SYS 6016/4582: Machine Learning **Spring 2017**

Lecture 6 & 7

Instructor: Quanquan Gu *Date: Feb 7th & 9th, 2017* *Scriber: Yujie Xu, Renkun Ni*

Last time we introduced the definitions of **Rademacher Complexity** and **Empirical Rademacher Complexity**, which are restated here.

Definition 1 (Rademacher Complexity) *For a given function class $\mathcal{F}$, its Rademacher complexity is defined as*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_Z \mathbb{E}_\sigma \bigg[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \bigg],$$

where σ_i's are i.i.d. Rademacher random variables, and Z_i's are i.i.d. random variables.

Definition 2 (Empirical Rademacher Complexity) *For a given function class $\mathcal{F}$ and a set of observations $Z_1, Z_2, \ldots, Z_n$, the Empirical Rademacher Complexity is defined as*

$$\widehat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \bigg[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \bigg| Z_{1:n} \bigg],$$

where σ_i's are i.i.d. Rademacher random variables, and $Z_{1:n}$ is a shorthand notation for $Z_1, Z_2, ..., Z_n$.

Recall that we want to bound supremum of the difference between the empirial risk and expected risk,

$$\sup_{h \in \mathcal{H}} \big| \widehat{R}_n(h) - R(h) \big| \leq \max \bigg\{ \sup_{h \in \mathcal{H}} \widehat{R}_n(h) - R(h), \sup_{h \in \mathcal{H}} R(h) - \widehat{R}_n(h) \bigg\}.$$

With the fact that

$$\mathbb{P}\bigg(\sup_{h \in \mathcal{H}} \big| \widehat{R}_n(h) - R(h) \big| \geq t \bigg) \leq \mathbb{P}\bigg(\underbrace{\sup_{h \in \mathcal{H}} \widehat{R}_n(h) - R(h)}_{G_n(\mathcal{H})} \geq t \bigg) + \mathbb{P}\bigg(\underbrace{\sup_{h \in \mathcal{H}} R(h) - \widehat{R}_n(h)}_{G'_n(\mathcal{H})} \geq t \bigg),$$

we can bound the term $\sup_{h \in \mathcal{H}} \big| \widehat{R}_n(h) - R(h) \big|$ by bounding $G_n(\mathcal{H})$ and $G'_n(\mathcal{H})$ respectively. Since $G_n(\mathcal{H})$ and $G'_n(\mathcal{H})$ are similiar, let's consider $G_n(\mathcal{H})$ first. From the last lecture, we introduced the roadmap to bound $G_n(\mathcal{H})$.
Step1: Bound $G_n(\mathcal{H})$ by Bounded Difference Inequality,i.e.

$$G_n(\mathcal{H}) \leq \mathbb{E}[G_n(\mathcal{H})] + \sqrt{\log(1/\sigma)/2n}.$$

Step2: Bound $\mathbb{E}[G_n(\mathcal{H})]$

$$\mathbb{E}[G_n(\mathcal{H})] \leq 2\mathbb{E}_{X,Y} \mathbb{E}_\sigma \bigg[\sup_{h \in \mathcal{H}} 1/n \sum_{i=1}^{n} \sigma_i \mathbb{1}\{h(X_i) \neq Y_i\} \bigg] = 2\mathcal{R}_n(\ell \circ \mathcal{H}),$$

where ℓ is the 0-1 loss function. We can see that the above bound applies to $G'_n(\mathcal{H})$ as well.

Before bounding the Rademacher Complexity, we first introduce some properties of Rademacher Complexity.

Property 1 *If there is only one function in the function class, namely $\mathcal{F} = \{f\}$, then given $Z_{1:n}$, $\mathcal{R}_n(\mathcal{F}) = 0$.*

Proof:

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_Z \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \middle| Z_{1:n} \right] = \mathbb{E}_Z \mathbb{E}_\sigma \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \middle| Z_{1:n} \right]$$
$$= \mathbb{E}_Z \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\sigma [\sigma_i f(Z_i)] \middle| Z_{1:n} \right] = 0.$$

The last equation holds due to the symmetry of Rademacher variables. ■

Property 2 (Singleton) *If there is only one function in the funcion class $\mathcal{F}$, then the Rademacher complexity of $\mathcal{F}$ is 0.*

Property 3 (Boundedness) *The Rademacher Complexity has the following trivial bound:*

$$\mathcal{R}_n(\mathcal{F}) \leq \sup_{f \in \mathcal{F}} \sup_{z \in \mathcal{Z}} f(z).$$

Property 4 (Monotonicity) *If $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then*

$$\mathcal{R}_n(\mathcal{F}_1) \leq \mathcal{R}_n(\mathcal{F}_2).$$

Remark 1 *This must be true because if plugging in $\mathcal{F}_1$ to the definition of Rademacher complexity, the supremum is over $\mathcal{F}_1$. If considering the Rademacher Complexity of $\mathcal{F}_2$ by changing $\mathcal{F}_1$ to $\mathcal{F}_2$, the supremum will increase because it takes supremum over a larger domain.*

Property 5 (Scaling) *For any function class $\mathcal{F}$ and a real number c, define function class $c \cdot \mathcal{F} := \{c \cdot f : f \in \mathcal{F}\}$. Then we have*

$$\mathcal{R}_n(c \cdot \mathcal{F}) = |c| \mathcal{R}_n(\mathcal{F}).$$

Property 6 (Lipschitz Composition) *If ϕ is a Lipschitz continuous function over $\mathbb{R}$ with parameter L, i.e.,*

$$\forall\ t, s \in \mathrm{dom}(\phi), \quad |\phi(t) - \phi(s)| \leq L \cdot |t - s|,$$

then

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \leq L \mathcal{R}_n(\mathcal{F}).$$

It can be seen that the Rademacher Complexity is the expectation of the Empirical Rademacher Complexity. According to the concentration inequality we introduced before, we can show that the Empirical Rademacher complexity concentrate around the Rademacher Complexity. If we can bound some term by Rademacher Complexity, then we can bound it by Empirical Rademacher Complexity plus some additional deviation term. That is why we introduce the Empirical Rademacher Complexity. And also for many machine learning problems, the Empirical Rademacher Complexity is very easy to calculate. Therefore we intend to calculate the Empirical Rademacher Complexity instead of the Rademacher Complexity in quite a few application scenarios.

The next lemma gives a bound of the Empirical Rademacher Complexity $\widehat{\mathcal{R}}_n(\mathcal{F})$ in the case that $\mathcal{F}$ is finite.

Lemma 1 (Massart's Finite Class Lemma) *Let $\mathcal{F}$ be a finite set of functions. Moreover, suppose that all functions in $\mathcal{F}$ are bounded, i.e., for any $f \in \mathcal{F}$, $|f(Z_i)| \leq M$ for any observation Z_i. Then its Empirical Rademacher Complexity is bounded by*

$$\widehat{\mathcal{R}}_n(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log|\mathcal{F}|}{n}}.$$

Proof: Recall that

$$\widehat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\middle| Z_{1:n}\right]$$

Define $W_f = 1/n \sum_{i=1}^{n} \sigma_i f(Z_i)$ then we have

$$\widehat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} W_f \middle| Z_{1:n}\right].$$

For any $t \geq 0$, we have

$$\begin{aligned}
\exp(t\widehat{\mathcal{R}}_n(\mathcal{F})) &= \exp\left(t\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} W_f \middle| Z_{1:n}\right]\right) \\
&\leq \mathbb{E}_\sigma\left[\exp\left(t\sup_{f\in\mathcal{F}} W_f\right)\middle| Z_{1:n}\right] \\
&\leq \mathbb{E}_\sigma\left[\sum_{f\in\mathcal{F}} \exp(tW_f)\middle| Z_{1:n}\right],
\end{aligned}$$

where first inequality holds due to Jensen's inequality and the convexity of exponential function, and the last inequality follows the fact that maximum is always smaller than the summation provided that each $\exp(tW_f)$ is positive. With the property of expectation, we obtain

$$\begin{aligned}
\exp(t\widehat{\mathcal{R}}_n(\mathcal{F})) &\leq \sum_{f\in\mathcal{F}} \mathbb{E}_\sigma\left[\exp(tW_f)\middle| Z_{1:n}\right] = \sum_{f\in\mathcal{F}} \mathbb{E}_\sigma\left[\exp\left(t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right)\middle| Z_{1:n}\right] \\
&= \sum_{f\in\mathcal{F}} \mathbb{E}_\sigma\left[\prod_{i=1}^{n}\exp\left(t\frac{1}{n}\sigma_i f(Z_i)\right)\middle| Z_{1:n}\right] \\
&= \sum_{f\in\mathcal{F}}\prod_{i=1}^{n} \mathbb{E}_\sigma\left[\exp\left(t\frac{1}{n}\sigma_i f(Z_i)\right)\middle| Z_{1:n}\right], \qquad (1)
\end{aligned}$$

where first two equalities follow from the definition of W_f and the property of exponential function. The last equality holds since σ_i's are independent and identically distributed. According to the Hoeffding's Lemma and the fact that $\mathbb{E}_\sigma[1/n\sigma_i f(Z_i)] = 0, -M/n \leq |1/n\sigma_i f(Z_i)| \leq M/n$, we achieve

$$\mathbb{E}_\sigma\left[\exp\left(t\frac{1}{n}\sigma_i f(Z_i)\right)\middle| Z_{1:n}\right] \leq \exp\left(\frac{t^2(2M/n)^2}{8}\right) = \exp\left(\frac{t^2M^2}{2n^2}\right). \qquad (2)$$

Submitting (2) into (1), we have

$$\exp\left(t\widehat{\mathcal{R}}_n(\mathcal{F})\right) \leq \sum_{f\in\mathcal{F}}\prod_{i=1}^{n}\exp\left(\frac{t^2M^2}{2n^2}\right) = \sum_{f\in\mathcal{F}}\exp\left(\frac{t^2M^2}{2n}\right) = |\mathcal{F}|\exp\left(\frac{t^2M^2}{2n}\right).$$

Taking logarithm and dividing by t on both sides, we have

$$\begin{aligned}\widetilde{\mathcal{R}}_n(\mathcal{F}) &\leq \frac{1}{t}\log\left[|\mathcal{F}|\exp\left(\frac{t^2M^2}{2n}\right)\right] \\ &= \frac{1}{t}\left[\log|\mathcal{F}| + \frac{t^2M^2}{2n}\right] = \frac{\log|\mathcal{F}|}{t} + \frac{tM^2}{2n}. \end{aligned} \tag{3}$$

Note that (3) holds for any $t > 0$. Therefore we can minimize the right hand side of (3) for t to get the tightest upper bound. To minimize the right hand side, we solve the following equation:

$$\frac{\log|\mathcal{F}|}{t} = \frac{tM^2}{2n}.$$

The solution is $t^* = \sqrt{(2n\log|\mathcal{F}|)/(M^2)}$. Submitting this solution into (3), we obtain

$$\widehat{\mathcal{R}}_n(\mathcal{F}) \leq \sqrt{\frac{2M^2\log|\mathcal{F}|}{n}}.$$

■