Homework 1

Fondations of Machine Learning

Xiang Pan (xp2030)

February 22, 2022

A Concentration of bound

1

1.1 (a)

According to the Hoeffding's Inequality, we have

$$\Pr[R(h) - \widehat{R}(h) \geq \epsilon] \leq e^{-2m\epsilon^2} \tag{1}$$

We use $\epsilon = \frac{1}{2}$ here, thus we have

$$\Pr[R(h) - \widehat{R}(h) \geq \frac{1}{2}] \leq e^{-2m\frac{1}{4}} = e^{-\frac{1}{2}m} < e^{-\frac{1}{3}m} \tag{2}$$

So we will not have h,

$$R(h) - \widehat{R}(h) \geq \frac{1}{2}, \tag{3}$$

with probability at least $e^{-\frac{1}{3}m}$.

1.2 (b)

The algorithm is that for all the samples in the training dataset, we assign label 1 to them, for any other samples, we assign label 0.

$$R(h) = \Pr_{x\sim D}[h(x) \neq c(x)] = \underset{x\sim D}{\mathrm{E}}\left[1_{h(x)\neq c(x)}\right]. \tag{4}$$

We have $\widehat{R}_S(h_S) = 0$, and $R(h_S) = 1$.

According to the definition of $R(h)$, the training dataset samples is limited/finity, for the expectation, the final error is 1. So we have,

$$R(h_S) - \widehat{R}_S(h_S) = 1 \tag{5}$$

1.3 (c)

The (a) part is about the probability and the Hypothesis h is not conditional on the data samples S. But for given Hypothesis and given data samples, we can design a algorithm to achieve the (b) part.

PAC-Bayesian bound

2

2.1 (a)

The Rademacher complexity bound is,

$$\mathrm{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g\left(z_i\right) + 2\Re_m(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{6}$$

We already know that $L(h, z) = l(h(x), y)$, L is a family of functions mapping from $\mathbb{R} \times y \rightarrow [0,1]$. Directly apply the Rademacher complexity bound,

$$\underset{\substack{h \sim Q \\ z \sim D}}{\mathbb{E}}[L(h, z)] \leq \underset{h \sim Q}{\mathbb{E}}\left[\frac{1}{m} \sum_{i=1}^{m} L\left(h, z_i\right)\right] + 2\Re_m\left(g_\mu\right) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{7}$$

2.2 (b)

Follow the result from (7), and the inequality from the statement,

$$\Re_m\left(\mathcal{G}_\mu\right) \leq \sqrt{\frac{2\mu}{m}}. \tag{8}$$

We have, with probability at least $1-\delta$, and $Q \in \mathcal{G}_\mu$

$$\underset{\substack{h \sim Q \\ z \sim D}}{\mathbb{E}}[L(h, z)] \leq \underset{h \sim Q}{\mathbb{E}}\left[\frac{1}{m} \sum_{i=1}^{m} L\left(h, z_i\right)\right] + 2\sqrt{\frac{2\mu}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{9}$$

Comsidering some a,

$$\Delta(\mathcal{H}) = \{Q \in \Delta(\mathcal{H}) : \mathrm{D}(Q \| P) \leq a\} \cup (\bigcup_{j=1}^{\infty} \left\{Q \in \Delta(\mathcal{H}) : a 2^{j-1} < D(Q \| P) \leq a 2^j\right\}) \tag{10}$$

For a = 1,

$$\Delta(\mathcal{H}) = \{Q \in \Delta(\mathcal{H}) : \mathrm{D}(Q\|P) \leq 1\} \cup (\bigcup_{j=1}^{\infty} \{Q \in \Delta(\mathcal{H}) : 2^{j-1} < D(Q\|P) \leq 2^j\}) \tag{11}$$

We denote $\Delta(H) = \Delta(H_0) \cup \Delta(H_1) \cup ...\Delta(H_\infty)$,
where $\Delta(H_0) = \{Q \in \Delta(\mathcal{H}) : \mathrm{D}(Q\|P) \leq 1\}$
and $\Delta(H_j)$ for $j > 0$ denote $Q \in \Delta(\mathcal{H}) : 2^{j-1} < D(Q\|P) \leq 2^j$.
The definition can be combined as,

$$\Delta(H_j) = \begin{cases} \{Q \in \Delta(\mathcal{H}) : 0 \leq D(Q\|P) \leq 1\} & j = 0 \\ \{Q \in \Delta(\mathcal{H}) : 2^{j-1} < D(Q\|P) \leq 2^j\} & j \geq 1 \end{cases} \tag{12}$$

Applying (9), with probability at most $\delta_j = \frac{1}{(2^{j+1})}\delta$, and $Q \in \mathcal{G}_{2^j}$, we have,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q \\ z\sim D}}[L(h,z)] > \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h, z_i)\right] + 2\sqrt{\frac{2(2^j)}{m}} + \sqrt{\frac{\log\frac{1}{\delta_j}}{2m}} \tag{13}$$

Note that $\Delta(H_j) \subseteq \mathcal{G}_{2^j}$, then for all $Q \in \Delta(H_j)$, (13) still holds. And we name this inequality (13) as,

$$LHS > RHS_j. \tag{14}$$

We want to prove (partial result, not final target): with at least probability 1-δ_j, we have,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q \\ z\sim D}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h, z_i)\right] + (4 + \frac{1}{\sqrt{e}})\sqrt{\frac{\max(1, D(Q\|P))}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{15}$$

We can convert it to: with at most probability δ_j, we have,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q \\ z\sim D}}[L(h,z)] > \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h, z_i)\right] + (4 + \frac{1}{\sqrt{e}})\sqrt{\frac{\max(1, D(Q\|P))}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{16}$$

We name it as

$$LHS > RHS'. \tag{17}$$

We have with at most probability δ_j, $LHS > RHS_j$ according to (13), if we can prove that $RHS' > RHS_j$, thus we have with at most probability δ_j, $LHS > RHS'$.

We want to prove currently is $RHS' > RHS_j$**,** when $Q \in \Delta(H_j)$.

$$\iff \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m}L\left(h,z_i\right)\right]+(4+\frac{1}{\sqrt{e}})\sqrt{\frac{\max\left(1,D(Q||P)\right)}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$
$$\geq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m}L\left(h,z_i\right)\right]+2\sqrt{\frac{2\mu}{m}}+\sqrt{\frac{\log\frac{1}{\delta_j}}{2m}} \tag{18}$$

$$\iff (4+\frac{1}{\sqrt{e}})\sqrt{\frac{\max\left(1,D(Q||P)\right)}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\geq(2\sqrt{\frac{2^{j+1}}{m}}+\sqrt{\frac{\log\frac{1}{\delta_j}}{2m}}) \tag{19}$$

$$\iff (4+\frac{1}{\sqrt{e}})\sqrt{\frac{\max\left(1,2^j\right)}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\geq(2\sqrt{\frac{2^{j+1}}{m}}+\sqrt{\frac{\log\frac{1}{\delta_j}}{2m}}), \tag{20}$$

$$\iff (4+\frac{1}{\sqrt{e}})\sqrt{\frac{2^j}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\geq(2\sqrt{\frac{2^{j+1}}{m}}+\sqrt{\frac{\log\frac{1}{\delta_j}}{2m}}) \tag{21}$$

$$\iff (4+\frac{1}{\sqrt{e}})\sqrt{2^j}+\sqrt{\frac{\log\frac{1}{\delta}}{2}}\geq(2\sqrt{2^{j+1}})+\sqrt{\frac{\log\frac{2^{j+1}}{\delta}}{2}}) \tag{22}$$

$$\iff (4+\frac{1}{\sqrt{e}})\sqrt{2^j}+\sqrt{\frac{\log\frac{1}{\delta}}{2}}\geq(2\sqrt{2^{j+1}})+\sqrt{\frac{(j+1)+\log\frac{1}{\delta}}{2}}) \tag{23}$$

$$\iff (\frac{1}{\sqrt{e}})\sqrt{2^j}+\sqrt{\frac{\log\frac{1}{\delta}}{2}}\geq\sqrt{\frac{(j+1)+\log\frac{1}{\delta}}{2}}) \tag{24}$$

where (20) is from $Q\in\Delta(H_j)$.
We have

$$\frac{t}{e}\geq\frac{\log(2t)}{2} \tag{25}$$

With $t=2^i$, we have

$$\frac{2^j}{e}\geq\frac{\log(2^{j+1})}{2} \tag{26}$$
$$\geq\frac{(j+1)}{2} \tag{27}$$

$$\sqrt{\frac{2^j}{e}}+\sqrt{\frac{\log\frac{1}{\delta}}{2}} \geq \sqrt{\frac{2^j}{e}+\frac{\log\frac{1}{\delta}}{2}} \geq \sqrt{\frac{(j+1)}{2}+\frac{\log\frac{1}{\delta}}{2}} \tag{28}$$

We have proved (28), thus we have proved the (16).

We name the inequality (15) holds as event $E_j = 1$, and the inequality does not hold as $E_j = 0$, the final inequality (39) holds as event $E = 1$, the final inequality does not hold as $E = 0$.

$$
\begin{aligned}
&\Pr\left[E=0\right] && (29)\\
&= \Pr\left[\bigcup_{j\in[0,\infty]} Q\in\Delta(H_j) : E_j = 0\right] && (30)\\
&\leq \sum_{j\in[0,\infty]} \Pr\left[Q\in\Delta(H_j) : E_j = 0\right] \qquad \text{(Union Bound)} && (31)\\
&= \sum_{j\in[0,\infty]} \Pr\left[Q\in\Delta(H_j) : LHS > RHS'\right] && (32)\\
&\leq \sum_{j\in[0,\infty]} \Pr\left[Q\in\Delta(H_j) : LHS > RHS_j\right] \qquad (RHS' > RHS_i) && (33)\\
&= \delta_0+\delta_1+...\delta_\infty \qquad \text{(Probability from (13))} && (34)\\
&= (1/2+1/4+1/8+...+1/2^\infty)\delta && (35)\\
&= (\sum_{j=0}^{\infty}\frac{1}{2^j})\delta && (36)\\
&= \delta && (37)
\end{aligned}
$$

With at most probability δ, we have proved,.

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim\mathcal{D}}}[L(h,z)] > \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L\left(h,z_i\right)\right]+\left(4+\frac{1}{\sqrt{e}}\right)\sqrt{\frac{\max\{\mathrm{D}(Q\|P),1\}}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{38}$$

With at least probability 1-δ, we have proved,

$$\mathop{\mathbb{E}}_{\substack{h\sim Q\\ z\sim D}}[L(h,z)] \leq \mathop{\mathbb{E}}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L\left(h,z_i\right)\right]+\left(4+\frac{1}{\sqrt{e}}\right)\sqrt{\frac{\max\{\mathrm{D}(Q\|P),1\}}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{39}$$

Rademacher complexity

3

3.1 (a)

$$\widehat{\Re}_{\mathcal{S}}(\mathcal{H}) = \underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h\left(\mathbf{x}_i\right)\right] \tag{40}$$

$$= \frac{1}{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}} \mathbf{w}\sum_{i=1}^{m}\sigma_i \mathbf{x_i}\right] \tag{41}$$

$$= \frac{\Lambda}{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{x}_i\right\|_{\infty}\right] \quad \text{(by the definition of dual norm)} \tag{42}$$

$$= \frac{\Lambda}{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\max_{j\in d}\left|\sum_{i=1}^{m}\sigma_i x_{ij}\right|\right] \tag{43}$$

$$= \frac{\Lambda}{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\max_{j\in d}\max_{s\in\{-1,+1\}} s\sum_{i=1}^{m}\sigma_i x_{ij}\right] \tag{44}$$

$$= \frac{\Lambda}{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\sup_{\mathbf{z}\in A}\sum_{i=1}^{m}\sigma_i z_i\right] \tag{45}$$

Where A is the column vector set in M.

$$A := \left\{s\left(x_{1j},\ldots,x_{mj}\right)^{\top} : j\in[d], s\in\{-1,+1\}\right\} \tag{46}$$

For any $z\in A$, we have,

$$\|\mathbf{z}\|_2 \le \sup_{\mathbf{z}\in A}\|\mathbf{z}\|_2 = \left\|\mathbf{X}^{\top}\right\|_{2,\infty} \tag{47}$$

Using Massart's lemma, A contains at most 2N elements, we have,

$$\widehat{\Re}_{S}(\mathcal{H}) \le \frac{\Lambda}{m}\sqrt{2\log(2N)}\left\|X^{\top}\right\|_{2,\infty}. \tag{48}$$

3.2 (b)

Case 1: $p \leq 2$ By directly applying Jensen's inequality, we get, $C_1 = C_2 = 1$,

$$\mathop{\mathbb{E}}_{\sigma}\left[\left|\sum_{i=1}^{m}\sigma_i a_i\right|^p\right] \leq \left(\sum_{i=1}^{m} a_i^2\right)^{\frac{p}{2}}. \tag{49}$$

Case 2: $p > 2$

$$\begin{aligned}
\mathbb{E}\left[e^{\sigma \mathbf{a}}\right] &= \prod_{i=1}^{m}\mathbb{E}\left[e^{\sigma_i a_i}\right] && (50)\\
&= \prod_{i=1}^{m}\frac{e^{a_i}+e^{-a_i}}{2} && (51)\\
&= \prod_{i=1}^{m}\cosh\left(a_i\right) && (52)\\
&\leq e^{\frac{1}{2}\sum_{i=1}^{m}a_i^2} \qquad (\cosh(x) \leq e^{\frac{1}{2}x^2}) && (53)\\
&\leq e^{\frac{1}{2}\|\boldsymbol{a}\|_2^2} && (54)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\cosh(\boldsymbol{a}\cdot\boldsymbol{\sigma})] &= \frac{1}{2}\mathbb{E}\left[e^{\boldsymbol{a}\cdot\boldsymbol{\sigma}}+e^{-\boldsymbol{a}\cdot\boldsymbol{\sigma}}\right]\\
&\leq e^{\frac{1}{2}\|\boldsymbol{a}\|_2^2}
\end{aligned} \tag{55}$$

Since $\cosh(x)$ grows faster than x^p, for any $0 < p < \infty$, there exists a positive constant B_p satisfying $|x|^p \leq B_p \cosh(x)$. Hence,

$$\mathbb{E}\left[|\mathbf{a}\cdot\boldsymbol{\sigma}|^p\right] \leq B_p \mathbb{E}[\cosh(\mathbf{a})] \leq B_p e^{\frac{1}{2}\|\mathbf{a}\|_2^2} \tag{56}$$

$$\mathbb{E}\left[|\mathbf{a}\cdot\boldsymbol{\sigma}|^p\right] \leq C_p\|\mathbf{a}\|_2^p \tag{57}$$

$$\mathop{\mathbb{E}}_{\sigma}\left[\left|\sum_{i=1}^{m}\sigma_i a_i\right|^p\right] \leq C_P\left(\sum_{i=1}^{m} a_i^2\right)^{\frac{p}{2}}. \tag{58}$$

$$C_p = B_p \exp\left(1/2\right) \tag{59}$$

3.3 (c)

$$c_p \left(\sum_{i=1}^{m} a_i^2 \right)^{\frac{p}{2}} \leq \mathop{\mathbb{E}}_{\sigma} \left[\left| \sum_{i=1}^{m} \sigma_i a_i \right|^p \right] \tag{60}$$

Case 1: $p \geq 2$
Directly applying Jensen's inequality, we get,

$$\mathbb{E}\left[|a \cdot \sigma|^p\right] \geq \mathbb{E}\left[(a \cdot \sigma)^2\right]^{p/2} = \|a\|_2^p \tag{61}$$

$c_p = 1$ for $p \geq 2$
Case 2: $p < 2$
We choose $q > 2$,

$$\begin{aligned} \|a\|_2^{2(q-p)} &= \mathbb{E}\left[|a \cdot \sigma|^2\right]^{q-p} \\ &\leq \mathbb{E}\left[|a \cdot \sigma|^p\right]^{q-2} \mathbb{E}\left[|a \cdot \sigma|^q\right]^{2-p} \\ &\leq \mathbb{E}\left[|a \cdot \sigma|^p\right]^{q-2} C_q^{2-p} \|a\|_2^{q(2-p)} \end{aligned} \tag{62}$$

$$\mathbb{E}\left[|a \cdot \sigma|^p\right] \geq C_q^{(p-2)/(q-2)} \|a\|_2^p, \tag{63}$$

By using the $\|a\|_2 = 1$.

$$c_p = C_q^{(p-2)/(q-2)}, \tag{64}$$

where C_q is the constant in part (b).

3.4 (d)

$$\begin{aligned} \widehat{\mathcal{R}}_{\mathcal{S}}(H) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\|_1 \leq \Lambda} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right] \\ &= \frac{\Lambda}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\|_1 \leq \Lambda} \left| \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right| \right] \\ &\geq c_1 \frac{\Lambda}{m} \left[\sup_{\|\mathbf{w}\|_1 \leq \Lambda} \|x\|_2^p \right] \\ &= c_1 \frac{\Lambda}{m} \left\| X^\top \right\|_{2,\infty} \end{aligned} \tag{65}$$

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \geq c_1 \frac{\Lambda}{m} \left\| X^\top \right\|_{2,\infty} \tag{66}$$

3.5 (e)

We consider the dataset with dimension $N = 2^m$ and all the elements of data $x_{ij} \in \{-1, 1\}$, so the matrix $X \in \{-1, 1\}^{2^m \times m}$.

$$
\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(H) &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\|_1 \le \Lambda} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right] && (67) \\
&= \frac{\Lambda}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left\| \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right\|_\infty \right] && (68) \\
&= \frac{\Lambda}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\max_{j \in [1,d]} \sum_{i=1}^{m} \sigma_i \left(\mathbf{x}_i\right)_j \right] && (69) \\
&= \frac{\Lambda}{m} m && (70) \\
&= \Lambda && (71) \\
&= \frac{\Lambda}{m} \sqrt{m} \sqrt{m} && (72) \\
&= \frac{\Lambda}{m} \sqrt{\log N} \|X^T\|\|_{2,\infty} && (73)
\end{aligned}
$$

$$
\begin{aligned}
\|X^T\|\|_{2,\infty} &= \left\| \left(\left\|X^T{}_1\right\|_2, \ldots, \left\|X^T{}_N\right\|_2 \right) \right\|_\infty && (74) \\
&= \sqrt{m} && (75)
\end{aligned}
$$

We have given a example and showed the upper bound is tight for $p = 1$.

4 Note

I notice something strange to me:

In paper [2], the (p,q) group norm is given as

$$
\|\mathbf{M}\|_{p,q} = \left\| \left(\|\mathbf{M}_1\|_1, \ldots, \|\mathbf{M}_d\|_p \right) \right\|_q \qquad (76)
$$

which is different from the (p,q) group norm definition here.

References

[1] The Khintchine Inequality — almostsuremath.com. `https://almostsuremath.com/2020/08/04/the-khintchine-inequality/`.

[2] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the rademacher complexity of linear hypothesis sets. *arXiv preprint arXiv:2007.11045*, 2020.