CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Bo Liu, Zhaozhuo Xu

Lecture #9
October 22, 2018

1 Review and Overview

This lecture mainly covers three topics:

- Wrap up the proof for the margin based generalization error of the two-layer neural network from the last lecture.
- For finite hypothesis, introduce the shattering coefficient (the growth function) that measures the richness of the hypothesis set and the Massart's lemma that relates the Rademacher Complexity to the growth function. Then introduce the concept of VC dimension and explains how it relates to the above.
- Lastly, recall the covering technique and use it to deal with infinite hypothesis.

2 Margin Theory of Neural Network (continued)

Recall from the last lecture, we define the λ-regularized exponential loss as:

$$L_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i f(x_i)) + \lambda||\theta||_2^2$$

Digression: here we briefly show why we can replace the logistic loss with the exponential loss for easier derivation.

Recall that

$$\text{logistic loss} = \sum_{i=1}^{n} \log\Big(1 + \exp\big(-y_i f_\theta(x_i)\big)\Big)$$

If the margin $y_i f_\theta(x_i)$ is large, it means $t = \exp\big(-y_i f_\theta(x_i)\big)$ is small. Since

$$\log\big(1+t\big) \approx t \quad \text{as} \quad t \to 0$$

the exponential loss is very close to the logistic loss while being easier to deal with.

Then we continue the proof from last lecture[1]. Recall the following notations:

$$\begin{aligned}
\text{the normalized margin}: &\quad \gamma_\theta \triangleq \min_i y_i f_{\bar{\theta}}(x_i) \qquad \text{where} \quad \bar{\theta} = \frac{\theta}{||\theta||} \\
\text{the max normalized margin}: &\quad \gamma^* = \max_{||\theta||\le 1} \gamma_\theta \\
\text{parameters that achieve the maximum}: &\quad \theta^* \in \arg\max_{||\theta||\le 1} \gamma_\theta
\end{aligned}$$

Definition 1 (Positive homogeneity)**.** A function f_θ is positive homogeneous if

$$\exists c > 0 \quad \text{such that} \quad \forall a > 0 \text{ and } \forall x \quad f_{a\theta}(x) = a^c f_\theta(x)$$

Notice that feed forward neural network with ReLU activation is positive homogeneous.

Theorem 1. *Let $\gamma_\lambda = \min_i y_i f_{\overline{\theta_\lambda}}(x_i)$ be the normalized margin of the global optimizer of L_λ. Assume $\gamma^* > 0$, then as $\lambda \to 0$, $\gamma_\lambda \to \gamma^*$*

Proof. (Assume $c = 2$ for simplicity)

Let $S = ||\theta_\lambda||_2$, according to the definition of θ^*, $||\theta^*||_2 = 1$ because otherwise we can always scale θ^* to norm 1 with a larger normalized margin. Then we have:

$$||S\theta^*||_2 = ||\theta_\lambda||_2 = S$$

As θ_λ is the global optimizer of L_λ,

$$L_\lambda(\theta_\lambda) \le L_\lambda(S\theta^*)$$
$$\frac{1}{n}\sum_{i=1} \exp\Big(-y_i f_{\theta_\lambda}(x_i)\Big) + \lambda||\theta_\lambda||_2^2 \le \frac{1}{n}\sum_{i=1} \exp\Big(-y_i f_{S\theta^*}(x_i)\Big) + \lambda||S\theta^*||_2^2$$
$$\overbrace{\frac{1}{n}\sum_{i=1}^n \exp\Big(-y_i f_{\theta_\lambda}(x_i)\Big)}^{(1)} \le \overbrace{\frac{1}{n}\sum_{i=1}^n \exp\Big(-y_i f_{S\theta^*}(x_i)\Big)}^{(2)}$$

As in the two layer neural network case, $f_{S\theta^*}(x_i) = S^2 f_{\theta^*}(x_i)$ by positive homogeneity when $c = 2$. Combining that with the definition of normalized margin over γ^*, we have

$$\begin{aligned}
(2) &= \frac{1}{n}\sum_{i=1}^n \exp\Big(-y_i S^2 f_{\theta^*}(x_i)\Big) && \text{(by homogeneity)}\\
&\le \frac{1}{n}\sum_{i=1}^n \exp\left(-S^2\gamma^*\right) && \text{(since } y_i f_{\theta^*}(x_i) \ge \gamma^*\text{)}\\
&= \exp\left(-S^2\gamma^*\right)\\
(1) &= \frac{1}{n}\sum_{i=1}^n \exp\Big(-y_i S^2 f_{\overline{\theta_\lambda}}(x_i)\Big) && \text{(define } \overline{\theta_\lambda} = \frac{\theta_\lambda}{S}\text{)}\\
&\ge \frac{1}{n}\exp\left(-S^2\gamma_\lambda\right) && \text{(since } \exists i,\ \ y_i f_{\overline{\theta_\lambda}}(x_i) = \gamma_\lambda\text{)}
\end{aligned}$$

Combined with the fact that (1) $\le$ (2) from above, we have

$$\frac{1}{n}\exp\left(-S^2\gamma_\lambda\right) \le \exp\left(-S^2\gamma^*\right)$$

which implies

$$\gamma^* - \gamma_\lambda \le \frac{\log(n)}{S^2}$$

Notice that while n is fixed, we can change S.

Claim 1. *As $\lambda \to 0$, $S \to \infty$.*

Proof. First we note that as $\lambda \to 0$, $L_\lambda(\theta_\lambda) \to 0$. This is because

$$L_\lambda(\theta_\lambda) \leq L_\lambda(B\theta^\star) \leq \exp(-B^2\gamma^\star) + \lambda B^2$$

and so we can choose B sufficiently large and λ sufficiently small to make the upper bound on $L_\lambda(\theta_\lambda)$ arbitrarily small.

Also note that from the same reasoning as before,

$$L_\lambda(\theta_\lambda) \geq \frac{1}{n}\exp(-S^2\gamma_\lambda) \geq \frac{1}{n}\exp(-S^2\gamma^\star)$$

where the last inequality follows because $\gamma_\lambda \leq \gamma^\star$. Thus if S remains bounded as $\lambda \to 0$, $L_\lambda(\theta_\lambda)$ will be lower bounded by some constant, a contradiction. □

With this claim,

$$\lim_{\lambda\to 0} \gamma^* - \gamma_\lambda \leq 0 \iff \lim_{\lambda\to 0} \gamma^* \leq \gamma_\lambda$$

Meanwhile, by definition,

$$\forall\lambda > 0,\ \gamma^* \geq \gamma_\lambda$$

Combine above two equations together, we reach the desired conclusion

$$\lim_{\lambda\to 0} \gamma^* = \gamma_\lambda$$

□

Recap: the above shows that minimizing logistic loss with weak regularization will result in a max margin solution. Good margin leads to better generalization by Rademacher Complexity.

3 Growth Function and VC Dimension

Consider the binary classification setup with binary loss function:

$$\mathcal{H} = \{\, x \longmapsto h(x) \in \{\pm 1\}\,\}$$

$$\mathcal{F} = \{(x,y) \longmapsto l((x,y),h) : h \in \mathcal{H}\} \quad \text{where} \quad l((x,y),h) \triangleq \mathbb{1}(y \neq h(x))$$

Here $\mathcal{H}$ is the hypothesis set with binary outcomes and $\mathcal{F}$ is the family of loss functions associated to $\mathcal{H}$. Let

$$Q_{z_1,\ldots,z_n} = \{\big(f(z_1), \ldots, f(z_n)\big) \in \mathbb{R}^n : f \in \mathcal{F}\}$$

be all possible outputs from $\mathcal{F}$ on the input sequence $z_1, z_2, \ldots, z_n$, where $z_i = (x_i, y_i)$. We can re-define the Rademacher Complexity as

$$R_s(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\sigma_i \in \{\pm 1\}^n}\left[\sup_{\boldsymbol{q}\in Q}\langle \boldsymbol{\sigma}, \boldsymbol{q}\rangle\right]$$

since

$$\boldsymbol{q} = \begin{bmatrix} f(z_1) \\ f(z_2) \\ \vdots \\ f(z_n) \end{bmatrix} \qquad \text{for some } f \in \mathcal{F}$$

We will then introduce the growth function and use the Massart's lemma to show its relationship to the Rademacher Complexity.

Definition 2 (Growth Function)**.** The growth function, a.k.a the shattering coefficient is

$$S(\mathcal{F}, n) \triangleq \max_{z_1,\ldots,z_n} \left| Q_{z_1,\ldots,z_n} \right| \qquad \text{with } Q \text{ defined as in above}$$

Lemma 1 (Equivalent Version of Massart's Lemma)**.** *Suppose $Q \subseteq [-1,1]^n$ is finite*

$$R_s(\mathcal{F}) = \frac{1}{n} \mathbb{E}_{\sigma_i \in \{\pm 1\}^n} \left[\sup_{\boldsymbol{q} \in Q} \langle \boldsymbol{\sigma}, \boldsymbol{q} \rangle \right] \leq \sqrt{\frac{2 \log |Q|}{n}}$$

Therefore, combine Massart's lemma with the growth function, we have

$$R_s(\mathcal{F}) \leq \sqrt{\frac{2 \log |Q|}{n}} \Longrightarrow R_n(\mathcal{F}) \leq \sqrt{\frac{2 \log S(\mathcal{F}, n)}{n}} \tag{1}$$

because

$$R_n(\mathcal{F}) = \mathbb{E}_z[R_s(\mathcal{F})] \leq \max_z R_s(\mathcal{F}) \leq \max_z \sqrt{\frac{2 \log |Q|}{n}} \leq \sqrt{\frac{2 \log S(\mathcal{F}, n)}{n}}$$

This implies that the growth function can be used to bound the Rademacher Complexity.

Furthermore, for binary loss, the shattering coefficient of loss is the same as the shattering coefficient of the hypothesis class. Because there exists a bijection between the loss and the hypothesis:

$$\Big(l((x_1, y_1), h), \ldots, l((x_n, y_n), h) \Big) = \Big(\mathbb{1}(y_1 \neq h(x_1)), \ldots, \mathbb{1}(y_n \neq h(x_n)) \Big)$$
$$\xleftrightarrow{\text{bijection}} \Big(h(x_1), \ldots, h(x_n) \Big)$$

As a sanity check, it is easy to observe that

$$\Big| l((x_1, y_1), h), \ldots, l((x_n, y_n), h) : h \in \mathcal{H} \Big| = \Big| h(x_1), \ldots, h(x_n) : h \in \mathcal{H} \Big|$$

As a result,

$$S(\mathcal{F}, n) = S(\mathcal{H}, n) \Longrightarrow R_n(\mathcal{F}) \leq \sqrt{\frac{2 \log S(\mathcal{H}, n)}{n}} \tag{2}$$

Intuitively, we want to know how fast $S(\mathcal{H}, n)$ grows so that we could get a bound on the Rademacher Complexity. A trivial bound results from the fact that $S(\mathcal{H}, n) \leq 2^n$ and thus $R_n(\mathcal{F}) \leq \sqrt{2}$. It makes us wonder if we could get a tighter bound than that, namely $S(\mathcal{H}, n) \ll c^n$, for some $c > 0$, which leads to the Sauer's lemma and the definition of VC dimension.

Definition 3 (VC dimension)**.** The VC dimension of a hypothesis set is defined as

$$VC(\mathcal{H}) = \sup\{n : S(\mathcal{H}, n) = 2^n\}$$

In other words, suppose $VC(\mathcal{H}) = d$, then $S(\mathcal{H}, d) = 2^d$ and $S(\mathcal{H}, d+1) < 2^{d+1}$.

Lemma 2 (Sauer's Lemma)**.** *If $S(\mathcal{H}, d) = 2^d$ and $S(\mathcal{H}, d) \leq 2^{d+1}$,*

$$S(\mathcal{H}, n) \leq \left(\frac{en}{d}\right)^d, \ \forall n > d$$

Equivalently, the above contends that ever beyond the VC dimension of a hypothesis set, the shattering coefficient grows in polynomial instead of exponentially with regard to the dimension. Notice that $\left(\frac{en}{d}\right)^d \approx n^d \ll 2^n$. If we then plug in the result of Sauer's lemma into equation (1) and (2) above, we have

$$R_n(\mathcal{H}) = R_n(\mathcal{F}) \leq \sqrt{\frac{2 \log S(\mathcal{H}, n)}{n}} \leq \sqrt{\frac{2d \log\left(\frac{en}{d}\right)}{n}}$$

We provide a simple example to further illustrate what VC dimension is.

Example (VC Dimension of Intervals) Consider the following hypothesis set

$$\mathcal{H} = \left\{ x \longmapsto \mathbb{1}\big(x \in [a,\, b]\big) : a, b \in \mathbb{R} \right\},$$

what is the VC dimension of $\mathcal{H}$?

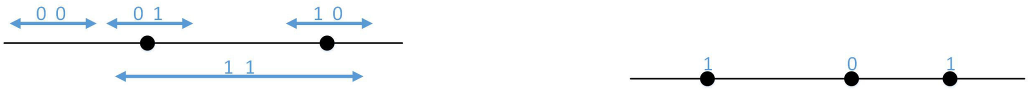
Answer: $VC(\mathcal{H}) = 2$. Let's consider the case when $n \in \{1, 2, 3\}$. We use 1/0, the value of $h(x_i)$, to denote whether $x_i \in [a, b]$.

$$S(\mathcal{H}, 1) = \Bigg| \{h(x_1) : h \in \mathcal{H}\} \Bigg| = 2^1 \qquad (1,\ 0)$$

$$S(\mathcal{H}, 2) = \Bigg| \{(h(x_1), h(x_2)) : h \in \mathcal{H}\} \Bigg| = 2^2 \qquad (11,\ 10,\ 01,\ 00)$$

$$S(\mathcal{H}, 3) = \Bigg| \{(h(x_1), h(x_2), h(x_3)) : h \in \mathcal{H}\} \Bigg| = 6 < 2^3 \qquad (111,\ 110,\ 100,\ 011,\ 001,\ 000)$$

Because it is impossible to have 101 or 010, as illustrated in the below picture.



(a)

(b)

Figure 1: VC dimension of intervals on the real line. (a) two points can always be shattered (b) three points cannot be shattered when it is $1, 0, 1$.

4 Covering Techniques

Although Massart's lemma handles the finite hypothesis well, what if the set Q is not finite? A simple solution would be making the set Q finite by discretization, which reminds us of the covering technique covered in previous lecture. However, different from before, the randomness comes from $\{\sigma_i\}s$ instead of the data.

Recall the definition of a ϵ-cover.

Definition 4 (ϵ-covering)**.** An ϵ cover of a set Q is a set C such that

$$\forall x \in Q,\ \exists x' \in C,\ \ s.t.\ \ \rho(x, x') \leq \epsilon \qquad \text{where } \rho \text{ is some metric}$$

What we want here is to discretize the space of the family of losses then apply Massart's lemma to bound the Rademacher Complexity of infinite hypothesis. In particular, notice that covering Q w.r.t metric $||\cdot||_2$ is equivalent as covering the family of losses $\mathcal{F}$ w.r.t. $\rho = L_2(P_n)$, or more specically:

$$\rho(f, f') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(f(z_i) - f'(z_i)\big)^2} \qquad \text{for } f, f' \in \mathcal{F}$$

Finally, we introduce the following theorem that completes the connection from the Rademacher Complexity to the ϵ covering of $\mathcal{F}$.

Theorem 2. *Let $\mathcal{F}$ be a family of functions from $z = (x, y) \longmapsto [-1, +1]$. Then*

$$R_s(\mathcal{F}) \leq \inf_{\epsilon} \underbrace{\sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, \rho)}{n}}}_{A} + \epsilon$$

Here $N(\epsilon, \mathcal{F}, \rho)$ denotes the minimum size of the ϵ-cover of $\mathcal{F}$ w.r.t the metric ρ. Notice that there is a trade-off between A and ϵ, e.g. as A goes up ϵ goes down and vice versa.

Proof. Let C be the ϵ-cover of $\mathcal{F}$ w.r.t ρ. In addition, let $B_\epsilon(g)$ denote the ϵ neighborhood of g w.r.t ρ. Then

$$\begin{aligned}
R_s(\mathcal{F}) &= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{g \in C} \sup_{f \in \mathcal{F} \cap B_\epsilon(g)} \sum_i \sigma_i f(z_i) \right] \qquad (\text{since } \mathcal{F} \subseteq \bigcup_{g \in C} B_\epsilon(g)) \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{g \in C} \sup_{f \in \mathcal{F} \cap B_\epsilon(g)} \left(\sum_i \sigma_i \big(f(z_i) - g(z_i)\big) + \sum_i \sigma_i g(z_i) \right) \right]
\end{aligned}$$

Notice that by Cauchy-Schwarz and the definition of ϵ-cover,

$$\sum_i \sigma_i \big(f(z_i) - g(z_i)\big) \leq \sqrt{\left(\sum_i \sigma^2 \right) \left(\sum_i \big(f(z_i) - g(z_i)\big)^2 \right)} \leq n\epsilon$$

Therefore

$$\begin{aligned}
R_s(\mathcal{F}) &= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{g \in C} \sup_{f \in \mathcal{F} \cap B_\epsilon(g)} \left(\sum_i \sigma_i \big(f(z_i) - g(z_i)\big) + \sum_i \sigma_i g(z_i) \right) \right] \\
&\leq \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{g \in C} \sup_{f \in \mathcal{F} \cap B_\epsilon(g)} \left(n\epsilon + \sum_i \sigma_i g(z_i) \right) \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{g \in C} \left(n\epsilon + \sum_i \sigma_i g(z_i) \right) \right] \qquad (\text{since there is no } f \text{ involved}) \\
&= \epsilon + \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{g \in C} \left(\sum_i \sigma_i g(z_i) \right) \right] \\
&= \epsilon + R_s(C)
\end{aligned}$$

Notice that we can pick whatever ϵ we want, combined with Massart's lemma, we have

$$R_s(\mathcal{F}) \leq \inf_{\epsilon} \Big(\epsilon + R_s(C) \Big) \leq \inf_{\epsilon} \sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, \rho)}{n}} + \epsilon$$

□

References

[1] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.