Tsinghua University
机器学习基础
Fall 2017

第 3 次作业

赵丰　2017310711　November 18, 2017

- **Acknowledgments:** This coursework refers to textbook.
- **Collaborators:**
  - 2.2 was solved with the help from `https://en.wikipedia.org/wiki/Sauer\%E2\%80\%93Shelah_lemma`.
  - 2.3 was solved with the help from `http://www.cs.nyu.edu/~mohri/ml07/sol2.pdf`
  - Discussion with the teaching assistant helped me finishing 3(b).

I use `enumerate` to generate answers for each question:

1. 1.1. (a)

$$\hat{\mathfrak{R}}_S(cG+d) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{\boldsymbol{\sigma}\cdot(c\boldsymbol{g}_S+d)}{m}\right] \tag{1}$$

因为 $\forall g \in G$,

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{d}{m}\sum_{i=1}^{m}\sigma_i\right] = 0 \tag{2}$$

所以

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{\boldsymbol{\sigma}\cdot(c\boldsymbol{g}_S+d)}{m}\right] = |c|\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}(c)\sigma_i g(z_i)\right] \tag{3}$$

因为 $\operatorname{sign}(c)\sigma_i$ 与 σ_i 具有相同的分布，所以

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{\boldsymbol{\sigma}\cdot(c\boldsymbol{g}_S+d)}{m}\right] = |c|\hat{\mathfrak{R}}_S(G) \tag{4}$$

从而推出：

$$\hat{\mathfrak{R}}_S(cG+d) = |c|\hat{\mathfrak{R}}_S(G) \tag{5}$$

(b)

$$\hat{\mathfrak{R}}_S(\operatorname{conv}(G)) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_i\in G}\frac{\boldsymbol{\sigma}\cdot(\sum_{i=1}^{n}\alpha_i\boldsymbol{g}_i)}{m}\right] \tag{6}$$

其中 $\boldsymbol{g}_i = (g_i(z_1),\ldots,g_i(z_m))^T$
一方面：

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_i\in G}\frac{\boldsymbol{\sigma}\cdot(\sum_{i=1}^{n}\alpha_i\boldsymbol{g}_i)}{m}\right] &\le \sum_{i=1}^{n}\alpha_i\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_i\in G}\frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_i}{m}\right]\\ &= \sum_{i=1}^{n}\alpha_i\hat{\mathfrak{R}}_S(G)\\ &= \hat{\mathfrak{R}}_S(G)\end{aligned}$$

另一方面，取 $n=1,\alpha_1=1$ 有

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_i\in G}\frac{\boldsymbol{\sigma}\cdot(\sum_{i=1}^{n}\alpha_i\boldsymbol{g}_i)}{m}\right]\geq\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_1\in G}\frac{\boldsymbol{\sigma}\cdot(\alpha_1\boldsymbol{g}_1)}{m}\right]=\hat{\mathfrak{R}}_S(G) \tag{7}$$

所以推出

$$\hat{\mathfrak{R}}_S(\mathrm{conv}(G))=\hat{\mathfrak{R}}_S(G) \tag{8}$$

(c) 可以证明 $\sup_{a\in A,b\in B}(a+b)=\sup_{a\in A}a+\sup_{b\in B}b$, 所以

$$\begin{aligned}\hat{\mathfrak{R}}_S(G_1+G_2)=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{g_1\in G_1\\g_2\in G_2}}\frac{\boldsymbol{\sigma}\cdot(\boldsymbol{g}_1+\boldsymbol{g}_2)}{m}\right]\\=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_1\in G_1}\frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_1}{m}\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_2\in G_2}\frac{\boldsymbol{\sigma}\cdot\boldsymbol{g}_2}{m}\right]\\=&\hat{\mathfrak{R}}_S(G_1)+\hat{\mathfrak{R}}_S(G_2)\end{aligned}$$

1.2.

$$\begin{aligned}\hat{\mathfrak{R}}_S(G)=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{\sum_{i=1}^{m}\sigma_i 1_{h(x_i)\neq y_i}}{m}\right]\\=&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{\sum_{i=1}^{m}\sigma_i\frac{1-h(x_i)y_i}{2}}{m}\right]\end{aligned} \tag{9}$$

因为 $|y_i|=1$, 由 **P1.1(a)** 的结论

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{\sum_{i=1}^{m}\sigma_i\frac{1-h(x_i)y_i}{2}}{m}\right]=&\frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\frac{\sum_{i=1}^{m}\sigma_i h(x_i)}{m}\right]\\=&\frac{1}{2}\hat{\mathfrak{R}}_{S_{\mathcal{X}}}(H)\end{aligned} \tag{10}$$

因此推出

$$\hat{\mathfrak{R}}_S(G)=\frac{1}{2}\hat{\mathfrak{R}}_{S_{\mathcal{X}}}(H) \tag{11}$$

2. (a) 对于直线上 m 个不同的点，考虑含 $k(1\leq k\leq m)$ 个正类有 $m+1-k$ 种分法，外加全为负类的一种分法，有

$$\Pi_H(m)=1+\sum_{k=1}^{m}(m+1-k)=1+\frac{m(m+1)}{2} \tag{12}$$

根据 Sauer 引理

$$\Pi_H(m)\leq\binom{m}{0}+\binom{m}{1}+\binom{m}{2}=1+\frac{m(m+1)}{2} \tag{13}$$

因此对于数轴上区间作为假设集的例子，引理给出的上界是紧的。

(b) 从 $\{1,2,\ldots,m\}$ 中选不超过 d 个数共有 $K=\sum_{k=0}^{d}\binom{m}{k}$ 中取法，对应于第 i 种取法，$h_i(x)$ 是这样一个向量，将取出的数对应的分量位置置 1，其余位置置 0。注意 h_i 虽然是从 $\mathcal{X}$ 到 $\{0,1\}^m$ 空间的一个映射但却不依赖于 $x\in\mathcal{X}$。将这 K 个映射构成我们考虑的函数空间 H，由定义可知 H 的增长函数为 K。当 $m=d$ 时 $K=2^d$，由 VC 维的定义可知 H 的 VC 维为 d。所以我们构造出了这样一个假设集合使得 Sauer 引理取到了等号。

(c) 只需对 $\mathbb{R}^n$ 中证明对任意 $n+3$ 个点只用一个闭球均不能完全分离两类。反设 $\exists \boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n+3}$，有 $\forall y_i\in\{0,1\}, 1\le i\le n+3$, 均 $\exists r,\boldsymbol{x}^*$, 使得 $y_i(||\boldsymbol{x}_i-\boldsymbol{x}^*||_2^2-r^2)>0, 1\le i\le n+3$

下面考虑 $\mathbb{R}^{n+1}$ 中的点集,$\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_{n+3}$, 其中 $\tilde{\boldsymbol{x}}_i=\binom{\boldsymbol{x}_i}{||\boldsymbol{x}_i||_2^2}$, 对于这 $n+3$ 个 $\mathbb{R}^{n+1}$ 中的点，由前面假设：

$\forall y_i\in\{0,1\}, 1\le i\le n+3$, 构造 $\boldsymbol{a}=\binom{-2\boldsymbol{x}^*}{1}, b=||\boldsymbol{x}^*||_2^2-r^2$，则有 $y_i(\boldsymbol{a}\cdot\tilde{\boldsymbol{x}}_i+b)>0, 1\le i\le n+3$，即用超平面在 $n+1$ 维空间中对任意 $n+3$ 个点都可以完全分离。这与超平面的 VC 维等于考虑的空间维数加 2 矛盾。

因此用 $\mathbb{R}^n$ 中的闭球分类 VC 维至多为 $n+2$。

3. (a) 对于给定的样本，由 $\hat{h}$ 定义我们有

$$\mathbb{E}[\widehat{\mathrm{err}}(\hat{h})]\le\mathbb{E}[\widehat{\mathrm{err}}(h^*)]=\mathrm{err}(h^*) \tag{14}$$

其中最后一式利用了对于给定的假设，经验误差的均值等于泛化误差。另一方面，

$$\mathbb{E}[\mathrm{err}(\hat{h})]\ge\mathbb{E}[\mathrm{err}(h^*)]=\mathrm{err}(h^*) \tag{15}$$

(b) 设 $\tilde{S}$ 与 S 只有一个样本点不同，设其指标为 (x_i',y_i'), 记函数

$$\Phi(S)=\widehat{\mathrm{err}}(\hat{h})-\mathbb{E}[\widehat{\mathrm{err}}(\hat{h})] \tag{16}$$

$$\Phi(S)-\Phi(S')=\min_{h\in\mathcal{H},S}\widehat{\mathrm{err}}(h)-\min_{h\in\mathcal{H},S'}\widehat{\mathrm{err}}(h) \tag{17}$$

$$\ge\min_{h\in\mathcal{H}}\frac{1}{m}(1_{h(x_i)\neq y_i}-1_{h(x_i')\neq y_i'}) \tag{18}$$

$$\ge\frac{-1}{m} \tag{19}$$

同理

$$\Phi(S')-\Phi(S)\ge\frac{-1}{m} \tag{20}$$

从而有

$$|\Phi(S)-\Phi(S')|\le\frac{1}{m} \tag{21}$$

由 McDiarmid 不等式，有 $1-\delta$ 的概率使得：

$$|\Phi(S)-\Phi(S')|\le\sqrt{\frac{\ln(2/\delta)}{2m}} \tag{22}$$

由于 δ 是小量,$\frac{\ln(2)}{m}$ 相比 $\frac{\ln(1/\delta)}{m}$ 为小量，因此

$$|\Phi(S) - \Phi(S')| = O(\frac{\sqrt{\ln(1/\delta)}}{m}) \tag{23}$$

常数为 $\frac{\sqrt{2}}{2}$

(c) (a) 中的结果说明，极小化泛化误差得到的最优估计器的统计平均性能要优于极小化经验误差的估计器（第二个不等式），但针对两种情况下的最优估计器，泛化误差要大于经验误差的平均（第一个不等式）。(b) 中的结果说明，当样本足够多时，经验误差以大概率集中在经验误差的平均值附近。