Vapnik-Chervonenkis Dimension

Ruixin Guo

Department of Computer Science
Kent State University

Mar 10, 2023

Contents

Contents

Recall

In machine learning, we want to use a function f to approximate the target function F. We assume f belongs to a function class $\mathcal{F}$, and search for the best f in $\mathcal{F}$ that approximates F.

The error in machine learning can be separated into Approximation Error and Estimation Error. We want to find a bound for Estimation Error.

$$R^{\text{true}}(f_m) - R^* = \underbrace{[R^{\text{true}}(f^*) - R^*]}_{\text{Approximation Error}} + \underbrace{[R^{\text{true}}(f_m) - R^{\text{true}}(f^*)]}_{\text{Estimation Error}}$$

The estimation error can be bounded by

$$R^{\text{true}}(f_m) - R^{\text{true}}(f^*) \le 2 \sup_{f \in \mathcal{F}} |R^{\text{true}}(f) - R^{\text{emp}}(f)|$$

And we can bound $|R^{\text{true}}(f) - R^{\text{emp}}(f)|$ by concentration inequalities like Hoeffding Inequality.

Usually we consider one-side bound for $R^{\text{true}}(f) - R^{\text{emp}}(f)$. The two-side bound is $P[|R^{\text{true}}(f) - R^{\text{emp}}(f)| \le \epsilon] = 2P[R^{\text{true}}(f) - R^{\text{emp}}(f) \le \epsilon]$, because the Hoeffding Bound is symmetric.

Recall

For a single $f \in \mathcal{F}$ we can bound the $R^{\text{true}}(f) - R^{\text{emp}}(f)$ as

$$P\left[R^{\text{true}}(f) - R^{\text{emp}}(f) \leq \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

However this bound is not useful because it is only for a single f. Not all fs in $\mathcal{F}$ will satisfy this bound. When we search for f, we try to minimize $R^{\text{emp}}(f)$, this tends to make the gap $R^{\text{true}}(f) - R^{\text{emp}}(f)$ big. So we are more likely to find an f that not satisfies the bound.

The solution is to use the union bound, i.e., we use the bound for all fs in $\mathcal{F}$, not just a single one. If the size of $\mathcal{F}$ is finite, let's say $|\mathcal{F}| = N$, then we can write the union bound as:

$$\forall f \in \mathcal{F} \quad P\left[R^{\text{true}}(f) - R^{\text{emp}}(f) \leq \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

The union bound tells us when N is bigger, the estimation error is bigger.

Question: How do we find a bound for the estimation error when $|\mathcal{F}|$ is infinite?

Growth Function

When functions in $\mathcal{F}$ is uncountable, we use a countable measure for $\mathcal{F}$. The key idea is to group functions based on the sample.

Given the samples $z_1, z_2, ..., z_m$, consider the set

$$\mathcal{F}_{z_1,z_2,...,z_m} = \{f(z_1), f(z_2), ..., f(z_m) : f \in \mathcal{F}\}$$

For binary classification, $f(z) \in \{0, 1\}$, then $|\mathcal{F}_{z_1,z_2,...,z_m}| \leq 2^m$, which means the set is always finite.

Here we put the functions that generate the same classification result in a group, and $\mathcal{F}$ is partitioned into $|\mathcal{F}_{z_1,z_2,...,z_m}|$ disjoint groups. Now we consider the maximum number of groups as a measure for $\mathcal{F}$:

Definition (Growth Function): The growth function is the maximum number of ways into which m points can be classified by the function class:

$$S_{\mathcal{F}}(m) = \sup_{(z_1,z_2,...,z_m)} |\mathcal{F}_{z_1,z_2,...,z_m}|$$

Shattering and VC dimension

In order to figure out how to compute $S_{\mathcal{F}}(m)$, we need to use VC dimension.

Definition (Shattering): We say $\mathcal{F}$ shatters an m-point dataset if $S_{\mathcal{F}}(m) = 2^m$.

- This means there is a dataset of size m points such that $\mathcal{F}$ can generate any classification on these points.

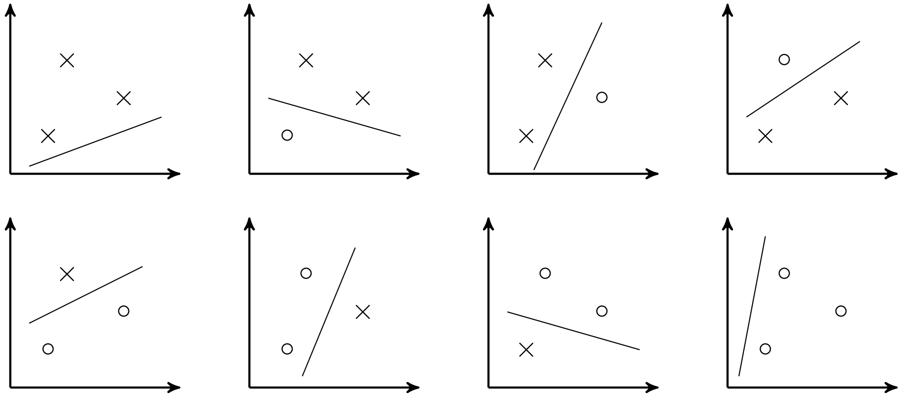
Definition (VC Dimension): The VC Dimension of a function class $\mathcal{F}$ is the largest h such that $S_{\mathcal{F}}(h) = 2^h$.

- The VC dimension of $\mathcal{F}$ is the maximum number of points that $\mathcal{F}$ can shatter.

Note that VC dimension is an attribute of the function class $\mathcal{F}$. Let h be the VC dimensions of $\mathcal{F}$ and m be the number of points of an dataset. If $m \leq h$, then $S_{\mathcal{F}}(m) = 2^m$; if $m > h$, then $S_{\mathcal{F}}(m) < 2^m$.

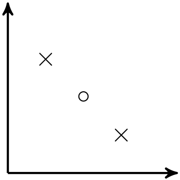
Example of VC Dimension

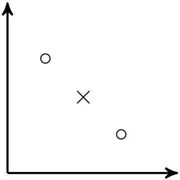
Example 1: When $\mathcal{F} = \{f(\boldsymbol{x}) = \mathbf{1}_{\boldsymbol{w}^T\boldsymbol{x}+b>0}, \boldsymbol{x} \in \mathbb{R}^2, \boldsymbol{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$, i.e., $\mathcal{F}$ is a set of all non-vertical lines in 2D space. Then the VC dimension of $\mathcal{F}$ is 3, because it can scatter at most 3 points:



Example of VC Dimension

One may argue that when the 3 points are on a line, the function class $\mathcal{F} = \{f(\boldsymbol{x}) = \mathbf{1}_{\boldsymbol{w}^T\boldsymbol{x}+b>0}, \boldsymbol{x} \in \mathbb{R}^2, \boldsymbol{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$ cannot shatter it. For example, we cannot find an $f \in \mathcal{F}$ that classifies the points into the following cases:





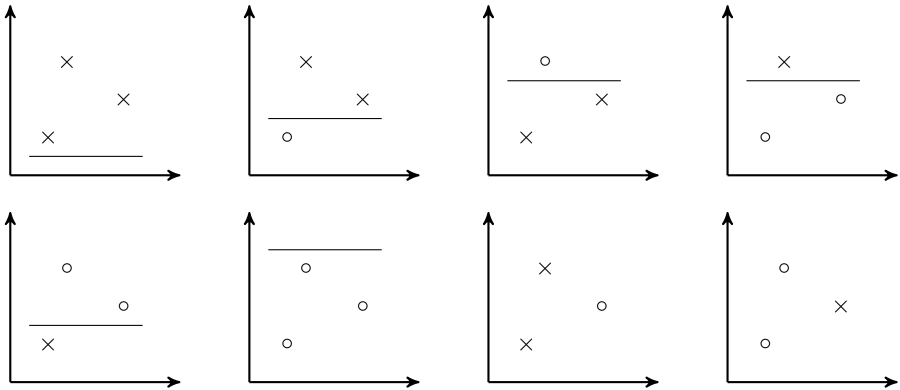
This is true. However, remember that VC dimension is data independent. If we can find at least one dataset of size m that $\mathcal{F}$ can shatter, then we say the VC dimension of $\mathcal{F}$ is m.

In Example 1 we have shown that $\mathcal{F}$ can shatter 3 points. However, for any 4 points dataset, $\mathcal{F}$ cannot shatter it (The Theorem on page 11 proves this). So the VC dimension of $\mathcal{F}$ cannot be 4.

Example of VC Dimension

Example 2: Consider the function class $\mathcal{F} = \{f(\boldsymbol{x}) = \mathbf{1}_{kx_2 \leq \theta}, \boldsymbol{x} = [x_1, x_2]^T \in \mathbb{R}^2, k \in \mathbb{R}, \theta \in \mathbb{R}\}$. That is, $\mathcal{F}$ is the set of all horizontal line classifiers. It can shatter at most 2 points but cannot shatter 3 points, so the VC dimension of $\mathcal{F}$ is 2.

When we use $\mathcal{F}$ to classify a 3 points dataset, it can only make 6 classification cases. So the Growth Number $S_{\mathcal{F}}(3) = 6$. (The last two cases below cannot be classified by $\mathcal{F}$.)



The VC Dimension of Linear Classifier

Theorem: Let $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\} \in \mathbb{R}^n$ be a point set in n-dimensional space. Let $\mathcal{F} = \{f(\boldsymbol{x}) = \mathbf{1}_{\boldsymbol{w}^T\boldsymbol{x}+b>0}, \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$ be the function class of all linear classifier in $\mathbb{R}^n$. Then $\mathcal{F}$ shatters S only when $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m$ are linearly independent and $m \leq n+1$.

Proof: Let $\boldsymbol{z}_i = [1\ \boldsymbol{x}_i^T]^T$, we can write $\mathcal{F}$ as $\mathcal{F} = \{f(\boldsymbol{x}) = I(\boldsymbol{\theta}^T\boldsymbol{z} > 0), \boldsymbol{\theta} \in \mathbb{R}^{n+1}\}$. Let $\boldsymbol{Z} = [\boldsymbol{z}_1\ \boldsymbol{z}_2\ ...\ \boldsymbol{z}_m] \in \mathbb{R}^{(n+1)\times m}$, we want the system of linear equations $\boldsymbol{Z}^T\boldsymbol{\theta} = \boldsymbol{y}$ solvable for arbitrary $\boldsymbol{y} \in \mathbb{R}^m$ (each element in $\boldsymbol{y}$ can either be > 0 or < 0 in order to cover all possible cases). This means rank$(\boldsymbol{Z}) =$ rank$(\boldsymbol{y}) = m$ and $\boldsymbol{\theta}$ has unique solution. Since rank$(\boldsymbol{Z}) = \min\{m, n+1\}$, we must have $m \leq n+1$. Since $\boldsymbol{Z}$ is of full rank and $m \leq n+1$, all its column vectors $\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m$ are linearly independent, which means $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m$ are linearly independent.

Sauer's Lemma

Consider such a case, when m is far greater than the VC dimension of $\mathcal{F}$, what will the growth function be like? The following theorem gives us a upper bound.

Theorem (Sauer's Lemma): Let $\mathcal{F}$ be a function class of binary output functions and its VC dimension is h. Then for all $m \in \mathbb{N}$:

$$S_{\mathcal{F}}(m) \leq \sum_{i=0}^{h} \binom{m}{i}$$

Furthermore, for all $m \geq h$, we have

$$S_{\mathcal{F}}(m) \leq (\frac{em}{h})^h$$

Proof: For the first inequality, we prove it by induction. When $m \leq h$, $S_{\mathcal{F}}(m) = 2^m = \sum_{i=0}^{h} \binom{m}{i}$, so this inequality holds.

When $m > h$, assume $S_{\mathcal{F}}(m-1) \leq \sum_{i=0}^{h} \binom{m-1}{i}$ is true. By induction, we want to prove $S_{\mathcal{F}}(m) \leq \sum_{i=0}^{h} \binom{m}{i}$ is true.

Consider a dataset $X' = \{x_1, x_2, ..., x_{m-1}\}$ having $m-1$ points, and $\mathcal{F}$ can generate $S_{\mathcal{F}}(m-1)$ classification results on X'. Let's add a new point x_m to the dataset, i.e., let the new dataset $X = X' \cup \{x_m\}$. Now consider the classification results of $\mathcal{F}$ on X.

	$\mathcal{F}$						$\mathcal{F}_1$					$\mathcal{F}_2$			
	x_1	x_2	x_3	x_4	x_5		x_1	x_2	x_3	x_4		x_1	x_2	x_3	x_4
f_1	0	1	1	0	0	$\rightarrow$	0	1	1	0					
f_2	0	1	1	0	1						$\rightarrow$	0	1	1	0
f_3	0	1	1	1	0	$\rightarrow$	0	1	1	1					
f_4	1	0	0	1	0	$\rightarrow$	1	0	0	1					
f_5	1	0	0	1	1						$\rightarrow$	1	0	0	1
f_6	1	1	0	0	1	$\rightarrow$	1	1	0	0					

We partition $\mathcal{F}$ into $\mathcal{F}_1$ and $\mathcal{F}_2$. Let f_is be the groups of $\mathcal{F}$. Each group corresponds to a classification result of X. If we ignore x_m, there can be at most two fs in $\mathcal{F}$ that generate the same label on $x_1, ..., x_{m-1}$. If there exists two, we put one into $\mathcal{F}_1$ and another into $\mathcal{F}_2$. If there exists only one, we put it into $\mathcal{F}_1$. The above table shows an example of $m = 5$. Therefore, we have the induction

$$S_{\mathcal{F}}(m) = S_{\mathcal{F}_1}(m-1) + S_{\mathcal{F}_2}(m-1)$$

It is obvious that the number of groups in $\mathcal{F}_1$ will be the same as $S_{\mathcal{F}}(m-1)$, i.e., $S_{\mathcal{F}_1}(m-1) = S_{\mathcal{F}}(m-1)$.

For $\mathcal{F}_2$, if there exists a set $T \subset X'$ such that $\mathcal{F}_2$ shatters T, then $\mathcal{F}$ shatters $T \cup \{x_m\}$. This is because for every f in $\mathcal{F}_2$, we can always find its counterpart in $\mathcal{F}_1$ such that they generate the same label on $x_1, x_2, ..., x_{m-1}$ but different on x_m. Therefore, VCDim$(\mathcal{F}_2) \leq$ VCDim$(\mathcal{F}) - 1$. Since VCDim$(\mathcal{F}) = h$, VCDim$(\mathcal{F}_2) \leq h - 1$. By assumption, we have $S_{\mathcal{F}_2}(m-1) \leq \sum_{i=0}^{h-1} \binom{m-1}{i}$.

Therefore,

$$\begin{aligned}
S_{\mathcal{F}}(m) &= S_{\mathcal{F}_1}(m-1) + S_{\mathcal{F}_2}(m-1) \\
&= S_{\mathcal{F}}(m-1) + S_{\mathcal{F}_2}(m-1) \\
&\leq \sum_{i=0}^{h} \binom{m-1}{i} + \sum_{i=0}^{h-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{h} \binom{m-1}{i} + \sum_{i=0}^{h} \binom{m-1}{i-1} \\
&= \sum_{i=0}^{h} \binom{m}{i}
\end{aligned}$$

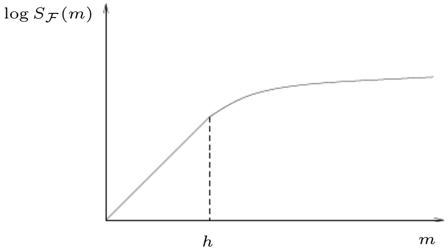
The last equality above uses the fact that $\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}$ for any $i \leq m-1$.

The second inequality $S_{\mathcal{F}}(m) \leq (\frac{em}{h})^h$ is because, when $m \geq h$,

$$S_{\mathcal{F}}(m) \leq \sum_{i=0}^{h} \binom{m}{i} \leq \left(\frac{m}{h}\right)^h \sum_{i=0}^{h} \binom{m}{i} \left(\frac{h}{m}\right)^i \leq \left(\frac{m}{h}\right)^h \left(1 + \frac{h}{m}\right)^m \leq \left(\frac{em}{h}\right)^h$$

The last inequality uses the fact that $(1 + \frac{1}{m})^m \leq e \Rightarrow (1 + \frac{h}{m})^m \leq e^h$.

The Sauer's Lemma shows the relationship between Growth Function and VC Dimension. Since $\log S_{\mathcal{F}}(m) = O(m)$ when $m \leq h$ and $\log S_{\mathcal{F}}(m) = O(\log m)$ when $m > h$, we can draw the figure as follows:



Contents

Vapnik-Chervonenkis Theorem

Theorem (Vapnik-Chervonenkis): For any $\delta > 0$, with respect to a random draw of the data,

$$\forall f \in \mathcal{F} \quad P\left[R^{\text{true}}(f) - R^{\text{emp}}(f) \leq 2\sqrt{2\frac{\log S_{\mathcal{F}}(2m) + \log \frac{4}{\delta}}{m}}\right] \geq 1 - \delta$$

Note that the above bound uses the Growth Function $S_{\mathcal{F}}$ instead of $|\mathcal{F}|$. $|\mathcal{F}|$ is infinite, but we can use $S_{\mathcal{F}}$ as a finite measure to bound $R^{\text{true}}(f) - R^{\text{emp}}(f)$.

The following pages will prove the Vapnik-Chervonenkis Theorem. I will first introduce some basics of statistics, then Symmetrization Lemma which is utilized in the proof, and finally make the proof by combining all the pieces.

Remember that $R^{\text{true}}(f) = E(R^{\text{emp}}(f))$. For simplicity, I denote $Z_i = \mathbf{1}_{f(z_i) \neq y_i} \in \{0, 1\}$, $Z = \frac{1}{m}\sum_{i=1}^{m} Z_i = R^{\text{emp}}(f)$ and $E(Z) = R^{\text{true}}(f)$.

Some Basics

- Let A and B be two events. If A happens, B will necessary happen (i.e., $A \Rightarrow B$), then $P(A) \leq P(B)$.

Proof: Let $\neg A$ be the event that A does not happen. Use the Law of Total Probability, we can write $P(B)$ as:

$$P(B) = P(B|A)P(A) + P(B|\neg A)P(\neg A)$$

Since $P(B|A) = 1$ and $P(B|\neg A)P(\neg A) \geq 0$, we have $P(B) \geq P(A)$.

- Let t_1 and t_2 be two variables, a and b be two constant, then $t_1 \geq a$ and $t_2 \geq b \Rightarrow t_1 + t_2 \geq a + b$, $t_1 + t_2 \geq a + b \Rightarrow t_1 \geq a$ or $t_2 \geq b$. That is,

$$P(t_1 \geq a \text{ and } t_2 \geq b) \leq P(t_1 + t_2 \geq a + b)$$

$$P(t_1 + t_2 \geq a + b) \leq P(t_1 \geq a \text{ or } t_2 \geq b)$$

Some Basics

- Uniform Bound: Let $S = \{s_1, s_2, ..., s_n\}$ be a set of random variables. Let t be a constant, then

$$P\left[\sup_{s\in S}\{s\} \le t\right] = P\left[s_1 \le t \text{ and } s_2 \le t \text{ and ... and } s_n \le t\right]$$

Or we can say $P\left[\sup_{s\in S}\{s\} \le t\right]$ means $\forall s \in S, P\left[s \le t\right]$.

Let $A : \{\sup_{s\in S}\{s\} \le t\}$ and $B : \{s_1 \le t \text{ and } s_2 \le t \text{ and ... and } s_n \le t\}$ be two events, we have $A \Leftrightarrow B$, thus $P(A) = P(B)$.

Similarly, $P\left[\sup_{s\in S}\{s\} \ge t\right]$ means $\exists s \in S, P\left[s \ge t\right]$. That is

$$\begin{aligned} P\left[\sup_{s\in S}\{s\} \ge t\right] &= 1 - P\left[s_1 \le t \text{ and } s_2 \le t \text{ and ... and } s_n \le t\right] \\ &= P\left[s_1 \ge t \text{ or } s_2 \ge t \text{ or ... or } s_n \ge t\right] \end{aligned}$$

Let $C : \{s_1 \ge t \text{ or } s_2 \ge t \text{ or ... or } s_n \ge t\}$ be another event, we have $C \Leftrightarrow \neg B$, thus $P(B) + P(C) = 1$.

Symmetrization Lemma

Lemma (Symmetrization): Let $Z_1, Z_2, ..., Z_m$ be m samples from D. Each $Z_i \in [0,1]$. Let $Z = \frac{1}{m}\sum_{i=1}^m Z_i$ be the sample mean of the m samples. Let $Z' = \frac{1}{m}\sum_{i=1}^m Z'_i$ where $Z'_1, Z'_2, ..., Z'_m$ are another m samples from D. Z and Z' are independent. Then for any $t > \sqrt{\frac{2}{m}}$,

$$P\left[E(Z) - Z \geq t\right] \leq 2P\left[Z' - Z \geq \frac{t}{2}\right]$$

This shows we can bound $E(Z) - Z$ using the difference of two samples $Z' - Z$. Here we call Z' a "ghost sample".

Proof: Since

$$\begin{aligned} P\left[E(Z) - Z \geq t\right] P\left[E(Z') - Z' \leq \frac{t}{2}\right] &= P\left[E(Z) - Z \geq t \text{ and } E(Z') - Z' \leq \frac{t}{2}\right] \\ &\leq P\left[(E(Z) - Z) - (E(Z') - Z') \geq t - \frac{t}{2}\right] \\ &= P\left[Z' - Z \geq \frac{t}{2}\right] \end{aligned}$$

And we can bound $P[E(Z') - Z' \leq \frac{t}{2}]$ using Chebyshev's Inequality:

$$P\left[E(Z') - Z' \geq \frac{t}{2}\right] \leq P\left[|E(Z') - Z'| \geq \frac{t}{2}\right] \leq \frac{4\mathsf{Var}(Z_i)}{mt^2} \leq \frac{1}{mt^2}$$

$$P\left[E(Z') - Z' \geq \frac{t}{2}\right] \leq P\left[|E(Z') - Z'| \geq \frac{t}{2}\right] \leq \frac{4\mathsf{Var}(Z_i)}{mt^2} \leq \frac{1}{mt^2}$$

The last inequality is because $Z_i \in [0,1] \Rightarrow Z_i^2 < Z_i$, thus

$$\mathsf{Var}(Z_i) = E(Z_i^2) - E^2(Z_i) \leq E(Z_i) - E^2(Z_i) \leq \frac{1}{4}$$

Therefore, $P\left[E(Z') - Z' \leq \frac{t}{2}\right] \geq \left(1 - \frac{1}{mt^2}\right)$, and

$$P\left[E(Z) - Z \geq t\right]\left(1 - \frac{1}{mt^2}\right) \leq P\left[Z' - Z \geq \frac{t}{2}\right]$$

$$P\left[E(Z) - Z \geq t\right] \leq \frac{mt^2}{mt^2 - 1} P\left[Z' - Z \geq \frac{t}{2}\right]$$

Since for any $t > \sqrt{\frac{2}{m}} \Rightarrow mt^2 > 2$ the above inequality holds, we have

$$P\left[E(Z) - Z \geq t\right] \leq 2P\left[Z' - Z \geq \frac{t}{2}\right]$$

Note that the Symmetrization Lemma also has a two-side form which can be proved similarly:

$$P\left[|E(Z) - Z| \geq t\right] \leq 2P\left[|Z' - Z| \geq \frac{t}{2}\right]$$

Proof of the VC Theorem

We can bound $P\left[Z' - Z \geq \frac{t}{2}\right]$ using the Hoeffding Bound:

$$
\begin{aligned}
P\left[Z' - Z \geq t\right] &= P\left[Z' - E(Z') + E(Z) - Z \geq t\right] \\
&\leq P\left[Z' - E(Z') \geq \frac{t}{2} \text{ or } E(Z) - Z \geq \frac{t}{2}\right] \\
&\leq P\left[Z' - E(Z') \geq \frac{t}{2}\right] + P\left[E(Z) - Z \geq \frac{t}{2}\right] \\
&\leq e^{-mt^2/2} + e^{-mt^2/2} \quad \text{[one-side Hoeffding Bound]} \\
&= 2e^{-mt^2/2}
\end{aligned}
$$

Therefore,

$$
P\left[Z' - Z \geq \frac{t}{2}\right] \leq 2e^{-mt^2/8}
$$

Proof of the VC Theorem

Now we put all the pieces together:

$$
\begin{aligned}
&P\left[\sup_{f\in\mathcal{F}}\left(R^{\mathsf{true}}(f)-R^{\mathsf{emp}}(f)\right)\geq t\right]\\
&\leq 2P\left[\sup_{f\in\mathcal{F}}\left(R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)\right)\geq t/2\right]\quad\text{[Symmetrization Lemma]}\\
&=2P\left[\sup_{f\in\mathcal{F}_{z_1,\ldots,z_m,z_1',\ldots,z_m'}}\left(R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)\right)\geq t/2\right]\quad\text{[restrict to data]}\\
&\leq 2\sum_{f\in\mathcal{F}_{z_1,\ldots,z_m,z_1',\ldots,z_m'}}P\left[\left(R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)\right)\geq t/2\right]\quad\text{[union bound]}\\
&\leq 2\sum_{f\in\mathcal{F}_{z_1,\ldots,z_m,z_1',\ldots,z_m'}}2e^{-mt^2/8}\quad\text{[Hoeffding bound]}\\
&=4e^{-mt^2/8}\sum_{f\in\mathcal{F}_{z_1,\ldots,z_m,z_1',\ldots,z_m'}}1\\
&=4S_{\mathcal{F}}(2m)e^{-mt^2/8}
\end{aligned}
$$

On previous page, the first inequality

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\mathsf{true}}(f)-R^{\mathsf{emp}}(f)\right)\geq t\right]\leq 2P\left[\sup_{f\in\mathcal{F}}\left(R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)\right)\geq t/2\right]$$

uses the Symmetrization Lemma for the uniform bound. To prove this, consider A, B, C, D are four independent events, $P(A)\leq 2P(C)$ and $P(B)\leq 2P(D)$, then it is easy to show $P(A\cup B)\leq 2P(C\cup D)$.

In the second equality

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)\right)\geq\frac{t}{2}\right]=P\left[\sup_{f\in\mathcal{F}_{z_1,\dots,z_m,z_1',\dots,z_m'}}\left(R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)\right)\geq\frac{t}{2}\right]$$

We project the fs in $\mathcal{F}$ on the double sample $z_1,...,z_m,z_1',...,z_m'$, thus $S_{\mathcal{F}}(2m)$ groups in total. Remember that $R^{\mathsf{emp}\prime}(f)=\frac{1}{m}\sum_{i=1}^m f(z_i')$ and $R^{\mathsf{emp}}(f)=\frac{1}{m}\sum_{i=1}^m f(z_i)$, thus $R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f)=\frac{1}{m}\sum_{i=1}^m(f(z_i')-f(z_i))$. We can consider this as a transformation $g:X\to Y$ where $X=\{(f(z_1),...,f(z_m),f(z_1'),...,f(z_m')):f\in\mathcal{F}\}$ and $Y=\{R^{\mathsf{emp}\prime}(f)-R^{\mathsf{emp}}(f):f\in\mathcal{F}\}$. The number of elements in X is known as $S_{\mathcal{F}}(2m)$.

Let $g : X \to Y$ be a transformation where X is the domain and Y is the codomain.

- The key idea is, since the number of elements in $\mathcal{F}$ is infinite, the probability $P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{true}}(f) - R^{\text{emp}}(f)\right) \geq t\right]$ cannot be calculated, however we can upperbound it by $P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{emp}\prime}(f) - R^{\text{emp}}(f)\right) \geq \frac{t}{2}\right]$. For the upperbound, we can find a domain $\mathcal{F}_{z_1,\ldots,z_m,z'_1,\ldots,z'_m}$ mapping to $\{R^{\text{emp}\prime}(f) - R^{\text{emp}}(f) : f \in \mathcal{F}\}$ such that the number of elements in the domain is finite, thus the probability of the upperbound can be calculated.

So why must we need the domain X to be finite? I will explain this by the Law of Total Probability. Generally, let $X = \{x_1, x_2, ..., x_m\}$, where m is the number of elements in X. Let A be the event that x_i happens. Suppose for each x_i, $P(A = x_i) = 1/m$. Let B be another event, we can write the probability as

$$P(B) = \sum_{i=1}^{m} P(B|A = x_i)P(A = x_i) = \frac{1}{m}\sum_{i=1}^{m} P(B|A = x_i)$$

Suppose $P(B|A = x_i)$ is identical for any x_i, we have $P(B) = P(B|A = x_i)$. This shows for a single B the probability $P(B)$ does not depend on m, even m is infinite. However, when calculating the union probability

$$P((B|A = x_1) \cup (B|A = x_2) \cup ... \cup (B|A = x_m)) \leq \sum_{i=1}^{m} P(B|A = x_i) = mP(B)$$

The union probability goes to infinity when $m \to \infty$. We do not wish this happen!

Therefore,

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{true}}(f)-R^{\text{emp}}(f)\right)\geq t\right]\leq 4S_{\mathcal{F}}(2m)e^{-mt^2/8}$$

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{true}}(f)-R^{\text{emp}}(f)\right)\leq t\right]\geq 1-4S_{\mathcal{F}}(2m)e^{-mt^2/8}$$

Let $\delta = 4S_{\mathcal{F}}(2m)e^{-mt^2/8}$, then $t=\sqrt{\frac{8}{m}\log\frac{4S_{\mathcal{F}}(2m)}{\delta}} = 2\sqrt{2\frac{\log S_{\mathcal{F}}(2m)+\log 4/\delta}{m}}$.
Then we have

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{true}}(f)-R^{\text{emp}}(f)\right)\leq 2\sqrt{2\frac{\log S_{\mathcal{F}}(2m)+\log\frac{4}{\delta}}{m}}\right]\geq 1-\delta$$

Combining Sauer's Lemma and VC Theorem

Remember in Sauer's Lemma we can bound the growth function by $S_{\mathcal{F}}(m) \leq (\frac{em}{h})^h$ when $m \geq h$. By plugging this into the bound of VC Theorem, we get

$$\forall f \in \mathcal{F} \quad P\left[R^{\mathsf{true}}(f) - R^{\mathsf{emp}}(f) \leq 2\sqrt{2\frac{h\log\frac{2em}{h} + \log\frac{4}{\delta}}{m}}\right] \geq 1 - \delta$$

This shows the difference between $R^{\mathsf{true}}(f)$ and $R^{\mathsf{emp}}(f)$ is at most of order $\sqrt{\frac{h\log m}{m}}$.

References

[1] Prediction: Machine Learning And Statistics. Lecture 14.
https://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-2012/pages/lecture-notes/

[2] Statistical Learning Theory. Lecture 5.
https://www.stat.purdue.edu/~jianzhan/STAT598Y/

[3] Introduction to Machine Learning. Lecture 16.
https://www.cs.cmu.edu/~epxing/Class/10701/lecture.html

[4] https://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0220.pdf

[5] https://courses.cs.washington.edu/courses/cse522/11wi/scribes/lecture9.pdf