# 机器学习理论研究导引
# 作业三

陈晟 MG21330006

2022 年 4 月 27 日

## 作业提交注意事项

(1) 本次作业提交截止时间为 **2022/05/03 23:59:59**, 截止时间后不再接收作业, 本次作业记零分;

(2) 作业提交方式：使用此 LaTex 模板书写解答，只需提交编译生成的 pdf 文件，将 pdf 文件上传到以下 ftp 服务器的指定位置：
地址：sftp://210.28.132.67:22，用户名：mlt2022，密码：mltspring2022@nju
文件夹位置：/C:/Users/mlt2022/hw_submissions/hw3_submission/；

(3) pdf 文件命名方式：学号-姓名-作业号-v 版本号, 例 MG1900000-张三-3-v1；如果需要更改已提交的解答，请在截止时间之前提交新版本的解答，并将版本号加一；

(4) 未按照要求提交作业，或 pdf 命名方式不正确，将会被扣除部分作业分数.

# 1 [50pts] Generalization Bound Based-on VC Dimension

在书中，为了推导基于 VC 维的泛化界，先推导出了基于增长函数的泛化界，再利用 VC 维和增长函数之间的关系（定理 3.1）完成证明。在本题中，我们将从基于 Rademacher 复杂度的泛化界出发，推导基于 VC 维的泛化界。若假设空间 $\mathcal{H}$ 的 VC 维为 $d$,

(1) **[30pts]** 试证明对特征空间上的任一分布 $\mathcal{D}$ 和任一正整数 $m$ 都有

$$\mathfrak{R}_m(\mathcal{H}) \leq \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} .$$

(2) **[20pts]** 试利用基于 Rademacher 复杂度的泛化界，结合 (1) 中的结果，推导基于 VC 维的泛化界。

**Proof.**

*(1)*

*VC 维:*$VCdim(\mathcal{H} = max\{m : \prod_{\mathcal{H}}(m) = 2^m\}$

*Let* $\mathcal{H}$ *be a hypothesis set with* $VCdim(\mathcal{H})$*.Then for all* $m \geq d$,

$\prod_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \binom{m}{i} (\frac{m}{d})^{d-i}$

$\leq \sum_{m}^{i=0} \binom{m}{i} (\frac{m}{d})^{d-i} = (\frac{m}{d})^d \sum_{i=0}^{m} \binom{m}{i} (\frac{d}{m})^i$

$= \left(\frac{m}{d}\right)^d (1 + \frac{d}{m})^m \leq (\frac{m}{d})^d e^d$

$\therefore \prod_{\mathcal{H}}(m) \leq (\frac{em}{d})^d = O(m^d)$

对于 $D = \{x_1, x_2 ... x_m\}$，$\mathcal{H}_{|\mathcal{D}}$ 为假设空间 $\mathcal{H}$ 在 $D$ 上的限制。由于 $h \in \mathcal{H}$ 的值域为 $\{-1, +1\}$，可知 $\mathcal{H}_{|\mathcal{D}}$ 中的元素为模长 $\sqrt{m}$ 的向量。因此由 *ppt* 中定理可得:

$\because \mathcal{R}_m(\mathcal{H}) = E_D[E_\sigma[sup_{u \in \mathcal{H}|D} \frac{1}{m} \sum_{m}^{i=1} \sigma_i u_i]]$

$\leq E_D[\frac{\sqrt{m}\sqrt{2ln|\mathcal{H}_{|D}|}}{m}]$

由 $|H_{|D}| \leq \prod_{\mathcal{H}}(m)$ 有

$\mathcal{R}_m(\mathcal{H}) \leq E_D[\frac{\sqrt{m}\sqrt{2ln\prod_{\mathcal{H}}(m)}}{m}] = \sqrt{\frac{2ln\prod_{\mathcal{H}}(m)}{m}}$

带入 $\prod_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ 即证: $\mathfrak{R}_m(\mathcal{H}) \leq \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}}$ .

*(2)*

设 $\mathcal{H}$ 为一族从 $\{-1, +1\}$ 中取值且 *VC* 维等于 $d$ 的函数。那么，对于任意的 $\delta > 0$，下面的不等式至少有 $1 - \delta$ 的概率对假设 $h \in \mathcal{H}$ 成立

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

因此，这种情况下的泛化界满足 $\mathcal{R}_m(h) \leq \hat{\mathcal{R}_m}(h) + \mathcal{O}(\sqrt{\frac{log(\frac{m}{d})}{\frac{m}{d}}})$

它说明了 $\frac{m}{d}$ 对于泛化的重要性。这个推论同样符合奥卡姆剃刀原则

(由 *Sauer* 引理，我们还能得到以下的 *VC* ，不通过 *Rademacher* 复杂度得到：$\mathcal{R}_m(h) \leq \hat{\mathcal{R}_m}(h) + \sqrt{\frac{8dlog\frac{2em}{d}+8log\frac{4}{\delta}}{m}}$)

## 2 [50pts] Generalization Bound and Covering Numbers

设 $\mathcal{H}$ 为一假设空间, 假设的定义域为 $\mathcal{X}$, 值域为 $\mathcal{Y}\subset\mathbb{R}$. $\forall\epsilon>0$, 可如下定义覆盖数 (Covering Number):

$$\mathcal{N}(\mathcal{H},\epsilon)=\min\left\{k\in\mathbb{N}|\exists\{h_1,\cdots,h_k\}\subset\mathcal{H},\ \text{s.t. }\forall h\in\mathcal{H},\exists i\in[k],\|h-h_i\|_\infty\leqslant\epsilon\right\},\tag{2.1}$$

其中 $\|h-h_i\|_\infty=\max_{x\in\mathcal{X}}|h(x)-h_i(x)|$. 覆盖数可以衡量一个假设空间的复杂度: 覆盖数越大, 意味着这一假设空间越复杂. 本题利用覆盖数证明了平方损失下的一个泛化界, 该结论也可以说明覆盖数可以衡量假设空间的复杂度. 令 $\mathcal{D}$ 为 $\mathcal{X}\times\mathcal{Y}$ 上的一个分布, 且有标记样本根据这一分布采样得到. 定义 $h\in\mathcal{H}$ 的泛化误差为

$$R(h)=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h(x)-y)^2\right].\tag{2.2}$$

训练集 $S=\{(x_1,y_1),\cdots,(x_m,y_m)\}$ 上的经验误差为

$$\hat{R}_S(h)=\frac{1}{m}\sum_{i=1}^m(h(x_i)-y_i)^2.\tag{2.3}$$

设 $\mathcal{H}$ 是有界的, 即 $\exists M>0$, 使得 $\forall(x,y)\in\mathcal{X}\times\mathcal{Y},h\in\mathcal{H},|h(x)-y|\leqslant M$.

(1) **[15pts]** 令 $L_S=R(h)-\hat{R}_S(h)$, 试证明 $\forall h_1,h_2\in\mathcal{H},S$,

$$|L_S(h_1)-L_S(h_2)|\leqslant 4M\|h_1-h_2\|_\infty.\tag{2.4}$$

(2) **[15pts]** 设 $\mathcal{H}$ 可以被 $k$ 个子集 $\mathcal{B}_1,\cdots,\mathcal{B}_k$ 覆盖, 即 $\mathcal{H}=\mathcal{B}_1\cup\cdots\cup\mathcal{B}_k$. 试证明 $\forall\epsilon>0$,

$$\Pr_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_S(h)|\geqslant\epsilon\right]\leqslant\sum_{i=1}^k\Pr_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right].\tag{2.5}$$

(3) **[20pts]** 令 $k=\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)$, $\mathcal{B}_1,\cdots,\mathcal{B}_k$ 为 $\mathcal{H}$ 的覆盖, 其中 $\forall i\in[k]$, $\mathcal{B}_i$ 的圆心为 $h_i$, 半径为 $\frac{\epsilon}{8M}$. 试证明

$$\Pr_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right]\leqslant\Pr_{S\sim\mathcal{D}^m}\left[|L_S(h_i)|\geqslant\frac{\epsilon}{2}\right].\tag{2.6}$$

并利用 Hoeffding 不等式证明

$$\Pr_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]\leqslant\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)2\mathrm{e}^{-\frac{m\epsilon^2}{2M^4}}.\tag{2.7}$$

**Proof.**

***(1)***

由于 $R(h)=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h(x)-y)^2\right]$. 训练集 $S=\{(x_1,y_1),\cdots,(x_m,y_m)\}$ 上的经验误差为 $\hat{R}_S(h)=\frac{1}{m}\sum_{i=1}^m(h(x_i)-y_i)^2$

$\therefore L_S=R(h)-\hat{R}_S(h)=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h(x)-y)^2\right]-\frac{1}{m}\sum_{i=1}^m(h(x_i)-y_i)^2$

$\forall h_1,h_2\in\mathcal{H},S,|L_S(h_1)-L_S(h_2)|=$

$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h_1(x)-y)^2\right]-\frac{1}{m}\sum_{i=1}^m(h_1(x_i)-y_i)^2-\left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h_2(x)-y)^2\right]-\frac{1}{m}\sum_{i=1}^m(h_2(x_i)-y_i)^2\right)\right|$

$\because \forall h_1, h_2 \in \mathcal{H}$

$\therefore \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h_1(x)-y)^2\right] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h_2(x)-y)^2\right]$

$\frac{1}{m}\sum_{i=1}^{m}(h_2(x_i)-y_i)^2 - \frac{1}{m}\sum_{i=1}^{m}(h_1(x_i)-y_i)^2 =$

$\frac{1}{m}\sum_{i=1}^{m}\left((h_2(x_i)-y_i)^2-(h_1(x_i)-y_i)^2\right) =$

$\frac{1}{m}\sum_{i=1}^{m}\left((h_2(x_i)+h_1(x_i)-2y_i)(h_2(x)-h_1(x))\right)$

$\because \mathcal{H}$ 是有界的，即 $\exists M>0$，使得 $\forall(x,y)\in\mathcal{X}\times\mathcal{Y}, h\in\mathcal{H}, |h(x)-y|\leqslant My$

$\therefore |(h_1(x_i)+h_2(x_i)-2y_i)^2| \le |(2*max(h_1(x_i),h_2(x_i))-2y_i)^2| \le 4M$

并且显然,$|h_1(x_i)-h_2(x_i)| \le \|h_1-h_2\|_\infty = \max_{x\in\mathcal{X}}|h_1(x)-h_2(x)|$

$\therefore |L_S(h_1)-L_S(h_2)| = |\frac{1}{m}\sum_{i=1}^{m}(h_2(x_i)-y_i)^2 - \frac{1}{m}\sum_{i=1}^{m}(h_1(x_i)-y_i)^2| =$

$|\frac{1}{m}\sum_{i=1}^{m}\left((h_2(x_i)+h_1(x_i)-2y_i)(h_2(x)-h_1(x))\right)| =$

$\frac{1}{m}\sum_{i=1}^{m}|(h_2(x_i)+h_1(x_i)-2y_i)(h_2(x)-h_1(x))| \le \frac{1}{m}\sum_{i=1}^{m}(4M\|h_1-h_2\|_\infty) = 4M\|h_1-h_2\|_\infty$

即证明 $\forall h_1,h_2\in\mathcal{H}, S, |L_S(h_1)-L_S(h_2)| \leqslant 4M\|h_1-h_2\|_\infty$

*(2)*

当 *k=1* 时，显然有 $\mathcal{H}=B_1$

所以 $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}|L_S(h)|\geqslant\epsilon\right] = \sum_{i=1}^{1}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right]$ 题目要求显然成立

假设 *k=n* 时，对任意 $\mathcal{H}$ $n$ $B_1,B_2...B_n$，都有 $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}|L_S(h)|\geqslant\epsilon\right] = \sum_{i=1}^{n}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right]$ 成立

则 *k=n+1* 时，会新增加一个 $B_{n+1}$，由上可知该 $B_{n+1}$ 可以是 *k=n* 的某一种子集划分情况下，某一个子集 $B_p$ 的子集，因此可以有

$\sum_{i=1}^{n}+1\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right] = \sum_{i=1}^{n}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right] + \underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_{n+1}}|L_S(h)|\geqslant\epsilon\right]$

$\therefore \underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}|L_S(h)|\geqslant\epsilon\right] \le \sum_{i=1}^{n+1}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right]$ 成立

由数学归纳法可知，对任意 *k(k>0)*，有

$\mathcal{H}=\mathcal{B}_1\cup\cdots\cup\mathcal{B}_k$. $\forall\epsilon>0$, 有 $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}|L_S(h)|\geqslant\epsilon\right] \leqslant \sum_{i=1}^{k}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right]$.

*(3)*

$\because k=\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)$

$\therefore k=\mathcal{N}(\mathcal{H},\frac{\epsilon}{8M}) = \min\left\{k\in\mathbb{N}|\exists\{h_1,\cdots,h_k\}\subset\mathcal{H},\ s.t.\ \forall h\in\mathcal{H},\exists i\in[k],\|h-h_i\|_\infty\leqslant\frac{\epsilon}{8M}\right\}$

由于 $\forall i\in[k]$, $\mathcal{B}_i$ 的圆心为 $h_i$，半径为 $\frac{\epsilon}{8M}$

$\therefore \underset{S\sim\mathcal{D}^m}{Pr}\left[|L_S(h_i)|\geqslant\frac{\epsilon}{2}\right] = R(h_i)-\hat{R}_S(h_i) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(h_i(x)-y)^2\right] - \frac{1}{m}\sum_{j=1}^{m}(h_i(x_j)-y_j)^2$

根据 *ppt* 对引理 *4.1* 证明过程中的 $Pr(sup_{h\in\mathcal{H}}|\hat{E}_D(h)-\hat{E}_{D'}(h)|\ge\frac{1}{2}\epsilon) \ge \frac{1}{2}Pr(sup_{h\in\mathcal{H}}|E(h)-\hat{E}_D(h)|>\epsilon)$

其中 $D$ 和 $D'$ 都是从 $D$ 独立同分采样的训练集

$\underset{S\sim\mathcal{D}^m}{Pr}\left[|L_S(h_i)|\geqslant\frac{\epsilon}{2}\right] \ge E_{B_i\in Q}[E_{S_i\sim D^m}[\mathcal{I}(sup_{h\in\mathcal{H}}(|\hat{E}_{B_i}(h)-\hat{E}_{D'}(h)|\ge\frac{1}{2}\epsilon)]]$

$\ge E_{h_i\sim B_i}[\mathcal{I}(|\hat{E}_D(h_0)-E(h_0)|-||\hat{E}_{h_i}(h_0)-E(h_0)|\ge\frac{1}{2}\epsilon)] \ge 1-Pr(|\hat{E}_{h_i}(h_0)-E(h_0)|>\frac{1}{2}\epsilon)$

由 *Chebyshev* 不等式 *(1.21)* 可得

$Pr(|\hat{E}_{h_i}(h_0)-E(h_0)|>\frac{1}{2}\epsilon) \le \frac{4(1-E(h_0))E(h_0)}{\epsilon^2 m} \le \frac{1}{\epsilon^2 m}$

由于 $\forall i\in[k]$, $\mathcal{B}_i$ 的圆心为 $h_i$，半径为 $\frac{\epsilon}{8M}$，于是可得 $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right] \leqslant \underset{S\sim\mathcal{D}^m}{Pr}\left[|L_S(h_i)|\geqslant\frac{\epsilon}{2}\right]$

*b.* 利用 *Hoeffding* 不等式证明 $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right] \leqslant \mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)2\mathrm{e}^{-\frac{m\epsilon^2}{2M^4}}$

由 *(2)* $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}|L_S(h)|\geqslant\epsilon\right] \leqslant \sum_{i=1}^{k}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}|L_S(h)|\geqslant\epsilon\right]$

且 $\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]=\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}\left|L_S(h)\right|\geqslant\epsilon\right]$

$\therefore\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]\leq\sum_{i=1}^{k}\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]$

$\leq\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]$

显然，由 *Hoeffding* 不等式，有：$\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]\leq 2\mathrm{e}^{-\frac{m\epsilon^2}{2M^4}}$

$\therefore\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{H}}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]\leq\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)\underset{S\sim\mathcal{D}^m}{Pr}\left[\sup_{h\in\mathcal{B}_i}\left|R(h)-\hat{R}_S(h)\right|\geqslant\epsilon\right]\leq\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{8M}\right)2\mathrm{e}^{-\frac{m\epsilon^2}{2M^4}}$