Machine learning theory

Hypothesis complexity measures

Hamid Beigy

Sharif University of Technology

April 27, 2020

Table of contents

Introduction

1. In last session, we showed that finite hypothesis class H is learnable in PAC model with the following sample complexity.
$$m \geq \frac{1}{\epsilon}\left[\log|H| + \log\frac{1}{\delta}\right]$$
where $|H|$ is the length of description of hypothesis class H.
2. In last session, we showed that finite hypothesis class H is learnable in Agnostic PAC model with the following sample complexity.
$$m \geq \frac{1}{\epsilon^2}\left[\log|H| + \log\frac{1}{\delta}\right]$$

1. How can we use these bounds for infinite hypothesis class H? (via discretization)
   - Let every $h \in H$ is parametrized by k parameters.
   - Let each parameter is represented by b bits in computer.
   - Then every $h \in H$ can be represented by 2^{kb} bits.
   - The bound for PAC model is

$$m \geq \frac{1}{\epsilon}\left[kb + \log\frac{1}{\delta}\right]$$

$$m = O\left(\frac{1}{\epsilon}\left[k + \log\frac{1}{\delta}\right]\right)$$

   - The bound for Agnostic PAC model is

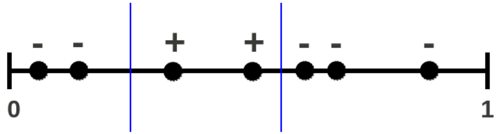
$$m \geq \frac{1}{\epsilon^2}\left[kb + \log\frac{1}{\delta}\right]$$

$$m = O\left(\frac{1}{\epsilon^2}\left[k + \log\frac{1}{\delta}\right]\right)$$

2. The above bounds show that the sample complexity is proportional to the number of parameters of hypothesis.
3. It will be shown that some hypothesis classes have one parameter but they aren't learnable in these model.
4. This shows that $|H|$ is not suitable measure of richness of a hypothesis class.

Growth function

1. Let $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ be training set and H be hypothesis class.



2. To define growth function, let us to define dichotomy.

Definition (Dichotomy)

Let $x_1, \dots, x_m \in \mathcal{X}$, the dichotomies generated by H on these points are defined by

$$H(x_1, \dots, x_m) = \{(h(x_1), \dots, h(x_m)) \mid h \in H\}$$

Definition (Growth function)

The growth function counts the maximum number of dichotomies on any m points.

$$\Pi_H(m) = \max_{x_1, \dots, x_m \in \mathcal{X}} |H(x_1, \dots, x_m)|$$

3. Thus, $\Pi_H(m)$ is the maximum number of ways m points can be classified using H.

1. Considering one-dimensional threshold function H with the following training set.

$$X = \{1, 2, 3, 4, 5, 6\}$$

2. We have 7 distinct hypothesis for this hypothesis class.

Lemma (Growth function for one-dimensional threshold function)

Let $X = \{x_1, x_2, \ldots, x_m\}$ be the training set. Then we have

$$\Pi_H(m) = m + 1$$

3. Let H be set of intervals. What is the growth function for this hypothesis class?

Theorem (Upper bound for growth function)

Let H be the hypothesis class, then for any training set of size m, the following inequality holds.

$$\Pi_H(m) \leq 2^m.$$

Theorem (For realizable case)

Let H be the hypothesis class. For all $h \in H$ and for all $\delta > 0$, with the probability of at least $1-\delta$, the following inequality holds.

$$\mathbf{R}(h) = O\left(\frac{\ln \Pi_H(2m) + \ln \frac{2}{\delta}}{m}\right).$$

Theorem (For unrealizable case)

Let H be the hypothesis class. For all $h \in H$ and for all $\delta > 0$, with the probability of at least $1-\delta$, the following inequality holds.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2\ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Homework: Prove the above theorems.

VC-dimension

1. We showed that $\Pi_H(m) \leq 2^m$. But in most cases, this bound is not tight.
2. If we choose the size of the training set such that

$$\Pi_H(m) \leq 2^m,$$

   the hypothesis class H can classify all different labeling of S.
3. This leads to the definition of new complexity measure, VC-dimension.

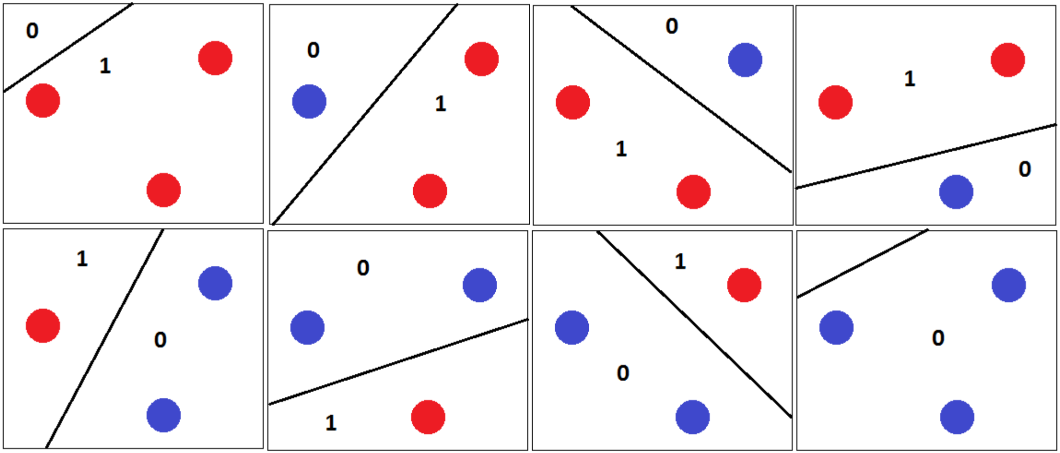
Definition (Dichotomy)

A dichotomy of a set S is a partition of S into two disjoint subsets.

Definition (Shattering)

A set S is shattered by hypothesis space H iff for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

Shattering



1. Formally, H shatters S if $\Pi_H(m) = 2^m$.

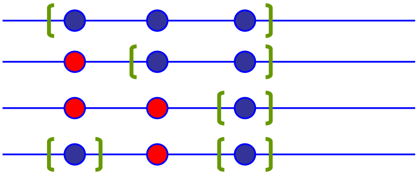
Definition (VC-dimension)

The Vapnik-Chervonenkis (VC) dimension of H, denoted as $VC(H)$, is the cardinality d of the largest set S shattered by H. If arbitrarily large finite sets can be shattered by H, then $VC(H) = \infty$ or

$$VC(H) = \max \{m \mid \Pi_H(m) = 2^m\}$$

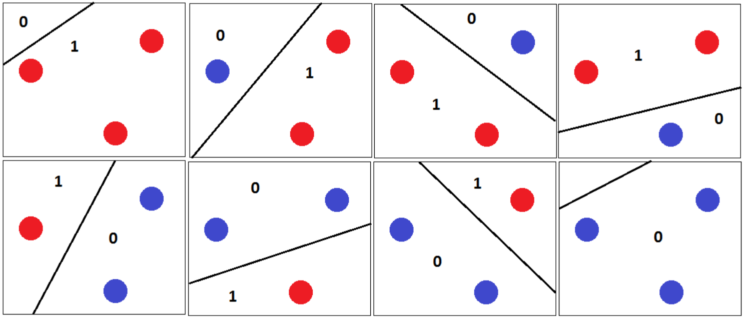
2. The definition of $VC(H)$ is: if there exists a set of d points that can be shattered by the classifier and there is no set of $d+1$ points that can be shattered by the classifier, then $VC(H) = d$.
3. The definition does not say: if any set of d points can be shattered by the classifier.

1. Let H be the set of intervals on the real line such that $h(x) = 1$ iff x is in the interval.
2. How many points can be shattered by H?

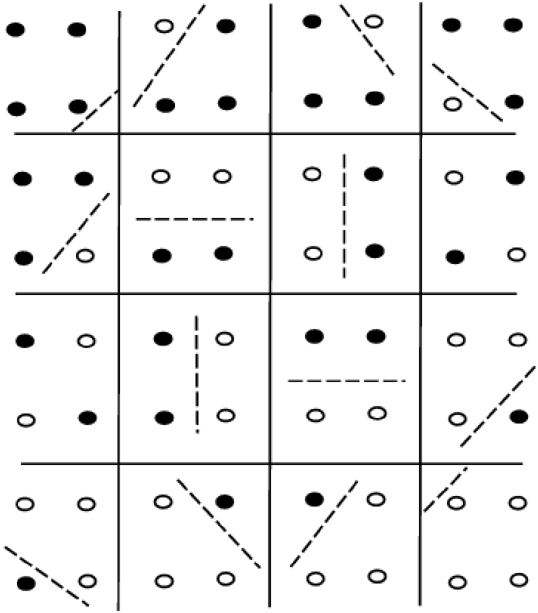


3. It can shatter 2 points. It cannot shatter 3 points. Thus $VC(H) = 2$.

1. Let H be the set of linear classifiers on the two-dimensional space.
2. How many points can be shattered by H?

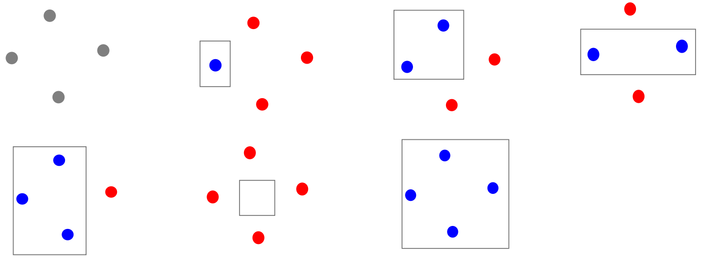


1. It can shatter 3 points. It cannot shatter 4 points. Thus $VC(H) = 3$.

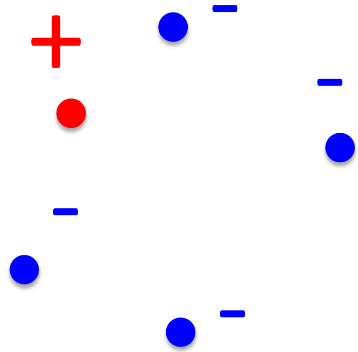


2. For d-dimensional linear classifier, we have $VC(H) = d + 1$

1. Let H be the set of axis aligned rectangle hypothesis class on the two-dimensional space.
2. How many points can be shattered by H?



1. It can shatter 4 points. It cannot shatter 5 points. Thus $VC(H) = 4$.



Theorem (VC-dimension of finite hypothesis classes)

For every finite hypothesis classes H, we have $VC(H) \leq \log|H|$.

Proof.

- Let $VC(H) = d$. Hence, we have
$$\Pi_H(d) = 2^d.$$
- In other hand, for every set with size $m > 1$, we have $\Pi_H(m) \leq |H|$.
- Hence, we have $2^d = \Pi_H(d) \leq |H|$.
- By taking log from both sides of $2^d = \Pi_H(d) \leq |H|$, the proof will be completed.

$\square$

Example (VC of conjunction)

Let H be the conjunction of at most n literals. Then, we have

$$n \leq VC(H) \leq n \log 3.$$

Lemma (Sauer-Shelah Lemma)

Let H be a hypothesis classes with $VC(H) = d$, then for $m \in \mathbb{N}$, we have

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

Homework: Prove this Lemma by using induction on $m + d$.

Corollary

Let H be a hypothesis classes with $VC(H) = d$, then for $m > d > 1$, we have

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d$$

Proof.

From Sauer-Shelah Lemma, we have

$$
\begin{aligned}
\Pi_H(m) &\leq \sum_{i=0}^{d} \binom{m}{i} \\
&\leq \sum_{i=0}^{d} \binom{m}{i} \underbrace{\left(\frac{m}{d}\right)^{d-i}}_{>1} \\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \quad \text{Using binomial distribution} \\
&= \left(\frac{m}{d}\right)^{d} \left(1 + \frac{d}{m}\right)^{m} \quad \text{Using inequality } (1-x) \leq e^{-x} \\
&\leq \left(\frac{m}{d}\right)^{d} \left(e^{d/m}\right)^{m} \\
&= \left(\frac{m}{d}\right)^{d} e^{d} = \left(\frac{me}{d}\right)^{d}
\end{aligned}
$$

□

Theorem (Generalization bound based on VC-dimension)

Let H be a hypothesis class with $VC(H) = d$, then for every $h \in H$ and every $\delta > 0$, with probability of at least $1 - \delta$, we have

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

This bound can be extended to nonrealizable case.

Proof.

From growth function, we have

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2 \ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

From Sauer-Shelah Lemma, we have

$$\begin{aligned}
\mathbf{R}(h) &\leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2 \ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \\
&\leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2 \ln \left(\frac{me}{d}\right)^d}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \\
&\leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2d \ln \frac{me}{d}}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.
\end{aligned}$$

$\square$

1. We showed that with probability at least $1-\delta$, and for all $h \in H$, if h is consistent, then

$$\mathbf{R}(h) = O\left(\frac{\ln \Pi_H(2m) + \ln\left(\frac{1}{\delta}\right)}{m}\right) \tag{1}$$

2. We also show that for all $m > d \geq 1$ and $VC(H) = d$, we have

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d$$

3. The above inequality says that
   - for $m \leq d$, $\Pi_H(m) = 2^m$. In this case, bound given in (1) is meaning less.
   - for $m \geq d$, $\Pi_H(m) = O(m^d)$. In this case, we have

$$\ln \Pi_H(m) = O(d \ln m)$$

   Hence, this bound is proportional to d and $\frac{1}{m}$

Theorem (Bound based on VC-dimension)

Let $VC(H) = d$, then for all consistent $h \in H$, with probability at least $1 - \delta$, we have

$$\mathbf{R}(h) = O\left(\frac{d \log m + \log \frac{1}{m}}{m}\right)$$

$$m = O\left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon}\right)$$

Example (One dimensional threshold function)

For one-dimensional threshold function, we showed $VC(H) = 1$ and $m \geq \frac{1}{\epsilon} \log \frac{2}{\delta}$. Using the above Theorem we have

$$m = O\left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{1}{\epsilon} \log \frac{1}{\epsilon}\right).$$

This shows that this bound is not bad.

Example (Axis aligned rectangle)

For axis aligned rectangle, we showed $VC(H) = 4$ and $m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$. Using the above Theorem we have

$$m = O\left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{4}{\epsilon} \log \frac{1}{\epsilon}\right).$$

The above two examples show that the sample complexity increases linearly with the number of parameters of hypothesis.

Example (Hypothesis class of $\operatorname{sgn}(\sin(\theta x))$)

We can show that $VC(H) = \infty$ but it has only one parameter.

Radamacher complexity

1. We use the following problem setting
   - The training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.
   - The label set $\mathcal{Y} = \{-1, +1\}$.
   - The hypothesis $h : \mathcal{X} \mapsto \{-1, +1\}$.
   - The empirical error $\hat{\mathbf{R}}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left[h(x_i) \neq y_i\right]$.
2. An alternative definition of empirical error is

$$
\begin{aligned}
\hat{\mathbf{R}}(h) &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left[h(x_i) \neq y_i\right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \begin{cases} 1 & \text{if } (h(x_i), y_i) = (+1, -1) \text{ or } (h(x_i), y_i) = (-1, +1) \\ 0 & \text{if } (h(x_i), y_i) = (+1, +1) \text{ or } (h(x_i), y_i) = (-1, -1) \end{cases} \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2} \\
&= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(x_i)
\end{aligned}
$$

1. The term $\frac{1}{2m}\sum_{i=1}^{m} y_i h(x_i)$ can be interpreted as correlation between the true and the predicted labels.
2. To find a hypothesis that minimizes the empirical error, we find a hypothesis that maximizes the correlation.

$$h = \operatorname*{argmax}_{h \in H} \frac{1}{m}\sum_{i=1}^{m} y_i h(x_i).$$

3. If we replace the true label with Radamacher random variables

$$\sigma_i = \begin{cases} +1 & \text{With probability of } \frac{1}{2} \\ -1 & \text{With probability of } \frac{1}{2} \end{cases}$$

   we obtain

$$h = \operatorname*{argmax}_{h \in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i).$$

4. Instead of selecting the hypothesis in H that correlates best with the labels, this now selects the hypothesis $h \in H$ that correlates best with the random noise variables σ_i.

1. Hypothesis h is dependent on the random variables σ_i. To measure how well H can correlate with random noise, we take the expectation of this correlation over the random variables σ_i and find

$$\mathbb{E}_{\sigma}\left[\max_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right]$$

2. This intuitively measures the expressiveness of H.
3. We can bound this expression using two extreme cases
   - When $|H| = 1$, the above expectation becomes zero.
   - When $|H| = 2^m$, the above expectation becomes one, because there always exists a hypothesis matching any set of σ_i's.

1. Instead of working with hypotheses $h : \mathcal{X} \mapsto \{-1, +1\}$, let's generalize our class of functions to the set of all real-valued functions.
2. Replace H with $\mathcal{F}$, which we define to be any family of functions $f : \mathcal{Z} \mapsto \mathbb{R}$.
3. Given sample $S = (z_1, \ldots, z_m)$ with $z_i \in \mathcal{Z}$, if we apply our expression from above to $\mathcal{F}$.

Definition (Empirical Rademacher complexity)

The empirical Rademacher complexity of a family of functions $\mathcal{F}$ with respect to a sample S is defined as

$$\hat{\mathcal{R}}_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i)\right]$$

4. This expression measures how well, on average, the function class $\mathcal{F}$ correlates with random noise over the sample S.

1. However, often we want to measure the correlation of $\mathcal{F}$ with respect to a distribution $\mathcal{D}$ over $\mathcal{X}$, rather than with respect to a sample S over $\mathcal{X}$.
2. To find this, we take the expectation of $\hat{\mathcal{R}}_S(\mathcal{F})$ over all samples of size m drawn according to $\mathcal{D}$.

Definition (Rademacher complexity/Expected Rademacher complexity)

The Rademacher complexity of a family of functions $\mathcal{F}$ with respect to a sample S is defined as

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

1. We first prove the following theorem as a general tools.

Theorem

Let $\mathcal{F}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$, and let sample $S = (z_1, \ldots, z_m)$ where $z_i \sim \mathcal{D}$ for some distribution $\mathcal{D}$ over $\mathcal{Z}$. Define $\hat{\mathbb{E}}_S[f] = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$, then with probability of at least $1-\delta$ for all $f \in \mathcal{F}$, we have

$$\begin{aligned}
\mathbb{E}[f] &\leq \hat{\mathbb{E}}_S[f] + 2\mathcal{R}_m(\mathcal{F}) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \\
\mathbb{E}[f] &\leq \hat{\mathbb{E}}_S[f] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right)
\end{aligned}$$

Proof:

We derive a bound for $\mathbb{E}[f] - \hat{\mathbb{E}}_S[f]$ for all $f \in \mathcal{F}$ or equivalently, bound $\sup_{f\in\mathcal{F}} \left\{\mathbb{E}[f] - \hat{\mathbb{E}}_S[f]\right\}$.

Note that this expression is a random variable that depends on S. So we want to bound the following random variable: $\phi(S) = \sup_{f\in\mathcal{F}} \left\{\mathbb{E}[f] - \hat{\mathbb{E}}_S[f]\right\}$.

Step 1: We show, with probability of at least $1-\delta$, inequality $\phi(S) \leq \mathbb{E}_S[\phi(S)] + \sqrt{\frac{\ln\frac{2}{\delta}}{2m}}$ holds.
This step allows us to go from working with $\phi(S)$ to working with $\mathbb{E}_S[\phi(S)]$.
Let $S = (z_1, z_2, \ldots, z_i, \ldots, z_m)$ and $S' = (z_1, z_2, \ldots, z_i', \ldots, z_m)$ be two training sets with only one different element.
Recall that McDiarmid's inequality states that, if for all i, we have

$$|f(z_1, z_2, \ldots, z_i, \ldots, z_m) - f(z_1, z_2, \ldots, z_i', \ldots, z_m)| \leq c_i$$

then the following inequality holds

$$\mathbb{P}\left[|f(S) - f(S')| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right)$$

From the definition of $\phi(S)$ we have

$$\begin{aligned}\phi(S) &= \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \right\} \\ &= \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}[f] - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \right\}.\end{aligned}$$

Since $f(z_i) \in [0, 1]$ for all i, changing any one example z_i to z_i' in the training set S will change $\frac{1}{m}\sum_{i=1}^{m} f(z_i)$ by at most $\frac{1}{m}$. Thus this changing of any one example affects $\phi(S)$ by at most this amount, implying that $|\phi(S) - \phi(S')| \leq \frac{1}{m}$.
This fits McDiarmid's inequality with $c_i = \frac{1}{m}$, so we can apply this inequality and arrive at the bound shown.

$$\begin{aligned}\mathbb{P}\left[|\phi(S) - E_S[\phi(S)]| \geq \epsilon\right] &\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\right) \\ &= 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}\left(\frac{1}{m}\right)^2}\right) \\ &= 2\exp\left(-2(m\epsilon)^2\right).\end{aligned}$$

If we let $\epsilon = \sqrt{\frac{\log 2\delta}{2m}}$, we obtain

$$\phi(S) \leq \mathbb{E}_S[\phi(S)] + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Step 2: Define a sample $S' = (z'_1, \ldots, z'_m)$, $z'_i \sim \mathcal{D}$, we show that $\mathbb{E}_S\left[\phi(S)\right] \leq \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathbb{E}}_{S'}\left[f\right] - \hat{\mathbb{E}}_S\left[f\right]\right)\right]$.

$$
\begin{aligned}
\mathop{\mathbb{E}}_S\left[\phi(S)\right] &= \mathop{\mathbb{E}}_S\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[f\right] - \hat{\mathop{\mathbb{E}}_S}\left[f\right]\right)\right] \\
&= \mathop{\mathbb{E}}_S\left[\sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{S'}\left[\hat{\mathop{\mathbb{E}}_{S'}}\left[f\right]\right] - \hat{\mathop{\mathbb{E}}_S}\left[f\right]\right] && \text{From definition of Radamacher complexity.} \\
&= \mathop{\mathbb{E}}_S\left[\sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{S'}\left[\hat{\mathop{\mathbb{E}}_{S'}}\left[f\right] - \hat{\mathop{\mathbb{E}}_S}\left[f\right]\right]\right] \\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathop{\mathbb{E}}_{S'}}\left[f\right] - \hat{\mathop{\mathbb{E}}_S}\left[f\right]\right)\right] && \text{Moving } S' \text{ outside of sup.}
\end{aligned}
$$

The last be done since the expectation of a max over some function is at least the max of that expectation over that function.

Step 3: We show $\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathbb{E}}_{S'}[f]-\hat{\mathbb{E}}_{S}[f]\right)\right]=\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\sum_i\sigma_i\left(f(z_i')-f(z_i)\right)\right]$, where $z_i'\sim\mathcal{D}$.

$$\mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathbb{E}}_{S'}[f]-\hat{\mathbb{E}}_{S}[f]\right)\right]=\mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\left(\sum_i f(z_i')-\sum_i f(z_i)\right)\right]$$
$$=\mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i\left(f(z_i')-f(z_i)\right)\right].$$

By adding Radamacher random variables, we obtain

$$\mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathbb{E}}_{S'}[f]-\hat{\mathbb{E}}_{S}[f]\right)\right]=\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i\sigma_i\left(f(z_i')-f(z_i)\right)\right]$$

Step 4: We show $\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\sum_i\sigma_i\left(f(z_i')-f(z_i)\right)\right]\leq 2\mathcal{R}_m(\mathcal{F})$.

$$\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i\sigma_i\left(f(z_i')-f(z_i)\right)\right]\leq\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i\sigma_i f(z_i')+\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i(-\sigma_i)f(z_i)\right]$$

This inequality was obtained from inequality $\sup(a+b)\leq\sup(a)+\sup(b)$.

$$\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i\sigma_i\left(f(z_i')-f(z_i)\right)\right]\leq\mathop{\mathbb{E}}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i\sigma_i f(z_i')\right]+\mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_i(-\sigma_i)f(z_i)\right]$$
$$=\mathcal{R}_m(\mathcal{F})+\mathcal{R}_m(\mathcal{F}).$$

The last inequality was obtained because $-\sigma_i$ has the same distribution as σ_i.

Conclusion: By combining all the pieces together, the theorem will be proved.
The second inequality can be proved in the same way.

1. The following result relates the empirical Rademacher complexities of a hypothesis set H and to the family of loss functions $\mathcal{F}$ associated to H in the case of binary loss (zero-one loss).

Theorem

Let H be a family of functions taking values in $\{-1, +1\}$ and let $\mathcal{F}$ be the family of loss functions associated to H for the zero-one loss: $f_h(x, y) = \mathbb{I}[h(x) \neq y]$. For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, +1\}$, let $S_{\mathcal{X}}$ denote its projection over $\mathcal{X}$, i.e. $S_{\mathcal{X}} = (x_1, \ldots, x_m)$. Then,the following relation holds between the empirical Rademacher complexities of $\mathcal{F}$ and H:

$$\hat{\mathcal{R}}_S(\mathcal{F}_H) = \frac{1}{2}\hat{\mathcal{R}}_{S_{\mathcal{X}}}(H)$$

Proof.

For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, +1\}$, by definition, the empirical Rademacher complexity of G can be written as:

$$
\begin{aligned}
\hat{\mathcal{R}}_S(F_H) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f_h \in \mathcal{F}_H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f_h(x_i, y_i)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \left(\frac{1 - y_i h(x_i)}{2}\right)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h \in H} \frac{1}{2m}\sum_{i=1}^{m} \sigma_i + \sup_{h \in H} \frac{1}{2m}\sum_{i=1}^{m} (-y_i\sigma_i) h(x_i)\right] \\
&= \frac{1}{2m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{\sigma}[\sigma_i] + \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] \\
&= \frac{1}{2}\hat{\mathcal{R}}_{S_X}(H).
\end{aligned}
$$

$\square$

Relating different bounds

1. The following Theorem relates Rademacher complexity and the size of hypothesis space .

Theorem

For any hypothesis space $|H| < \infty$, the following inequality holds.

$$\hat{\mathcal{R}}_S(H) = \sqrt{\frac{2\ln|H|}{m}}$$

Lemma (Massart's Lemma)

Let $A \subseteq \mathbb{R}^m$ be a finite set of vectors with $\|\mathbf{a}\| \leq 1$ for all $\mathbf{a} \in A$. Then

$$\mathop{\mathbb{E}}_{\sigma}\left[\max_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right] \leq \sqrt{2\ln|A|},$$

where σ_i are independent Rademacher variables and $a_1, a_2, \ldots, a_m$ are components of vector $\mathbf{a}$.

Proof.

- Let us to define the space A as $A = \left\{ \frac{1}{\sqrt{m}} (h(x_1), h(x_2), \ldots, h(x_m)) \right\}$.
- Then $A \subseteq \mathbb{R}^m$ and for all $\mathbf{a} \in A$ we have $\|\mathbf{a}\| = 1$.
- From Rademacher complexity, we have

$$\begin{aligned}
\hat{\mathcal{R}}_S(H) &= \mathbb{E}_{\sigma}\left[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] \\
&= \mathbb{E}_{\sigma}\left[\sup_{\mathbf{a}\in A} \frac{\sqrt{m}}{m}\sum_{i=1}^{m} \sigma_i a_i\right] \\
&= \frac{1}{\sqrt{m}} \mathbb{E}_{\sigma}\left[\max_{\mathbf{a}\in A} \sum_{i=1}^{m} \sigma_i a_i\right] \\
&\le \frac{1}{\sqrt{m}} \sqrt{2\ln|A|} \\
&= \sqrt{\frac{2\ln|A|}{m}}.
\end{aligned}$$

- Since A is the set of classifiers for the set S, hence $A \subset H$ and $|A| \le |H|$.

□

1. The following Theorem relates Rademacher complexity and the Growth function .

Theorem

For any hypothesis space $|H|$, the following inequality holds.

$$\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{\ln \Pi_H(m)}{m}}.$$

Proof.

- We only need to consider behavior of hypotheses on training set S.
- Let $H' = \{$ one representative from H for each behaviors on $S\}$.
- Thus $H' \subset H$ and $|H'| = \Pi_H S \leq \Pi_H m \leq 2^m < \infty$.
- From definition of Rademacher complexity, we have

$$\hat{\mathcal{R}}_S(H) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right]$$

- Since, for every $h \in H$ that maximizes $\hat{\mathcal{R}}_S(H)$, there exists an $h' \in H'$ that results in the same value. Hence, we have

$$\begin{aligned}\hat{\mathcal{R}}_S(H) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h'\in H'} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h'(x_i)\right] \\ &= \hat{\mathcal{R}}_S(H').\end{aligned}$$

- This implies that the sup over H is no greater than the sup over H' and vice versa. Hence these two sup are equal and

$$\begin{aligned} \hat{\mathcal{R}}_S(H) &= \hat{\mathcal{R}}_S(H') \\ &\leq \sqrt{\frac{2\ln|H'|}{m}} \\ &= \sqrt{\frac{2\ln \Pi_H(S)}{m}} \end{aligned}$$

The following Theorem relates Rademacher complexity and VC dimension .

Theorem

Let $d = VC(H)$, then for $m \geq d \geq 1$, we have $\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{2d \ln\left(\frac{em}{d}\right)}{m}}$

Proof.

From Sauer Lemma, we have $\Pi_H(m) \leq \left(\frac{em}{d}\right)^d$ and using the previous Theorem, we have

$$\begin{aligned}
\hat{\mathcal{R}}_S(H) &\leq \sqrt{\frac{2\ln \Pi_H(m)}{m}} \\
&\leq \sqrt{\frac{2\ln\left(\frac{em}{d}\right)^d}{m}} \\
&= \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} \\
&= \sqrt{\frac{2\ln\left(\frac{em}{d}\right)}{\left(\frac{m}{d}\right)}}.
\end{aligned}$$

$\square$

Fundamental Theorem of Statistical Learning

Theorem (Fundamental Theorem of Statistical Learning)

Let H be hypothesis class from a domain $\mathcal{X}$ to $\{0,1\}$ and the loss function be the 0/1 loss. Then, the following are equivalent:

1. *H has uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for H.*
3. *H is agnostic PAC learnable.*
4. *H is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for H.*
6. *H has finite VC dimension.*

For the proof, please read section 6.4 of Ben-David book.

1. Chapter 3 of Mehryar Mohri and Afshin Rostamizadeh and Ameet Talwalkar Book[1].

[1] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Second Edition. MIT Press, 2018.

References

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Second Edition. MIT Press, 2018.

Questions?