B(\epsilon/2))} \tag{44}$$

即所证明的结论. □

定理 7. *对于任意 ϵ，设 $\phi : \mathbb{R} \mapsto \mathbb{R}$ 是一个 L-Lipschitz 函数，则*

$$N(\epsilon, \phi \circ \mathcal{V}, \rho) \leq N(\epsilon/L, \mathcal{V}, \rho). \tag{45}$$

2.2.2 Dubley's Theorem

接下来，我们将通过 covering number 的角度得到 Rademacher 复杂度的上界. 这个结论由 Dudley's Theorem 给出：

定理 8 (Dudley's Theorem)**.** *若 $\mathcal{F}$ 表示函数族 $f : \mathcal{X} \mapsto \mathbb{R}$，则*

$$\mathfrak{R}_S(\mathcal{F}) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \tag{46}$$

若 $\mathcal{F}$ 中每个函数值域都在 $[-1, 1]$，则有

$$\mathfrak{R}_S(\mathcal{F}) \leq 12 \int_0^1 \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \tag{47}$$

事实上，更一般的形式是

定理 9 (Localized Dudley's Theorem)**.** *若 $\mathcal{F}$ 表示函数族 $f : \mathcal{X} \mapsto \mathbb{R}$，则*

$$\mathfrak{R}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left(4\alpha + 12 \int_\alpha^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \right). \tag{48}$$

可以看出，一般的 Dudley's Theorem 是当 $\alpha = 0$ 的一个特殊情况，所以 Localized Dudley's Theorem 可以得到比 Dudley's Theorem 的更紧的上界.

参考资料

[1] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning[M]. MIT press, 2018.

[2] Richard Xu, Machine Learning Notes, `https://github.com/roboticcam/machine-learning-notes/`.