Lecture-08: Rademacher Complexity

1 Introduction

PAC learning guarantees were for finite hypothesis sets. However typical hypothesis sets in machine learning problems are infinite, e.g. set of all hyperplanes in SVM. We will generalize existing results and derive general learning guarantees for infinite hypothesis sets.

We will reduce the infinite hypothesis set to a finite set depending on the notion of complexity. First notion is *Rademacher complexity*, which is difficult to compute empirically for many hypothesis sets. We then study combinatorial notions of complexity, *growth function* and the *VC-dimension*. We relate Rademacher complexity to growth function, and then bound the growth function by the VC-dimension, which are easy to bound or compute in many cases.

2 Rademacher complexity

Consider a hypothesis set $H \subset \mathcal{Y}^{\mathcal{X}}$ and loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, then for each hypothesis $h \in H$, we can associate a function $g : \mathcal{Z} \to \mathbb{R}$ such that $g(x,y) = L(h(x),y)$ which captures the corresponding loss L. The family of loss function associated to hypothesis set H is defined as

$$G \triangleq \left\{ g \in \mathbb{R}^{\mathcal{Z}} : g(x,y) = L(h(x),y) \text{ for all } (x,y) \in \mathcal{X} \times \mathcal{Y}, h \in H \right\}.$$

The Rademacher complexity captures the richness of a family of functions by measuring the degree to which a hypothesis set can fit random noise.

Definition 2.1 (Rademacher random variable). A uniform random variable $X : \Omega \to \{-1,1\}$ is called a **Rademacher random variable**.

For any $g \in G$ and m-sized sample $z \in \mathcal{Z}^m$, we denote by $g_z \triangleq (g(z_1),\ldots,g(z_m)) \in \mathbb{R}^m$.

Definition 2.2 (Empirical Rademacher complexity). Let $G \subseteq [a,b]^{\mathcal{Z}}$ be a family of functions and a fixed labeled sample $z = (z_1,\ldots,z_m) \in \mathcal{Z}^m$ of size m. Then, the **empirical Rademacher complexity** of G with respect to the labeled sample z is defined as

$$\hat{\mathcal{R}}_z(G) \triangleq \mathbb{E}\left[\sup_{g \in G} \frac{1}{m} \langle \sigma, g_z \rangle \right],$$

where $\sigma : \Omega \to \{-1,1\}^m$, is an m-length vector of independent Rademacher variables.

Remark 1. The inner product $\langle \sigma, g_z \rangle$ measures the correlation of g_z with random noise σ, and the supremum over all $g \in G$ measures how well the hypothesis class H correlates with σ over the labeled sample z. This is a measure of richness/complexity of class G, since richer families can generate more g_z and better correlate with random noise on average.

Definition 2.3 (Rademacher complexity). Let D be the unknown fixed distribution according to which labeled sample $z \in \mathcal{Z}^m$ is drawn in an *i.i.d.* fashion. For any $m \in \mathbb{N}$, the **Rademacher complexity** of a family of loss functions G is mean of empirical Rademacher complexity for sample z, and denoted by

$$\mathcal{R}_m(G) \triangleq \mathbb{E}\hat{\mathcal{R}}_z(G).$$

Lemma 2.4. *Let $G \subseteq [0,1]^{\mathcal{Z}}$ be a family of functions. Then, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$*

$$\mathcal{R}_m(G) \leqslant \hat{\mathcal{R}}_z(G) + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Proof. We observe that $\mathbb{E}\hat{\mathcal{R}}_z(G)$, and that $\hat{\mathcal{R}}_G$ satisfies the bounded difference property with bounding vector $\frac{1}{m}\mathbf{1}$. The result follows from the McDiarmid's inequality. □

Theorem 2.5. *Let $G \subseteq [0,1]^{\mathcal{Z}}$ be a family of functions. Then, for any $\delta > 0$, with probability at least $1-\delta$, both the inequalities hold for all $g \in G$*

$$\mathbb{E}g(z) \leqslant \frac{1}{m}\langle \mathbf{1}, g_z\rangle + 2\mathcal{R}_m(G) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}, \qquad \mathbb{E}g(z) \leqslant \frac{1}{m}\langle \mathbf{1}, g_z\rangle + 2\hat{\mathcal{R}}_z(G) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$

Proof. For any labeled sample $z \in \mathcal{Z}^m$ and loss function $g \in G$, we denote the empirical average of g over labeled sample z as

$$\hat{\mathbb{E}}_z[g] \triangleq \frac{1}{m}\langle \mathbf{1}, g_z\rangle.$$

We consider the following function $\Phi : \mathcal{Z}^m \to \mathbb{R}$,

$$\Phi(z) \triangleq \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_z[g]).$$

Consider two samples z, z' differing at a single example z_m in z and z'_m in z'. Then, we can write

$$\Phi(z') - \Phi(z) \leqslant \sup_{g\in G}(\hat{\mathbb{E}}_{z'}[g] - \hat{\mathbb{E}}_z[g]) = \sup_{g\in G}\frac{g(z_m) - g(z'_m)}{m} \leqslant \frac{1}{m}.$$

Similarly, we can obtain $\Phi(z) - \Phi(z') \leqslant \frac{1}{m}$. Hence, the function Φ has the bounded difference property with bounding vector $\frac{1}{m}\mathbf{1}$. By McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$, we have

$$\Phi(z) \leqslant \mathbb{E}\Phi(z) + \sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$

We next bound the mean of the $\Phi(z)$ by the difference of empirical average for samples z, z', sampled *i.i.d.* from the fixed unknown distribution D, by applying the Jensen's inequality to convex function supremum. We get

$$\mathbb{E}\Phi(z) = \mathbb{E}\left[\sup_{g\in G}(\mathbb{E}[g] - \hat{\mathbb{E}}_z[g])\right] = \mathbb{E}\left[\sup_{g\in G}\mathbb{E}\left[\hat{\mathbb{E}}_{z'}[g] - \hat{\mathbb{E}}_z[g]\right]\right] \leqslant \mathbb{E}\left[\sup_{g\in G}(\hat{\mathbb{E}}_{z'}[g] - \hat{\mathbb{E}}_z[g])\right].$$

Since z, z' are *i.i.d.* , the inner product $\langle \sigma, g_{z'} - g_z\rangle$ for *i.i.d.* Rademacher vector $\sigma \in \{-1,1\}^m$ has same distribution as $\langle \mathbf{1}, g_{z'} - g_z\rangle$. Therefore, we have

$$\mathbb{E}\Phi(z) \leqslant \mathbb{E}\left[\sup_{g\in G}\frac{1}{m}\langle \sigma, g_{z'} - g_z\rangle\right] \leqslant \mathbb{E}\left[\sup_{g\in G}\frac{1}{m}\langle \sigma, g_{z'}\rangle\right] + \mathbb{E}\left[\sup_{g\in G}\frac{1}{m}\langle -\sigma, g_z\rangle\right] = 2\mathcal{R}_m(G).$$

□

Lemma 2.6. *Let $\mathcal{Y} = \{-1,1\}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and the hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$ be a family of functions and let G be the family of loss functions associated to the hypothesis set H for the zero-one loss, i.e.*

$$G = \left\{(x,y) \mapsto \mathbb{1}_{\{h(x)\neq y\}} : h \in H\right\}.$$

For any labeled sample $z \in \mathcal{Z}^m$, let $x = z_{\mathcal{X}}$ denote its projection over $\mathcal{X}$, i.e. $x = (x_1, \ldots, x_m) \in \mathcal{X}^m$. Then,

$$\hat{\mathcal{R}}_z(G) = \frac{1}{2}\hat{\mathcal{R}}_x(H).$$

Proof. For any sample $z = ((x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [m])$ where $\mathcal{Y} = \{-1,1\}$, we have $\mathbb{1}_{\{h(x_i)\neq y_i\}} = \frac{1-y_i h(x_i)}{2}$. Therefore, we can write

$$\hat{\mathcal{R}}_z(G) = \mathbb{E}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^m \sigma_i \mathbb{1}_{\{h(x_i)\neq y_i\}}\right] = \mathbb{E}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^m \sigma_i\Big(\frac{1 - y_i h(x_i)}{2}\Big)\right].$$

Since $\sum_{i=1}^{m} \sigma_i$ remains constant for all $h \in H$ and its mean is zero, we can ignore this term. Further, $\sigma \circ y = (\sigma_i y_i \in \mathcal{Y} : i \in [m])$ has same distribution as $\sigma = (\sigma_i \in \mathcal{Y} : i \in [m])$, and therefore

$$\hat{\mathcal{R}}_z(G) = \frac{1}{2}\mathbb{E}\left[\sup_{h\in H} \frac{1}{m} \langle -\sigma, y \circ h(x)\rangle\right] = \frac{1}{2}\mathbb{E}\left[\sup_{h\in H} \frac{1}{m} \langle \sigma, h(x)\rangle\right] = \frac{1}{2}\hat{\mathcal{R}}_x(H).$$

□

Theorem 2.7 (Rademacher complexity bounds – binary classification). *Let $H \subseteq \mathcal{X}^{\mathcal{Y}}$ be a family of functions for $\mathcal{Y} = \{-1, +1\}$ and let D be the fixed and unknown distribution over the input space $\mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample $z \in \mathcal{Z}^m$ of size m drawn* i.i.d. *according to D, each of the following holds for any hypothesis $h \in H$*

$$R(h) \leqslant \hat{R}(h) + \mathcal{R}_m(H) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}, \qquad R(h) \leqslant \hat{R}(h) + \hat{\mathcal{R}}_S(H) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Proof. The result follow from Theorem **??** and Lemma **??**. □

Remark 2. The second learning bound is data dependent, and very useful if we can efficiently compute the empirical Rademacher complexity $\hat{\mathcal{R}}_S(H)$. Since σ and $-\sigma$ have the same distribution, we get

$$\hat{\mathcal{R}}_S(H) = \mathbb{E}\left[\sup_{h\in H} \frac{1}{m} \langle -\sigma, h_S\rangle\right] = -\mathbb{E}\left[\inf_{h\in H} \frac{1}{m} \langle \sigma, h_S\rangle\right].$$

for a fixed value of σ, computing $\inf_{h\in H} \frac{1}{m} \langle \sigma, h_S\rangle$ is equivalent to an *empirical risk minimization* problem, which is known to be computationally hard for some hypothesis sets.

A McDiarmid's inequality

Definition A.1 (Martingale difference). A sequence of random variables $(V_n \in \mathbb{R} : n \in \mathbb{N})$ is a **martingale difference sequence** with respect to a random sequence $(X_n \in \mathbb{R} : n \in \mathbb{N})$ if V_n is a function of $X_1, \ldots, X_n$ for all $n \in \mathbb{N}$, and

$$\mathbb{E}[V_{n+1} \mid X_1, \ldots, X_n] = 0.$$

Lemma A.2. *Let V and Z be random variables satisfying $\mathbb{E}[V \mid Z] = 0$ and $f(Z) \leqslant V \leqslant f(Z) + c$ for some function f and constant $c \geqslant 0$. Then, for all $t > 0$, we have*

$$\mathbb{E}[e^{sV} \mid Z] \leqslant e^{t^2 c^2/8}.$$

Proof. The result follows from Hoeffding's Lemma for conditional expectation given Z, where $[a, b] = [f(Z), f(Z) + c]$. □

Theorem A.3 (Azuma's inequality). *Let $(V_n : n \in \mathbb{N})$ be a martingale difference sequence with respect to the random variables $(X_n : n \in \mathbb{N})$ and assume that for all $n \in \mathbb{N}$ there is a constant $c_n \geqslant 0$ and random variable Z_n, which is a function of $X_1, \ldots, X_{i-1}$, that satisfy $Z_i \leqslant V_i \leqslant Z_i + c$. Defining $\sigma^2 \triangleq \sum_{i=1}^{m} c_i^2 = \|c\|_2^2$, we have for all $\epsilon > 0$ and $m \in \mathbb{N}$,*

$$P\left\{\sum_{i=1}^{m} V_i \geqslant \epsilon\right\} \leqslant e^{-2\epsilon^2/\sigma^2}, \qquad P\left\{\sum_{i=1}^{m} V_i \leqslant -\epsilon\right\} \leqslant e^{-2\epsilon^2/\sigma^2}.$$

Proof. For any $k \in \mathbb{N}$, we can define $S_k \triangleq \sum_{i=1}^{k} V_i$, then by Chernoff bound, we have

$$P\{S_m \geqslant \epsilon\} \leqslant e^{-t\epsilon}\mathbb{E}[e^t S_m] = e^{-t\epsilon}\mathbb{E}[e^t S_{m-1}\mathbb{E}[e^{tV_m} | X_1, \ldots, X_{m-1}]] \leqslant e^{-t\epsilon}\mathbb{E}[e^{tS_{m-1}}]e^{t^2 c_m^2/8} \leqslant \exp\left(-t\epsilon + \frac{t^2\sigma^2}{8}\right).$$

The result for the first part follows by taking $t^* = \frac{4\epsilon}{\sigma^2}$. The second part can be proved similarly. □

Definition A.4 (Bounded difference property). A function $f : \mathcal{X}^m \to \mathbb{R}$ is said to have the **bounded difference property**, if for all $i \in [m]$ there exists a constant $c_i > 0$ such that for any $x, y \in \mathbb{R}^m$ differing only at the ith location, we have

$$|f(x) - f(y)| \leqslant c_i. \tag{1}$$

The vector $c \in \mathbb{R}_+^m$ is called the **bounding vector**.

Theorem A.5 (McDiarmid's inequality). *Let $f : \mathcal{X}^m$ be a function with the bounded difference property with bounding vector $c \in \mathbb{R}_+^m$, and $(X_i \in \mathcal{X} : i \in [m])$ be a set of m independent random variables. Denoting $f(S) \triangleq f(X_1, \ldots, X_m)$, for all $\epsilon > 0$, we have*

$$P\{f(S) - \mathbb{E}f(S) \geqslant \epsilon\} \leqslant e^{-2\epsilon^2/\|c\|_2^2}, \qquad P\{f(S) - \mathbb{E}f(S) \leqslant -\epsilon\} \leqslant e^{-2\epsilon^2/\|c\|_2^2}.$$

Proof. It suffices to show that $f(S) - \mathbb{E}f(S) = \sum_{i=1}^m V_i$ for some martingale difference sequence $(V_i : i \in [m])$ with respect to the sequence $(X_i : i \in [m])$ and $Z_i \leqslant V_i \leqslant Z_i + c_i$ for some random variable Z_i a function of $X_1, \ldots, X_{i-1}$.

Let $V = f(S) - \mathbb{E}f(S)$, then we define such a sequence $(V_1, \ldots, V_m)$ as

$$V_k = \mathbb{E}[V \mid X_1, \ldots, X_k] - \mathbb{E}[V \mid X_1, \ldots, X_{k-1}],\ k \in [m], \qquad \sum_{k=1}^m V_k = V.$$

We can verify that $(V_i : i \in [m])$ is martingale difference equation, since V_k is a function of $X_1, \ldots, X_k$ and $\mathbb{E}[V_k | X_1, \ldots, X_{k-1}] = 0$ for each $k \in [m]$. Since $\mathbb{E}f(S)$ is not random, we can write

$$V_k = \mathbb{E}[f(S) \mid X_1, \ldots, X_k] - \mathbb{E}[f(S) \mid X_1, \ldots, X_{k-1}],$$

and define upper and lower bounds for V_k as

$$W_k \triangleq \sup_x \mathbb{E}[f(S) \mid X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f(S) \mid X_1, \ldots, X_{k-1}], \quad U_k \triangleq \inf_x \mathbb{E}[f(S) \mid X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f(S) \mid X_1, \ldots, X_{k-1}].$$

Then the result follows from the hypothesis (**??**), which implies that

$$W_k - U_k = \sup_{x,y \in \mathcal{X}} E[f(S) \mid X_1, \ldots, X_{k-1}, x] - E[f(S) \mid X_1, \ldots, X_{k-1}, y] \leqslant c_k.$$

□