CS 7545: Machine Learning Theory **Fall 2019**

Lecture 24: Generalization Bounds + Neural Network

Lecturer: Jacob Abernethy *Scribes: Yiwen Chen, Qinsheng Zhang*

24.1 Generalization Bounds

24.1.1 Summary of past few lectures

We have a sample $\mathrm{S} = \{(x_1, y_1), \cdots, (x_m, y_m)\} \sim D$, where $y_i \in \{0, 1\}$ and loss l is 0-1 loss.

We define the risk for a function h as

$$R(h) = \mathbb{E}_{(x,y)\sim D}[l(h(x_i), y)]. \tag{24.1}$$

We define empirical risk minimization as

$$\hat{h}^{ERM} = \arg\min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i). \tag{24.2}$$

Next we can calculate the estimation error as

$$\begin{aligned} R(\hat{h}^{ERM}) - \min_{h^*\in\mathcal{H}} R(h^*) &\leq c_1 \sup_{h\in\mathcal{H}} |R(h) - \hat{R}(h)| \\ &\leq c_1 \sup_{g\in\mathcal{G}_{\mathcal{H}}} |\hat{\mathbb{E}}_s g - \mathbb{E} g| \\ &\leq c_2 \sup_{g} |\hat{\mathbb{E}}_s g - \mathbb{E} g| \\ &\leq c_3 (\mathcal{R}_m(\mathcal{G}_{\mathcal{H}}) + \sqrt{\frac{\log 1/\delta}{2m}}) \\ &= c_3 (\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2m}}) \\ &\leq c_4 (\sqrt{\frac{\log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}) \\ &\leq c_5 (\sqrt{\frac{\log(m^d)}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}) \\ &= c_5 \sqrt{\frac{d\log(m)}{m}} + c_5 \sqrt{\frac{\log 1/\delta}{2m}} \end{aligned} \tag{24.3}$$

where d is the VC-dimension of $\mathcal{H}$.

This shows that the estimation error is of order $O(\frac{d\log(m)}{m})$, we will next show that the lower bound of the error is $O(\frac{d}{m})$.

24.1.2 Lower Bound

Claim 24.1 *Given a Hyphothesis class $\mathcal{H}$ with VC dim d, there exists a family of distribution D_σ, $\sigma \in \Gamma$ such that for any algorithm $\mathcal{A}$ ($\mathcal{A} : \{(x_1, y_1), \cdots, (x_m, y_m)\} \to h$), we have*

$$\Pr(\mathcal{R}(\hat{h}) - \min_{h \in \mathcal{H}} \mathcal{R}(h) > \sqrt{\frac{d}{320m}}) > \frac{1}{64} \tag{24.4}$$

Proof: Read the book.
Sketch: sample $\sigma_1, \cdots, \sigma_d$ as i.i.d Rademacher distribution random variable.
Let $x_1, \cdots, x_d$ be the shattered set of X and define D_σ on $X \times \{0, 1\}$ as follows:

$$\begin{aligned} \Pr_{(x,y)\sim D_\sigma}((x, y) = (x_i, 1)) &= (\frac{1}{2} + c\sigma_i \frac{d}{m})\frac{1}{d} \\ \Pr_{(x,y)\sim D_\sigma}((x, y) = (x_i, 0)) &= (\frac{1}{2} - c\sigma_i \frac{d}{m})\frac{1}{d} \end{aligned} \tag{24.5}$$

■

We also need the following lemma.

Lemma 24.2 *If a coin has distribution Bernoulli($\frac{1}{2} \pm r$), then we need $O(\frac{1}{r^2})$ samples to get $\frac{2}{3}$ chance of correctly guessing the bias direction.*

Combine this lemma and the above proof sketch we can get a rough sense of the complete proof.

24.2 Property of Growth Function

Recall $\Pi_{\mathcal{H}}(m) = \sup_{S \subset \mathcal{X} \; |\mathcal{S}| = m} |\mathcal{H}|s|$. Let $\mathcal{H}_1$, $\mathcal{H}_2$ be two function classes,

Theorem 24.3 $\Pi_{\mathcal{H}_1 \times \mathcal{H}_2}(m) \le \Pi_{\mathcal{H}1}(m) \cdot \Pi_{\mathcal{H}2}(m)$
Proof: Homework ■
Suppose $\Pi_{\mathcal{H}1}(m)$ map $\mathcal{X} \to \mathcal{Y}$ and $\Pi_{\mathcal{H}2}(m)$ map $\mathcal{Y} \to \mathcal{Z}$, then we have

Theorem 24.4 $\Pi_{\mathcal{H}_1 \circ \mathcal{H}_2}(m) \le \Pi_{\mathcal{H}1}(m) \cdot \Pi_{\mathcal{H}2}(m)$
Proof: Homework ■

24.3 Neural Network

Definition 24.5 *Let $x = \mathbb{R}^{d_0}$ A neural network with respect to binary activatoin is a composition of f_n. $f_l \circ f_{f-1} \cdots f_2 \circ f_1 : \quad x \to \{1, -1\}$*

$$f_i : \mathbb{R}^{d_{i-1}} \to \{-1, 1\}^{d_i} \quad i = 1, \cdots l - 1$$
$$f_l : \mathbb{R}^{d_{l-1}} \to \{-1, 1\}$$

We have l layers and layer i has d_i nodes:

$$f_{i,j}(u) = sign(w^{i,j}u - \theta^{i,j}) \quad j = 1, \cdots, d_i$$

Based on the definition of the above neural network, $f_{i,j} \in \mathcal{H}_{i,j}$, $\mathcal{H}_{i,j}$ stands for class of linear thredshold on d_{i-1} dimensions.

$$VC - dim(\mathcal{H}_{i,j}) = d_{i-1} + 1$$

$$\mathcal{H}_i = \mathcal{H}_{i,1} \times \mathcal{H}_{i,2} \cdots \times \mathcal{H}_{i,d_i}$$

f_n class for all d_i outputs at layer i.
The entire class of Neural Network $\mathcal{F} = \mathcal{H}_1 \circ \mathcal{H}_2 \cdots \circ \mathcal{H}_l$

$$\begin{aligned} \Pi_{\mathcal{F}} &\leq \prod_{i=1}^{l} \Pi_{\mathcal{H}_i}(m) \\ &\leq \prod_{i=1}^{l} \prod_{j=1}^{d_i} \Pi_{\mathcal{H}_{i,j}}(m) \\ &\leq \prod_{i=1}^{l} \prod_{j=1}^{d_i} m^{d_{i-1}+1} \\ &= m^{\sum_{i=1}^{l} \sum_{j=1}^{d_i} (d_{i-1}+1)} \end{aligned} \tag{24.6}$$

Claim 24.6 *If* $\Pi_{\mathcal{H}} \leq m^N$, *then*

$$VC - dim(\mathcal{H}) = \mathcal{O}(N \log N)$$

N stands for total number of parameters in the Neural Network.

The proof is a homework. The above analysis is based on binary activation. However, when it comes to **Sigmoid** or **Relu** activation, we can prove its VC dimension equals $\mathcal{O}(T^2)$,T stands for the total operation on the input to genreate the final output value.