刻画复杂度

Rademacher complexity

Jiaxuan Zou

西安交通大学数学与统计学院

2025 年 5 月 4 日

1. 3.1 拉德马赫复杂度
2. 参考文献

基本符号与假设

- 用 H 表示假设集，与前几章一致。
- h 代表 H 中的一个元素。
- 损失函数 $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ ，具有一般性。
- 对于每个 $h : \mathcal{X} \to \mathcal{Y}$，关联函数 g，将 $(x, y) \in \mathcal{X} \times \mathcal{Y}$ 映射到 $L(h(x), y)$ 。
- G 通常为与 H 相关联的损失函数族。

拉德马赫复杂度的意义

- 拉德马赫复杂度用于刻画函数族的丰富度。
- 通过衡量假设集拟合随机噪声的程度来实现。

复杂度的正式定义

接下来将给出经验拉德马赫复杂度和平均拉德马赫复杂度的正式定义。

经验拉德马赫复杂度定义

设 G 是从 Z 映射到 $[a,b]$ 的函数族，$S=(z_1,\ldots,z_m)$ 是大小为 m 的固定样本，元素来自 Z 。

$$
\begin{aligned}
\hat{\mathcal{R}}_S(G) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{\boldsymbol{\sigma}\cdot\mathbf{g}_S}{m}\right]
\end{aligned}
$$

其中 $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_m)^\top$ ，σ_i 是取值为 $\{-1,+1\}$ 的独立均匀随机变量，即拉德马赫变量，$\mathbf{g}_S=(g(z_1),\ldots,g(z_m))^\top$ 。

- $\boldsymbol{\sigma}\cdot\mathbf{g}_S$ 衡量 $\mathbf{g}_S$ 与随机噪声向量 $\boldsymbol{\sigma}$ 的相关性。
- $\sup\limits_{g\in G}\frac{\boldsymbol{\sigma}\cdot\mathbf{g}_S}{m}$ 衡量函数类 G 在样本 S 上与 $\boldsymbol{\sigma}$ 的相关程度。
- 经验拉德马赫复杂度衡量函数类 G 在样本 S 上与随机噪声的平均相关程度，刻画函数族 G 的丰富度。

拉德马赫复杂度定义

设 D 为抽取样本的分布，对于整数 $m \geq 1$，函数族 G 的拉德马赫复杂度定义为：

$$\mathcal{R}_m(G) = \mathop{\mathbb{E}}_{S \sim D^m}\left[\hat{\mathcal{R}}_S(G)\right]$$

它是经验拉德马赫复杂度 $\hat{\mathcal{R}}_S(G)$ 在根据分布 D 抽取的所有大小为 m 的样本上的期望值。

- 基于此定义，后续将给出基于拉德马赫复杂度的泛化界。

定理 3.1 基于拉德马赫复杂度的泛化界

设 G 是从 Z 映射到 $[0,1]$ 的函数族。对于任意 $\delta>0$，至少以 $1-\delta$ 的概率，对所有 $g\in G$：

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.3}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathcal{R}}_S(G) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3.4}$$

麦克迪尔米德（McDiarmid）不等式

设 $X_1, X_2, \ldots, X_n$ 是相互独立的随机变量，$f : \mathcal{X}^n \to \mathbb{R}$ 是一个函数。如果对于所有的 $i = 1, 2, \ldots, n$ ，以及任意的 $x_1, \ldots, x_n, x_i' \in \mathcal{X}$ ，有：

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

则对于任意 $\epsilon > 0$ ，有：

$$\Pr\left(f(X_1, X_2, \ldots, X_n) - \mathbb{E}[f(X_1, X_2, \ldots, X_n)] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

定理 3.1 证明 - 定义与初步推导

定义 $\Phi(S)=\sup_{g\in G}\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]$ ，其中 $\widehat{\mathbb{E}}_S[g]=\frac{1}{m}\sum_{i=1}^{m}g(z_i)$ 。对于仅一点不同的样本 S 和 S' ，有 $|\Phi(S)-\Phi(S')|\leq 1/m$ 。根据麦克迪尔米德不等式，至少以 $1-\delta/2$ 的概率：

$$\Phi(S)\leq \mathbb{E}_{S}[\Phi(S)]+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

定理 3.1 证明 - 期望界定（上）

对 $\mathbb{E}_S[\Phi(S)]$ 进行界定：

$$
\begin{aligned}
\mathbb{E}_S[\Phi(S)] &= \mathbb{E}_S\left[\sup_{g\in H}\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]\right] \\
&= \mathbb{E}_S\left[\sup_{g\in H}\mathbb{E}_{S'}[\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]]\right] && \text{（利用独立同分布采样性质，}\mathbb{E}[g]=\mathbb{E}_{S'}[\widehat{\mathbb{E}}_{S'}[g]]\text{）} \\
&\le \mathbb{E}_{S,S'}\left[\sup_{g\in H}\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]\right] && \text{（由詹森不等式和上确界函数的凸性得到）} \\
&= \mathbb{E}_{S,S'}\left[\sup_{g\in H}\frac{1}{m}\sum_{i=1}^{m}(g(z_i')-g(z_i))\right] && \text{（根据经验均值}\widehat{\mathbb{E}}_S[g]=\frac{1}{m}\sum_{i=1}^{m}g(z_i)\text{展开）}
\end{aligned}
$$

定理 3.1 证明 - 期望界定（下）

继续对 $\mathbb{E}_{S}[\Phi(S)]$ 进行界定：

$$= \mathbb{E}_{\boldsymbol{\sigma},S,S'}\left[\sup_{g\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z_i')-g(z_i))\right]$$

（引入拉德马赫变量σ_i，其不改变期望值）

$$\leq \mathbb{E}_{\boldsymbol{\sigma},S'}\left[\sup_{g\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i')\right] + \mathbb{E}_{\boldsymbol{\sigma},S}\left[\sup_{g\in H}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(z_i)\right]$$

（由上确界函数的次可加性，即 $\sup(U+V)\leq\sup(U)+\sup(V)$）

$$= 2\mathbb{E}_{\boldsymbol{\sigma},S}\left[\sup_{g\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] = 2\mathcal{R}_m(G)$$

（根据拉德马赫复杂度定义及 σ_i 和 $-\sigma_i$ 分布相同）

定理 3.1 证明 - 最终推导

由前面结果，用 δ 代替 $\delta/2$ ，得到式 (3.3) 的界。又因为改变 S 中一点，$\hat{\mathcal{R}}_S(G)$ 最多改变 $1/m$ ，再用麦克迪尔米德不等式得：

$$\mathcal{R}_m(G) \leq \hat{\mathcal{R}}_S(G) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

结合上述不等式，至少以 $1-\delta$ 的概率得到：

$$\Phi(S) \leq 2\hat{\mathcal{R}}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

即式 (3.4)。

定理 3.1 基于拉德马赫复杂度的泛化界

设 G 是从 Z 映射到 $[0,1]$ 的函数族。对于任意 $\delta > 0$ ，至少以 $1-\delta$ 的概率，对所有 $g \in G$ ：

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.3}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathcal{R}}_S(G) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3.4}$$

引理 3.1

本引理将假设集 H 和二元损失下相关损失函数族 G 的经验拉德马赫复杂度联系起来。

- 设 H 是取值在 $\{-1,+1\}$ 的函数族，$G = \{(x,y) \mapsto 1_{h(x)\neq y} : h \in H\}$ 为 0-1 损失相关损失函数族。
- 样本 $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in \mathcal{X} \times \{-1,+1\}$ ，$S_{\mathcal{X}} = (x_1,\ldots,x_m)$ 是其在 $\mathcal{X}$ 上的投影。

则 G 和 H 的经验拉德马赫复杂度关系为：

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

引理 3.1 证明

对于样本 $S=((x_1,y_1),\dots,(x_m,y_m))\in\mathcal{X}\times\{-1,+1\}$，$G$ 的经验拉德马赫复杂度：

$$
\begin{aligned}
\hat{\mathcal{R}}_S(G) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i 1_{h(x_i)\neq y_i}\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \frac{1-y_ih(x_i)}{2}\right] && (\text{利用}1_{h(x_i)\neq y_i}=\frac{1-y_ih(x_i)}{2}) \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i y_i h(x_i)\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] && (\text{利用}\sigma_i\text{ 和 }-y_i\sigma_i\text{ 分布相同}) \\
&= \frac{1}{2}\hat{\mathcal{R}}_{S_{\mathcal{X}}}(H)
\end{aligned}
$$

取期望可得对于任意 $m\geq 1$，$\mathcal{R}_m(G)=\frac{1}{2}\mathcal{R}_m(H)$，可用于推导二元分类泛化界。

定理 3.2 - 拉德马赫复杂度界 - 二元分类

设 H 是取值在 $\{-1,+1\}$ 的函数族，D 为输入空间 $\mathcal{X}$ 上的分布。对于任意 $\delta>0$，在大小为 m 的样本 S（依 D 抽取）上，至少以 $1-\delta$ 的概率，对任意 $h\in H$：

$$R(h)\leq \hat{R}(h)+\mathcal{R}_m(H)+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$R(h)\leq \hat{R}(h)+\hat{\mathcal{R}}_S(H)+3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

定理 3.2 证明

此定理结果可由定理 3.1 和引理 3.1 直接推导得出。

定理 3.2 泛化界分析

定理 3.2 为二元分类提供两个基于拉德马赫复杂度的泛化界。

- 式 (3.18) 依赖数据，$\hat{\mathcal{R}}_S(H)$ 是特定样本 S 的函数，若能计算它，该界很有价值。
- 计算 $\hat{\mathcal{R}}_S(H)$ 时，利用 σ_i 和 $-\sigma_i$ 分布相同，有：

$$\hat{\mathcal{R}}_S(H) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i h(x_i)\right] = -\mathbb{E}_{\boldsymbol{\sigma}}\left[\inf_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right]$$

- 固定 $\boldsymbol{\sigma}$，计算 $\inf_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)$ 等价于经验风险最小化问题，对某些假设集计算困难。后续将把拉德马赫复杂度与易计算的组合度量联系。

定义 3.3 增长函数

我们将展示如何用增长函数界定拉德马赫复杂度。对于假设集 H，增长函数 $\Pi_H : \mathbb{N} \to \mathbb{N}$ 定义为：

$$\forall m \in \mathbb{N}, \Pi_H(m) = \max_{\{x_1,\dots,x_m\} \subseteq \mathcal{X}} |\{(h(x_1), \dots, h(x_m)) : h \in H\}|$$

$\Pi_H(m)$ 是用 H 中假设对 m 个点分类的不同方式的最大数量，是假设集 H 丰富度的一种度量，且不依赖分布，是纯粹组合性质的。后续将用马斯萨特引理 (Massart’s lemma) 关联拉德马赫复杂度与增长函数。

定理 3.3 马斯萨特引理（Massart's lemma）

设 $A \subseteq \mathbb{R}^m$ 为有限集，$r = \max_{\mathbf{x}\in A} \|\mathbf{x}\|_2$ ，则：

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{r\sqrt{2\log|A|}}{m}$$

这里 σ_i 是取值 $\{-1, +1\}$ 的独立均匀随机变量，$x_1, \ldots, x_m$ 是向量 $\mathbf{x}$ 分量。该引理后续将用于关联拉德马赫复杂度与增长函数。

定理 3.3 马斯萨特引理证明

对于任意 $t>0$：

$$\exp\left(t\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right)\le\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left(\exp\left[t\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right)$$
$$=\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left(\sup_{x\in A}\exp\left[t\sum_{i=1}^{m}\sigma_i x_i\right]\right)\le\sum_{x\in A}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left(\exp\left[t\sum_{i=1}^{m}\sigma_i x_i\right]\right)$$

（利用詹森不等式、整理项及上确界界定）

霍夫丁引理（Hoeffding's lemma）

设 Y 是一个随机变量，满足 $a \leq Y \leq b$（a 和 b 为常数），则对于任意实数 s，有：

$$\mathbb{E}_{Y}[\exp(sY)] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

定理 3.3 马斯萨特引理证明（续）

利用 σ_i 独立性，应用霍夫丁引理并根据 r 定义：

$$\begin{aligned}
\exp\left(t\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leq \sum_{x\in A}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma_i}\left(\exp\left[t\sigma_i x_i\right]\right) \\
&\leq \sum_{x\in A}\prod_{i=1}^{m}\exp\left[\frac{t^2(2x_i)^2}{8}\right] \\
&= \sum_{x\in A}\exp\left[\frac{t^2}{2}\sum_{i=1}^{m}x_i^2\right] \leq \sum_{x\in A}\exp\left[\frac{t^2r^2}{2}\right] = |A|e^{\frac{t^2r^2}{2}}
\end{aligned}$$

两边取对数除以 t 得：

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{\log|A|}{t} + \frac{tr^2}{2}$$

选 $t = \frac{\sqrt{2\log|A|}}{r}$ 使上界最小化，得：

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq r\sqrt{2\log|A|}$$

两边除以 m 即得引理结论，即设 $A\subseteq\mathbb{R}^m$ 为有限集，$r=\max_{\mathbf{x}\in A}\|\mathbf{x}\|_2$，则：

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{r\sqrt{2\log|A|}}{m}$$

推论 3.1

设 G 是取值在 $\{-1,+1\}$ 的函数族，则：

$$\mathcal{R}_m(G) \leq \sqrt{\frac{2\log \Pi_G(m)}{m}}$$

推论 3.1 证明

对于固定样本 $S=(x_1,\ldots,x_m)$，$G_{|S}$ 为 $g\in G$ 对应的函数值向量集合。因 g 取值 $\{-1,+1\}$，向量范数 $\le\sqrt{m}$，应用马斯萨特引理：

$$\mathcal{R}_m(G)=\mathop{\mathbb{E}}_S\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{u\in G_{|S}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i u_i\right]\right]\le\mathop{\mathbb{E}}_S\left[\frac{\sqrt{m}\sqrt{2\log|G_{|S}|}}{m}\right]$$

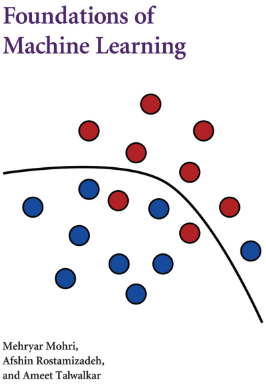
又 $|G_{|S}|$ 由增长函数界定，所以：

$$\mathcal{R}_m(G)\le\mathop{\mathbb{E}}_S\left[\frac{\sqrt{m}\sqrt{2\log\Pi_G(m)}}{m}\right]=\sqrt{\frac{2\log\Pi_G(m)}{m}}$$

推论 3.2 - 增长函数泛化界

设 H 是取值在 $\{-1,+1\}$ 的函数族。对于任意 $\delta > 0$ ，至少以 $1-\delta$ 的概率，对任意 $h \in H$ ：

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2\log \Pi_H(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$



https://github.com/SmartPig-Joe/Foundations-of-Machine-Learning

Thanks!