Learnability with Rademacher Complexities

Daniel Khashabi

Fall 2013
Last Update: September 26, 2016

1 Introduction

Our goal in study of passive supervised learning is to find a hypothesis h based on a set of examples that has small error with respect to some target function.

One can improve generalization by controlling the complexity of the concept class $\mathcal{H}$ from which we are choosing a hypothesis. One way to achieve this is via the ideas in VC dimension [*]. Here we will introduce Rademacher complexity as another way of handling hypothesis space complexity, and as a result, deriving generalization bounds. Here are some major differences our results will have with those in the discussion of VC dimension:

- One observation in the dicusssion of VC dimension is that it is independent of the data distribution. In other words, its guranttees hold for any data distribution; on the other hand, the bound that it gives might not be tight for certains data distributions.
- The analysis of VC dimension bound apply to discrete problems (such as classification), and it does not state anything about problems like regression.

2 Rademacher Averages/Complexities

Here we define Rademacher complexity which will be used in bounding risk functions.

Definition 2.1 (Rademacher Average)**.** *If $\mathcal{H} \subset \mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of functions we are exploring defined on domain $X \subset \mathcal{X}$, and $S = \{x_i\}_{i=1}^n$ be the set of samples generated by some unknown distribution $D_{\mathcal{X}}$ on the same domain $\mathcal{X}$. Define σ_i to be uniform random variable on ± 1, for any i. The "empirica" Rademacher average or complexity is defined as following:* [†]

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| \Big| \{x_i\}_{i=1}^n \right] = \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \Big| \{x_i\}_{i=1}^n \right]$$

and the expectation of the above measure, with respect to the random samples $\{x_i\}_{i=1}^n$, is called the Rademacher average or complexity:

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E} \hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| \right]$$

[*] http://web.engr.illinois.edu/~khashab2/learn/vc.pdf

[†] Implicit assumption: suppremum over the function class $\mathcal{H}$ is measurable.

There is a similar definition without the absolutes, which have similar properties as above:

$$\hat{\mathfrak{R}}_S^a(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^n \sigma_i f(x_i)\,|\{x_i\}_{i=1}^n\right]$$

and

$$\mathfrak{R}_n^a(\mathcal{H}) = \mathbb{E}\hat{\mathfrak{R}}_S^a(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^n \sigma_i f(x_i)\right]$$

Another way of writing the Rademacher complexiity is the following

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\left|\frac{\mathbf{f}_S.\sigma}{n}\right|\,|\{x_i\}_{i=1}^n\right]$$

where $\mathbf{f}_S = (f(x_1),\ldots,f(x_n))^\top$, and $\sigma = (\sigma_1,\ldots,\sigma_n)$. The dot product $\mathbf{f}_S.\sigma$ measures the correlation between the function values, and the radnom noise vector. In overall, the Rademacher complexity measures how well the function class $\mathcal{H}$ can correlate with random noise. The richer the hypothesis class it, the better it will correclate with the random noise.

Here are some useful properties of the Rademacher averages.

Lemma 2.1. *For any $\{x_i\}_{i=1}^n$ and for any function class $\mathcal{F}$ and $\mathcal{H}$, that map $\mathcal{X}\to\mathbb{R}$:*

1. *If $\mathcal{H}\subseteq\mathcal{F}$ then $\hat{\mathfrak{R}}_S(\mathcal{H}) \le \hat{\mathfrak{R}}_S(\mathcal{F})$.*
2. *For <u>any</u> function $h:\mathcal{X}\to\mathbb{R}$, then $\hat{\mathfrak{R}}_S^a(\mathcal{F}+h) = \hat{\mathfrak{R}}_S^a(\mathcal{F})$.*
3. *If $cvx(\mathcal{F}) = \{x\to\mathbb{E}_{f\sim\pi}[f(x)], \pi\in\Delta(\mathcal{F})\}$ then $\hat{\mathfrak{R}}_S^a(\mathcal{F}) = \hat{\mathfrak{R}}_S^a(cvx(\mathcal{F}))$.*
4. *$\hat{\mathfrak{R}}_S^a(\mathcal{F}+\mathcal{H}) = \hat{\mathfrak{R}}_S^a(\mathcal{F}) + \hat{\mathfrak{R}}_S^a(\mathcal{H})$.*

Proof of this proposition is included in Section 7.

Proof. We prove each proposition:

1.

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}}\left|\sum_{i=1}^n\sigma_i f(x_i)\right|\,|\{x_i\}_{i=1}^n\right] \le \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\sigma_i f(x_i)\right|\,|\{x_i\}_{i=1}^n\right] = \hat{\mathfrak{R}}_S(\mathcal{F})$$

2.

$$\begin{aligned}
\hat{\mathfrak{R}}_S^a(\mathcal{F}+h) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i\left(f(x_i)+h(x_i)\right)|\{x_i\}_{i=1}^n\right]\\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i f(x_i) + \sum_{i=1}^n\sigma_i h(x_i)\,|\{x_i\}_{i=1}^n\right]\\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i f(x_i)\,|\{x_i\}_{i=1}^n\right] + 0 = \hat{\mathfrak{R}}_S^a(\mathcal{F})
\end{aligned}$$

3.

$$\begin{aligned}
\hat{\mathfrak{R}}_S^a(\text{cvx}(\mathcal{F})) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\pi\in\Delta(\mathcal{F})}\sum_{i=1}^n \sigma_i \mathbb{E}_{f\in\pi}\left[f(x_i)\right]|\{x_i\}_{i=1}^n\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\pi\in\Delta(\mathcal{F})}\mathbb{E}_{f\in\pi}\left[\sum_{i=1}^n \sigma_i f(x_i)\right]|\{x_i\}_{i=1}^n\right] \text{ (swap only in the corners of the convex set)} \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i)\,|\{x_i\}_{i=1}^n\right] = \hat{\mathfrak{R}}_S^a(\mathcal{F})
\end{aligned}$$

4.

$$\begin{aligned}
\hat{\mathfrak{R}}_S^a(\mathcal{F}+\mathcal{H}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F},h\in\mathcal{H}}\sum_{i=1}^n \sigma_i\left(f(x_i)+h(x_i)\right)|\{x_i\}_{i=1}^n\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i) + \sup_{h\in\mathcal{H}}\sum_{i=1}^n \sigma_i h(x_i)\,|\{x_i\}_{i=1}^n\right] = \hat{\mathfrak{R}}_S^a(\mathcal{F}) + \hat{\mathfrak{R}}_S^a(\mathcal{H})
\end{aligned}$$

■

Lemma 2.2. *Given real-valued CDF function $F(x)$, and $\mathcal{F}$ being class of indicator functions on half-intervals which define the empirical CDF function:*

$$\hat{F}_S(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\le x\}}$$

we can show that,

$$\mathbb{E}_S\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_S(x) - F(x)\right|\right] \le 2\mathfrak{R}_n(\mathcal{F})$$

with $S = (X_1 = x_1, ..., X_n = x_n)$.

Proof. The trick that is commonly used for this is converting expectation to empirical mean by introducing fake/ghost samples S' and symmetrization:

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_S(x) - F(x)\right|\right] &= \mathbb{E}_S\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_S(x) - \mathbb{E}_{S'}\left[\hat{F}_{S'}(x)\right]\right|\right] \le \mathbb{E}_{S,S'}\left[\sup_{x\in\mathbb{R}}\left|\hat{F}_S(x) - \hat{F}_{S'}(x)\right|\right] \\
&= \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^n\left[\mathbf{1}_{\{X_i\le x\}} - \mathbf{1}_{\left\{X_i'\le x\right\}}\right]\right] \overset{d}{=} \mathbb{E}_{S,S',\sigma}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\left[\mathbf{1}_{\{X_i\le x\}} - \mathbf{1}_{\left\{\bar{X}_i\le x\right\}}\right]\right] \\
&\le 2\mathbb{E}_{S,\sigma}\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{1}_{\{X_i\le x\}}\right| = 2\mathfrak{R}_n(\mathcal{F}),\ \text{for } \mathcal{F} = \text{half-intervals}
\end{aligned}$$

■

It turns out that this observation is general for any loss function. The following bounding technique could be generalized to any loss function.

Lemma 2.3. *Given a class functions $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$ defined on domain $X \subset \mathcal{X}$, we have the following general bound on the Rademacher average:*

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{H}} \left| \mathbb{E}f - \hat{\mathbb{E}}_S f \right| \right] \leq 2\mathfrak{R}_n(\mathcal{H})$$

with $S = (x_1, ..., x_n)$ and $\hat{\mathbb{E}}_S f = \frac{1}{n} \sum_{i=1}^n f(x_i)$.

Proof. The steps for the previous proof hold for this proof, with some minor changes. Again, we convert expectation to empirical mean by introducing fake/ghost samples S' and symmetrization:

$$\begin{aligned}
\mathbb{E}_S \left[\sup_{f \in \mathcal{H}} \left| \mathbb{E}f - \hat{\mathbb{E}}_S f \right| \right] &= \mathbb{E}_S \left[\sup_{f \in \mathcal{H}} \left| \mathbb{E}_{S'} \hat{\mathbb{E}}_{S'} f - \hat{\mathbb{E}}_S f \right| \right] \\
&= \mathbb{E}_{S,S'} \left[\sup_{f \in \mathcal{H}} \left| \hat{\mathbb{E}}_{S'} f - \hat{\mathbb{E}}_S f \right| \right] = \mathbb{E}_{S,S',\sigma} \left[\frac{1}{n} \sum_{i=1}^n \sigma_i \left[f(x_i) - f(x_i') \right] \right] \\
&\leq 2\mathbb{E}_{S,\sigma} \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| = 2\mathfrak{R}_n(\mathcal{F})
\end{aligned}$$

■

With the following lemma we show how to generalize Rademacher averages using Lipchitz maps.

Lemma 2.4 (Ledoux-Talagrand contraction)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex and increasing function. Also let $\phi_i(x) : \mathbb{R} \to \mathbb{R}$, s.t. it satisfies $\phi_i(0) = 0$ with Lipchitz constant L (for any $x, y \in \mathbb{R} \Rightarrow |\phi_i(x) - \phi_i(y)| \leq L |x - y|$). For any $T \subset \mathbb{R}^n$,*

$$\mathbb{E}_\sigma f \left(\frac{1}{2} \sup_{t \in T} \left| \sum_{i=1}^n \sigma_i \phi_i(t_i) \right| \right) \leq \mathbb{E}_\sigma f \left(L. \sup_{t \in T} \left| \sum_{i=1}^n \sigma_i t_i \right| \right)$$

Proof. Proof with definition of Rademacher average and properties of convex functions. ■

The above lemma will result the following bound:

Corollary 2.5. *Let $\mathcal{F}$ be a class of functions with domain X and $\phi(.)$ be a L-Lipchitz map from $\mathbb{R}$ to $\mathbb{R}$ with $\phi(0) = 0$. The composition of the map on the functions is defined as $\phi \circ \mathcal{F} = \{\phi \circ f | f \in \mathcal{F}\}$. Then*

$$\mathfrak{R}_n(\phi \circ \mathcal{F}) \leq 2L\mathfrak{R}_n(\mathcal{F})$$

Proof. In the previous lemma, take the convex increasing function be the identity function. ■

2.1 Rademacher complexity of linear class

Here we analyze the Rademacher complexity of the following linear classes. These results will come handy in analyzing the generalization bounds of many forthcoming problems which involve linear models.

Define the following classes:

$$\mathcal{H}_1 = \{\mathbf{x} \to \langle \mathbf{x}, \mathbf{w} \rangle : \|\mathbf{w}\|_1 \leq 1\}, \quad \mathcal{H}_2 = \{\mathbf{x} \to \langle \mathbf{x}, \mathbf{w} \rangle : \|\mathbf{w}\|_2 \leq 1\}$$

Lemma 2.6. *Let $S = (x_1, \ldots, x_n)$, then*

$$\mathfrak{R}_n(\mathcal{H}_2 \circ S) \leq \frac{\max_i \|\mathbf{x}_i\|_2}{\sqrt{n}}$$

Proof.

$$\begin{aligned}
\mathfrak{R}_n^a(\mathcal{H}) &= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq 1} \sum_{i=1}^n \sigma_i \langle \mathbf{x}_i, \mathbf{w} \rangle \right] \\
&= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq 1} \left\langle \mathbf{w}, \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\rangle \right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\|_2 \right]
\end{aligned}$$

Due to Jensen inequality:

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\|_2 \right] \leq \frac{1}{n} \sqrt{\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\|_2^2 \right]}$$

Since the Rademacher random variables are independent of each other, we have:

$$\begin{aligned}
\frac{1}{n} \sqrt{\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\|_2^2 \right]} &= \frac{1}{n} \sqrt{\mathbb{E}_\sigma \left[\sum_{i,j} \sigma_i \sigma_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right]} \\
&= \frac{1}{n} \sqrt{\sum_{i,j,i \neq j} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \mathbb{E}_\sigma [\sigma_i \sigma_j] + \sum_i \langle \mathbf{x}_i, \mathbf{x}_i \rangle \mathbb{E}_\sigma \left[\sigma_i^2\right]} \\
&= \frac{1}{n} \sqrt{\sum_i \|\mathbf{x}_i\|^2} \\
&\leq \frac{1}{n} \sqrt{n \max_i \|\mathbf{x}_i\|^2} = \frac{\max_i \|\mathbf{x}_i\|}{\sqrt{n}}
\end{aligned}$$

■

Lemma 2.7. *Let $S = (x_1, \ldots, x_n)$, then*

$$\mathfrak{R}_n^a(\mathcal{H}_1 \circ S) \leq \max_i \|\mathbf{x}_i\|_\infty \frac{2 \log 2n}{n}$$

Proof.

$$\begin{aligned}
\mathfrak{R}_n^a(\mathcal{H}) &= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\mathbf{w}:\|\mathbf{w}\|_1 \leq 1} \sum_{i=1}^n \sigma_i \langle \mathbf{x}_i, \mathbf{w} \rangle \right] \\
&= \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\mathbf{w}:\|\mathbf{w}\|_1 \leq 1} \left\langle \mathbf{w}, \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\rangle \right] \\
&\leq \frac{1}{n}\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \right\|_\infty \right]
\end{aligned}$$

The last step is done via the finite class lemma (see Lemma 4.1). ■

3 Generalization bounds

Here is the main theorem, which contains the generalization bounds via Rademacher complexity:

Theorem 3.1. *Let $\mathcal{F}$ be a class of functions, defined on domain X and mapping to $[0,1]$. For some $\delta \in (0,1)$, and for any $f \in \mathcal{F}$:*

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}, \text{ with probability at least } 1-\delta$$

Also for any $f \in \mathcal{F}$:

$$\mathbb{E}f(X) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i) + 2\mathfrak{R}_S(\mathcal{F}) + 5\sqrt{\frac{\log 2/\delta}{2n}}, \text{ with probability at least } 1-\delta$$

Similar results can be found with slightly different definition of the Radmacher average:

Theorem 3.2. *Let $\mathcal{F}$ be a class of functions, defined on domain X and mapping to $[0,1]$. For some $\delta \in (0,1)$, and for any $f \in \mathcal{F}$:*

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2\mathfrak{R}^a_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}, \text{ with probability at least } 1-\delta$$

Also for any $f \in \mathcal{F}$:

$$\mathbb{E}f(X) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i) + 2\mathfrak{R}^a_S(\mathcal{F}) + 3\sqrt{\frac{\log 2/\delta}{2n}}, \text{ with probability at least } 1-\delta$$

A side note before jumpin into the proof: usually in practice the set $\mathcal{F}$ is a composition of input space $\mathcal{X}$, hypothesis functions $\mathcal{H}$ and the loss family ℓ whcih measures the quality of the learning:

$$\mathcal{F} = \ell \circ \mathcal{H} \circ S$$

For example for SVM, $\mathcal{H}$ is space of linear classifiers, and ℓ is margin based (hard/soft) loss.

Another issue worhty to point out is that, here we assumed that the range of the function $\mathcal{F}$ is bounded inside $[0,1]$. However if the function is ranged between $[0,c]$, a c coefficient would appear before $\sqrt{\frac{\log 2/\delta}{2n}}$ (easy to verify through the proof).

Proof. For a sample set $S = (x_1, x_2, ..., x_n)$, define the following function

$$\Phi_S(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left\{ \mathbb{E}\left[f\right] - \frac{1}{n}\sum_{x_i \in S} f(x_i) \right\}$$

Proof uses the McDiarmid's bound on the function $\Phi_S(\mathcal{F})$; define the sample set S' to be exactly the same as S, except one differing sample.

$$\Phi_S(\mathcal{F}) - \Phi_{S'}(\mathcal{F}) \leq \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{x_i \in S'} f(x_i) - \frac{1}{n}\sum_{x_i \in S} f(x_i) \right\} = \sup_{f\in\mathcal{F}} \frac{f(x_j) - f(x'_j)}{n} \leq \frac{1}{n}$$

We used the fact that supremum of diffence is bigger than the difference of supremums. Also we implicitly assumed that the function is bounded between 0 and 1. Hence we proved that

$$|\Phi_S(\mathcal{F}) - \Phi_{S'}(\mathcal{F})| \leq 1$$

Using the boundedness property of $\Phi(.)$ and using the McDiarmid's inequality we have:

$$\Phi_S(\mathcal{F}) \leq \mathbb{E}_S\left[\Phi_S(\mathcal{F})\right] + \sqrt{\frac{\log 2/\delta}{2n}}, \quad \text{with probably at least } 1 - \delta/2$$

Note that using Lemma 2.3 we know:

$$\mathbb{E}_S\left[\Phi_S(\mathcal{F})\right] \leq 2\mathfrak{R}_n(\mathcal{H})$$

which would give us the first inequality (with $\delta/2$ replaced with δ). To get the second inequaly, we apply the McDiarmid bound on the Rademacher definition:

$$\mathfrak{R}_n(\mathcal{H}) \leq \hat{\mathfrak{R}}_S(\mathcal{H}) + \sqrt{\frac{\log 2/\delta}{2n}}, \quad \text{with probably at least } 1 - \delta/2$$

Combine this with the previous result and we will the 2nd inequality the in the defintion of the theorem. ■

3.1 Concentration bounds for binary classification

We start with a few examples, and then move to more general theorems.

Example 3.3. Let $f : \mathsf{X} \to \{0, 1\}$, and let $(X, Y) \in \mathsf{X} \times \{0, 1\}$ be n random i.i.d. samplings from the joint distribution P_{XY}. Consider the empirical risk defined as,

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{f(X_i) \neq Y_i\right\}$$

1. Prove that for any $f \in \mathcal{F}$,

$$L(f) \leq L_n(f) + \sqrt{\frac{2L(f)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n} \tag{1}$$

   probability at least $1 - \delta$.
   Hint: Use Berinstein's inequality.

2. Use the result of the previous part to show that, for any $f \in \mathcal{F}$,

$$L(f) \leq L_n(f) + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{n}$$

   with probability at least $1 - \delta$. Use this to prove that if the ERM solution predicts every test data correctly, i.e., if $L_n(\hat{f}_n) = 0$, then,

$$L(\hat{f}_n) \leq \frac{4\log(|\mathcal{F}|/\delta)}{n}$$

   with probability at least $1 - \delta$. This bound also holds with the relationship between X and Y is deterministic.
   Hint: Use the fact that, for any $a, b, c \in \mathbb{R}^+$ and $a \leq b + c\sqrt{a}$, then we have $a \leq b + c^2 + c\sqrt{b}$.

Lemma 3.4 (Berinstein's inequality)**.** *If $U_1, ..., U_n$ are n i.i.d. Bernoulli random variables with parameter p, then,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} U_i < p - \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2p + 2\epsilon/3}\right) \tag{2}$$

3.2 Generalization bound for hard SVM using Rademacher complexity

Here we prove generalization bound for hard SVM. [‡]. We will resort to Thereom 3.2 whcih contains the generalization bounds based on the definition of the Rademacher average. For SVM, the hypothesis space is a class of ilnear predictors:

$$\mathcal{H} = \left\{\langle \mathbf{w}, \mathbf{x}\rangle : \forall \mathbf{w} \in \mathbb{R}^d\right\}$$

with hinge loss $\ell(\mathbf{x}, y; \mathbf{w}) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}$ as the loss functin. Define

$$\mathcal{F} = \ell \circ \mathcal{H} \circ S = \{\ell(\mathbf{x}_1, y_1; \mathbf{w}), \ell(\mathbf{x}_2, y_2; \mathbf{w}), \ldots, \ell(\mathbf{x}_n, y_n; \mathbf{w})\}$$

since the hinge loss is 1-Lipchitz, and assuming that $\|\mathbf{x}\| \leq R, \|\mathbf{w}\| \leq B$, using Lemma 2.5 we have:

$$\mathfrak{R}_n(\mathcal{F}) \leq BR/\sqrt{n}$$

In general for any ρ-Lipchitz function, $\mathfrak{R}_n(\mathcal{F}) \leq \rho BR/\sqrt{n}$. Plugging this into Theorem 3.2 we get the following risk bound for SVM:

$$\mathbb{E}f(X) \leq \mathbb{E}_n f(X) + 2BR/\sqrt{n} + \sqrt{\frac{\log 1/\delta}{2n}}, \text{ with probability at least } 1 - \delta$$

So how should we interpret this? Suppose we make the assumption that we know the minimize of the empirical risk, which we denote with $\mathbf{w}^*$ which has zero empirical risk. Also $B = \|\mathbf{w}^*\|$, $\mathcal{H}$ can simply be the set of linear classifier which have norm smaller than B. Then the risk bound can be refined to

$$\mathbb{E}f(X) \leq \hat{L} + \frac{2R\|\mathbf{w}^*\|}{\sqrt{n}} + (1 + R\|\mathbf{w}^*\|)\sqrt{\frac{\log 1/\delta}{2n}}, \text{ with probability at least } 1 - \delta$$

And note that $\mathcal{F}$ is $(1+B\|\mathbf{w}^*\|)$-Lipchitz. With risk bound, one can show that the sample complexity of hard-SVM $\frac{R^2\|\mathbf{w}^*\|^2}{\epsilon^2}$.

In practice $\mathbf{w}^*$ is not known. One way to fix this, is to use the doubling trick on the weight vector size bound B. Suppose $B_i = 2^i$, $\mathcal{H}$ be all the linear models with weight norm less than B_i, $\delta_i = 2/$. For each i we can write an inequaliy for the risk. A union bound over all of the inequalities would give a unified bound which holds for all $\mathbf{w}$s.

4 Glivenko-Cantelli Theorem

The Glivenko-Cantelli guarantees uniform convergence bounds on empirical risk of the distributions. Our characterization of GC is based on Rademacher and Finite Class lemma, though this is not the only way to derive these results. First we introduce the finite class lemma which is a tool for bounding Rademacher averages.

[‡]Details on basic formulations here: http://web.engr.illinois.edu/~khashab2/learn/svm.pdf

Lemma 4.1 (Finite Class Lemma (Massart)). *Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^n$ and $\{\sigma_i\}_{i=1}^m$ independent Rademacher random variables, and $L = \sup_{a\in A} \mathcal{A}$,*

$$R_n(\mathcal{A}) = \frac{1}{n}\mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^m \sigma_i x_i\right] \leq \frac{2L\sqrt{\log|\mathcal{A}|}}{n}$$

Proof. Define,

$$\mu = \mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^m \sigma_i x_i\right] = m \times R_n(\mathcal{A})$$

For any $\lambda \in \mathbb{R}^+$,

$$\begin{aligned} e^{\lambda\mu} &\leq \mathbb{E}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}} \sum_{i=1}^m \sigma x_i\right)\right] = \mathbb{E}\left[\sup_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^m \sigma x_i\right)\right] \leq \mathbb{E}\left[\sum_{a\in\mathcal{A}} \exp\left(\lambda \sum_{i=1}^m \sigma x_i\right)\right] \\ &= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^m \sigma x_i\right)\right] = \sum_{a\in\mathcal{A}} \prod_{i=1}^m \mathbb{E}\left[\exp\left(\lambda\sigma x_i\right)\right] = \sum_{a\in\mathcal{A}} \prod_{i=1}^m \frac{\exp\left(-\lambda x_i\right) + \exp\left(\lambda x_i\right)}{2} \\ &\leq \sum_{a\in\mathcal{A}} \prod_{i=1}^m \exp\left(\lambda^2 {x_i}^2/2\right) \leq \sum_{a\in\mathcal{A}} \prod_{i=1}^m \exp\left(\lambda^2 L^2/2\right) \leq |\mathcal{A}| \prod_{i=1}^m \exp\left(\lambda^2 L^2/2\right) \end{aligned}$$

■

$$\Rightarrow \mu \leq \frac{\ln|\mathcal{A}|}{\lambda} + \frac{\lambda L^2}{2}.$$

Set $\lambda = \sqrt{2\frac{\ln|\mathcal{A}|}{L^2}}$, and we will have, $\mu \leq L\sqrt{2\ln|\mathcal{A}|}$

[More details: TBW]

The finite class lemma could be generalized to the class of binary-valued functions. Now define $\mathcal{F}$ be class of binary valued functions,

$$\mathcal{F} = \{f : \mathsf{Z} \to \{0, 1\}\}.$$

In other words, given random samples $\{Z_i\}_{i=1}^n$, and $\mathcal{F}(Z^n) \triangleq \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}$,

We generalize the bound using the Rademacher bound for this class of functions,

Lemma 4.2 (Rademacher bound for binary-valued functions). *For class of binary-valued functions $\mathcal{F}$,*

$$R_n\left(\mathcal{F}\left(Z^n\right)\right) \leq 2\sqrt{\frac{\log|\mathcal{F}(Z^n)|}{n}}$$

Proof. Proof in the Section 7. ■

Theorem 4.3 (Glivenko-Cantelli). *Let,*

$$F_n(x) \triangleq \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}$$

if $n \to \infty$, *then*

$$\sup_x |F_n(x) - F(x)| \stackrel{a.s}{\to} 0$$

for n *big enough.*

Proof. The proof consists of two main pars. First using the Rademacher for bounding the risk, and the second, using the Finite-Class lemma for bounding the Rademacher average. [More details for later] ■

5 Bibliographical notes

The first use of Rademacher complexity for risk bounds is probably due to [1, 2].

References

[1] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[2] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50, 2002.

6 Appendix: Union bound for risk

Let's assume we have proven the following bound for any $f \in \mathcal{F}$,

$$p(L(f) - L_n(f) \geq a(\delta)) \leq \delta, \quad \text{for any } f \in \mathcal{F}$$

which is equivalent to,

$$L_n(f) \geq L(f) + b(\delta) \quad \text{with probability at least } 1 - \delta \tag{3}$$

for some values a, b (functions of parameters). Then,

$$p(\exists f \in \mathcal{F} \wedge L_n(f) = 0 \wedge L(f) \geq a) \leq |\mathcal{F}|\delta$$

or, equivalently,

$$L_n(f) \geq L(f) + b(\delta/|\mathcal{F}|) \quad \text{with probability at least } 1 - \delta$$

Proof.

$$\begin{aligned} p(\exists f \in \mathcal{F} \wedge L_n(f) = 0 \wedge L(f) \geq a) &\leq p\left(\cup_{f \in \mathcal{F}} \left(L_n(f) = 0 \wedge L(f) \geq a\right)\right) \\ &\leq \sum_{f \in \mathcal{F}} p\left(\left(L_n(f) = 0 \wedge L(f) \geq a\right)\right) \\ &\leq |\mathcal{F}|\delta \end{aligned}$$

Now define $\delta' = \frac{\delta}{|\mathcal{F}|}$, and then using 3 we have

$$L_n(f) \geq L(f) + b(\delta') = L(f) + b(\delta/|\mathcal{F}|) \quad \text{with probability at least } 1 - \delta$$

which proves our desired statement. ■

7 Proofs

7.1 Proof of lemma 4.2

Proof. Since each f is a binary-valued function, $\mathcal{F} \subset \{0,1\}^n$. For any set of samples $\{Z_i\}_{i=1}^n$, and any function $f \in \mathcal{F}$, we know,

$$\sqrt{\sum_{i=1}^{n} |f(Z_i)|} \leq \sqrt{\sum_{i=1}^{n} 1} = \sqrt{n}$$

For a fixed set of random samples,$\{Z_i\}_{i=1}^n$, the set $\mathcal{F}(Z^n) \triangleq \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}$ is equivalent to the set $\mathcal{A}$, in Lemma 4.1, as $N = |\mathcal{F}(Z^n)| \leq 2^n$ and $L = \sqrt{n}$. As such,

$$R_n\left(\mathcal{F}\left(Z^n\right)\right) \leq 2\sqrt{\frac{\log|\mathcal{F}(Z^n)|}{n}}$$

■

8 Answers

Here answers to some of the questions are included. The answers are mostly by the authors, and might be buggy. Therefore, read cautiously!

8.1 Answer to example 3.3

8.1.1 First part :

We first use the Bernstein's inequality and simplify it. Consider the Equation 2 and take $\delta = \exp\left(-\frac{n\epsilon^2}{2p+2\epsilon/3}\right)$. Then,

$$\Rightarrow n\epsilon^2 - \left(\frac{2}{3}\ln\frac{1}{\delta}\right)\epsilon - 2p\ln\frac{1}{\delta} = 0$$

$$\Rightarrow \epsilon = \frac{\frac{2}{3}\ln\frac{1}{\delta} \pm \sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8np\ln\frac{1}{\delta}}}{2n}$$

Based on the assumption of the inequality the $\epsilon \geq 0$ and we can choose the value with the + sign in the about equation. Using this simplification, we can rewrite the Bernstein inequality in the following equivalent form:

$$EU \leq \frac{1}{n}\sum_{i=1}^{n} U_i + \frac{\frac{2}{3}\ln\frac{1}{\delta} + \sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8np\ln\frac{1}{\delta}}}{2n}, \quad \text{probability at least } 1-\delta$$

Now, for a specific $f \in \mathcal{F}$, we can consider $U_i = \mathbf{1}\{y_i \neq f(x_i)\}$ as a Bernoulli distribution, with the probability of success defined by $p = EU = L(f)$. The empirical estimation is the Bernoulli distribution is

$$\frac{1}{n}\sum_{i=1}^{n} U_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \neq f(X_i)\} = L_n(f).$$

This we can rewrite the bound as:

$$L(f) \leq L_n(f) + \frac{\ln\frac{1}{\delta}}{3n} + \frac{\sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8nL(f)\ln\frac{1}{\delta}}}{2n}, \quad \text{probability at least } 1-\delta$$

Now we use the fact that, $\sqrt{a}+\sqrt{b} \geq \sqrt{a+b}$

$$\begin{aligned}
L(f) &\leq L_n(f) + \frac{\ln\frac{1}{\delta}}{3n} + \frac{\sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2 + 8nL(f)\ln\frac{1}{\delta}}}{2n} \\
&\leq L_n(f) + \frac{\ln\frac{1}{\delta}}{3n} + \frac{\sqrt{\left(\frac{2}{3}\ln\frac{1}{\delta}\right)^2} + \sqrt{8nL(f)\ln\frac{1}{\delta}}}{2n} \\
&\leq L_n(f) + \frac{2\ln\frac{1}{\delta}}{3n} + \sqrt{\frac{2L(f)\ln\frac{1}{\delta}}{n}}, \quad \text{probability at least } 1-\delta
\end{aligned}$$

Which proves the desired result.

8.1.2 Second part :

We use the hint on the bound which we found in the previous part, in Equation 1, with the following definitions:

$$a = L(f), \quad b = L_n(f) + \frac{2\log(1/\delta)}{3n}, \quad c = \sqrt{\frac{2\log(1/\delta)}{n}}$$

This would imply the following inequality:

$$L(f) \leq L_n(f) + \frac{2\log(1/\delta)}{3n} + \left(\sqrt{\frac{2\log(1/\delta)}{n}}\right)^2 + \left(\sqrt{\frac{2\log(1/\delta)}{n}}\right)\sqrt{L_n(f) + \frac{2\log(1/\delta)}{3n}}$$

$$\Rightarrow L(f) \leq L_n(f) + \frac{8\log(1/\delta)}{3n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n} + \frac{4}{3}\left(\frac{\log(1/\delta)}{n}\right)^2}$$

We use the inequality $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$,

$$\begin{aligned}
L(f) \leq & L_n(f) + \frac{8\log(1/\delta)}{3n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n} + \frac{4}{3}\left(\frac{\log(1/\delta)}{n}\right)^2} \\
\leq & L_n(f) + \frac{8\log(1/\delta)}{3n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} + \sqrt{\frac{4}{3}\left(\frac{\log(1/\delta)}{n}\right)^2} \\
\leq & L_n(f) + \left(\frac{2}{\sqrt{3}} + \frac{8}{3}\right)\frac{\log(1/\delta)}{n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} \\
= & L_n(f) + \frac{3.83\log(1/\delta)}{n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} \\
\leq & L_n(f) + \frac{4\log(1/\delta)}{n} + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}}
\end{aligned}$$

Which proves the desired result. Now using this bound, we prove the last part of the question. Before that we state the union bound for risk.

Since this bound holds for any $f \in \mathcal{F}$, this also holds for $\hat{f} \in \mathcal{F}$. Based on the assumption of the question, the risk for this function is zero. For a fixed $\hat{f} \in \mathcal{F}$, if we have $L_n(\hat{f}) = 0$,

$$L(f) \leq \frac{4\log(1/\delta)}{n}$$

since $\hat{f}$ is not known a priori and it can any function in the class of functions $\mathcal{F}$, we need to use the union bound, as in Equation 3:

$$L(f) \leq \frac{4\log(|\mathcal{F}|/\delta)}{n}$$