EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 13: Rademacher Complexity and Massart's Lemma

Lecturer: Jacob Abernethy *Scribes: Yi-Jun Chang, Editors: Yitong Sun and David Hong*

13.1 Rademacher Complexity

Given a function class $G : \mathcal{X} \to \mathbb{R}$, let $\sigma_1, \ldots, \sigma_m$ be i.i.d. Rademacher random variables, that is $\sigma_i \in \{-1, 1\}$ with $\mathbb{P}(\sigma_i = 1) = 1/2$, and let $S = (x_1, \ldots, x_m)$ be a sample from $\mathcal{X}$. Then the **empirical Rademacher complexity** is defined as:

$$\hat{\mathfrak{R}}_S(G) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{1}{m} \sum \sigma_i g(x_i)\right],$$

and the **Rademacher complexity** is defined as:

$$\mathfrak{R}_m(G) = \mathbb{E}_{S \sim \mathcal{D}^m}\left[\hat{\mathfrak{R}}_S(G)\right].$$

We note that the Rademacher complexity is distribution-specific.

Based on Rademacher complexity, we can show the following generalization bound:

Theorem 13.1. *Let G be a function class mapping $\mathcal{X}$ to $[0, 1]$. Then, with probability at least $1 - \delta$ and for all $g \in G$,*

$$\mathbb{E}_{x \sim \mathcal{D}}[g(x)] \leq \frac{1}{m} \sum_{i=1}^{m} g(x_i) + 2\mathfrak{R}_m(G) + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $S = (x_1, \ldots, x_m) \sim \mathcal{D}^m$.

The proof of the above theorem requires **McDiarmid's inequality**, which is presented as following:

Theorem 13.2 (McDiarmid's inequality)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{X}$, and let f be a function taking finite subsets of $\mathcal{X}$ as input. Suppose that f satisfies bounded difference condition with the uniform constant c, i.e.,*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c$$

. Then with probability at least $1 - \delta$,

$$f(S) - \mathbb{E}_{S \sim \mathcal{D}^m}[f(S)] \leq \sqrt{\frac{mc^2}{2} \log(1/\delta)},$$

where $S \sim \mathcal{D}^m$.

Proof: (Sketch) Let $S = (x_1, \ldots, x_m) \sim \mathcal{D}^m$. We define a martingale $Z_i = \mathbb{E}[f(S) - \mathbb{E}[f(S)] | x_1, \ldots, x_{i-1}]$. It is easy to see that $|Z_i - Z_{i-1}| \leq c$ for all i. Then applying Azuma's inequality to the martingale difference sequence $\{Z_i\}$ yields the desired result. See Appendix D of the textbook *Foundation of Machine Learning* for a full proof. ■

We are ready to prove Theorem 13.1.

Proof of Theorem 13.1: To ease some notations, we define: $\mathbb{E}g := \mathbb{E}_{x\sim\mathcal{D}}[g(x)]$, $\hat{\mathbb{E}}_S g := \frac{1}{|S|}\sum_{x_i\in S} g(x_i)$, and $\Phi(S) := \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_S g)$.

The proof is composed of two parts:

1. $\Phi(S) \le \mathbb{E}_{S\sim\mathcal{D}^m}[\Phi(S)] + \sqrt{\frac{\log(1/\delta)}{2m}}$.
2. $\mathbb{E}_{S\sim\mathcal{D}^m}[\Phi(S)] \le 2\mathfrak{R}_m(G)$.

For part 1, we begin with showing that $|\Phi(S) - \Phi(S')| \le \frac{1}{m}$ when S and S' differ by one element (and let it be the i^{th} one):

$$\begin{aligned}
\Phi(S) - \Phi(S') &= \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_S g) - \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_{S'} g) \\
&\le \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_S g - \mathbb{E}g + \hat{\mathbb{E}}_{S'} g) \\
&= \sup_{g\in G} \frac{g(x_i') - g(x_i)}{m} \\
&\le \frac{1}{m}
\end{aligned}$$

The first inequality holds since supremum of difference is greater than difference of supremum.

By symmetry, we have $|\Phi(S) - \Phi(S')| \le \frac{1}{m}$. Then, by setting $c = \frac{1}{m}$, applying McDiarmid's inequality yields the desired inequality.

For part 2, we use the two-sample trick. Let $S' = (x_1', \ldots, x_n') \sim \mathcal{D}^m$.

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[\Phi(S)] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_S g)\right] \\
&\le \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in G}(\hat{\mathbb{E}}_{S'} g - \hat{\mathbb{E}}_S g)\right] \\
&= \mathbb{E}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m (g(x_i') - g(x_i))\right] \\
&= \mathbb{E}_{S,S'\sim\mathcal{D}^m,\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m \sigma_i(g(x_i') - g(x_i))\right] \\
&\le \mathbb{E}_{S'\sim\mathcal{D}^m,\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m \sigma_i g(x_i')\right] + \mathbb{E}_{S\sim\mathcal{D}^m,\boldsymbol{\sigma}}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m -\sigma_i g(x_i)\right] \\
&= 2\mathfrak{R}_m(G)
\end{aligned}$$

Combining the two parts gives us

$$\mathbb{E}_{x\sim\mathcal{D}}[g(x)] \le \frac{1}{m}\sum_{i=1}^m g(x_i) + 2\mathfrak{R}_m(G) + \sqrt{\frac{\log(1/\delta)}{2m}}.$$

The proof is complete. ■

13.2 Generalization Bound for Binary Classification

Given a hypothesis class $\mathcal{H}$ with functions taking ± 1 values, the associated **loss class** of $\mathcal{H}$ is defined as:

$$G := \{g_h(x,y) = \mathbf{1}[h(x) \neq y] | h \in \mathcal{H}\}.$$

Lemma 13.3. *For any sample $S = ((x_1, y_1), \dots, (x_m, y_m))$, we have $\hat{\mathfrak{R}}_S(G) = \frac{1}{2}\hat{\mathfrak{R}}_{S\restriction\mathcal{X}}(\mathcal{H})$, where $S \restriction \mathcal{X} = (x_1, \dots, x_m)$.*

Proof: The proof is easy. See Lemma 3.1 in the textbook. ■

The following theorem demonstrates an application of Rademacher complexity that provides us a generalization bound for binary classification.

Theorem 13.4. *For binary classification with 0-1 loss, let $\mathcal{H}$ be a class hypothesis mapping $\mathcal{X}$ to $\{-1, 1\}$. Then with probability $\geq 1 - \delta$, for any $h \in \mathcal{H}$, we have:*

$$R(h) \leq \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $S \sim \mathcal{D}^m$.

Proof: This directly follows from Theorem 13.1 and Lemma 13.3. ■

13.3 Massart's Lemma

Lastly, we present **Massart's lemma**, which gives us a better expression of $\mathfrak{R}_m(\cdot)$.

Theorem 13.5 (Massart's lemma)**.** *Let $A \subseteq \mathbb{R}^m$ be a finite set of points with $r = \max_{\mathbf{x}\in A} \|\mathbf{x}\|_2$. Then we have*

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{x}\in A} \sum_{i=1}^{m} x_i \sigma_i\right] \leq r\sqrt{2\log(|A|)}\,,$$

where $(x_1, \dots, x_n)$ is a vector in A.

Proof: Let $t > 0$ be a number to be chosen later.

$$\begin{aligned}
\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{x}\in A} \mathbf{x}^\top \boldsymbol{\sigma}\right]\right) &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp(t \max_{\mathbf{x}\in A} \mathbf{x}^\top \boldsymbol{\sigma})\right] && \text{(Jensen's inequality)} \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{\mathbf{x}\in A} \exp(t\mathbf{x}^\top \boldsymbol{\sigma})\right] && \text{(summation} \geq \text{maximum)} \\
&= \sum_{\mathbf{x}\in A} \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp(t\mathbf{x}^\top \boldsymbol{\sigma})\right] \\
&= \sum_{\mathbf{x}\in A} \mathbb{E}_{\boldsymbol{\sigma}}\left[\prod_{i=1}^{m} \exp(t x_i \sigma_i)\right] \\
&= \sum_{\mathbf{x}\in A} \prod_{i=1}^{m} \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp(t x_i \sigma_i)\right] \\
&\leq \sum_{\mathbf{x}\in A} \prod_{i=1}^{m} \exp\left(\frac{(2t x_i)^2}{8}\right) && \text{(applying Hoeffding's lemma)} \\
&= \sum_{\mathbf{x}\in A} \exp\left(\frac{t^2}{2}\sum_{i=1}^{m} {x_i}^2\right) \\
&\leq |A| \exp\left(\frac{t^2 r^2}{2}\right) && \text{(recall that } r = \max_{\mathbf{x}\in A} \|\mathbf{x}\|_2)
\end{aligned}$$

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

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{x}\in A}\mathbf{x}^\top\boldsymbol{\sigma}\right] \leq \frac{\log(|A|)}{t} + \frac{tr^2}{2}.$$

It is minimized when taking $t = \sqrt{\frac{\log(|A|)}{r^2/2}} = \frac{\sqrt{2\log(|A|)}}{r}$, and it leads to the bound:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{x}\in A}\mathbf{x}^\top\boldsymbol{\sigma}\right] \leq r\sqrt{2\log(|A|)}.$$

■