Introduction to VC dimension

Matthieu R. Bloch

1 Motivation: revisiting PAC learnability

For a hypothesis set $\mathcal{H}$ with $|\mathcal{H}| < \infty$ and $h^* = \operatorname{argmax}_{h\in\mathcal{H}} \widehat{R}_N(h)$, the key result behind our previous PAC learnability analysis is the inequality

$$\forall\epsilon > 0 \quad \mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geqslant \epsilon\right) \leqslant 2\,|\mathcal{H}|\exp(-2N\epsilon^2). \tag{1}$$

In particular, the factor $|\mathcal{H}|$ is the result of the *union bound*, which is used to show that for $\epsilon > 0$

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geqslant \epsilon\right) &\leqslant \mathbb{P}\left(\exists h \in \mathcal{H} : \left|\widehat{R}_N(h) - R(h)\right| \geqslant \epsilon\right) &(2)\\ &\leqslant \sum_{j=1}^{|\mathcal{H}|}\mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geqslant \epsilon\right). &(3)\end{aligned}$$

The second inequality is tight when the events $\mathcal{E}_j \triangleq \{\left|\widehat{R}_N(h_j) - R(h_j)\right| \geqslant \epsilon\}$ are *disjoint*, but this is rarely the case in our classification setup. This is illustrated in Fig. 1 below, in which the two linear classifier in $\mathbb{R}^2$ shown are distinct but have exactly the same empirical risk on the training set.

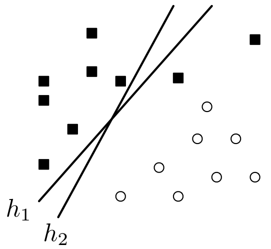


Figure 1: Two distinct classsifiers with the same empirical risk

This observations suggests that our bound might be extremely loose and that $|\mathcal{H}|$ may not necessarily be the right measure of the *richness* of the hypothesis set $\mathcal{H}$. Most of our work in the next few lectures will be devoted to finding a suitable replacement for $|\mathcal{H}|$, which will enable use to prove a generalization bound even in settings for which $|\mathcal{H}| = \infty$, as is the case for linear classifiers.

2 Dichotomy and growth function

Motivated by the situation in Fig. 1, where *many* classifier have the same empirical risk, we will attempt to assess the number of hypotheses that lead to *distinct* labelings for a given dataset. Intuitively, we are hoping that the number of distinct labelings is a quantity that better captures the richness of the hypothesis class $\mathcal{H}$. Formally, we introduce the notion of *dichotomy*. In what follows we restrict ourselves to the binary classification problem with labels $\{\pm 1\}$.

Definition 2.1 (Dichotomy). *For a dataset $\mathcal{D} \triangleq \{\mathbf{x}_i\}_{i=1}^N$ and set of hypotheses $\mathcal{H}$, the set of dichotomies generated by $\mathcal{H}$ on $\mathcal{D}$ is the set of labelings that can be generated by classfiers in $\mathcal{H}$ on the dataset, i.e.,*

$$\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N) \triangleq \{\{h(\mathbf{x}_i)\}_{i=1}^N : h \in \mathcal{H}\}. \tag{4}$$

Note that, as illustrated in Fig. 1, many sets $\{\{h(\mathbf{x}_i)\}_{i=1}^N$ for distinct h are actually identical because the labelings induced on the dataset are identical. By definition, for our binary labeling problem, $\left|\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)\right| \leqslant 2^N$ and in general $\left|\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)\right| \ll |\mathcal{H}|$. Unfortunately, $\left|\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)\right|$ is not a particularly useful quantity because it is not only potentially difficult to compute but also dependent on a specific dataset. This motivates the definition of the *growth function* as follows.

Definition 2.2 (Growth function). *For a set of hypotheses $\mathcal{H}$, the* growth function *of $\mathcal{H}$ is*

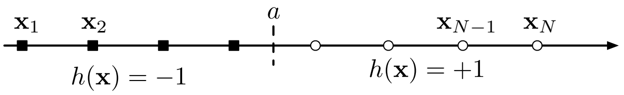
$$m_{\mathcal{H}}(N) \triangleq \max_{\{\mathbf{x}_i\}_{i=1}^N} \left|\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)\right|. \tag{5}$$

Note that the growth function depends on the number of datapoints N but not on the exact datapoints $\{\mathbf{x}_i\}_{i=1}^N$. The growth function measures the maximum number of dichotomies that $\mathcal{H}$ can generate over *all* possible datasets, and by definition, it still holds that $m_{\mathcal{H}}(N) \leqslant 2^N$.

Example 2.3 (Positive rays). *Consider a binary classification problem in $\mathbb{R}$ with the set of positive rays*

$$\mathcal{H} \triangleq \{h_a : \mathbb{R} \to \{\pm 1\} : x \mapsto \text{sign}\,(x-a)\,|a \in \mathbb{R}\}. \tag{6}$$

As illustrated below, the threshold a defines a classifier such that all points to the left are assigned label -1 while all points to the right are assigned label $+1$.



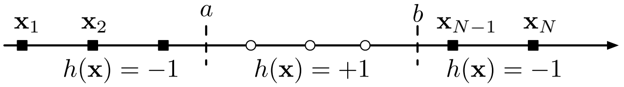
Although $\mathcal{H} = \infty$, the number of dichotomies is still finite, and one can actually compute the growth function exactly. In general, this is challenging because we need to identify the worst case *dataset that generates the highest number of dichotomies; here, this is only tractable because the situation is simple.*

Without losing generality, we can assume that all N points $\{x_i\}_{i=1}^N$ are distinct. Let us introduce $x_0 \triangleq -\infty$ and $x_{N+1} \triangleq \infty$. For any $i \geqslant 0$, all classifiers h_a with $x_i \leqslant a < x_{i+1}$ induce the same labeling. Consequently, the number of distinct labelings is at most $N+1$ and $m_{\mathcal{H}}(N) = N+1$. Interestingly, the growth function is growing polynomially in N, which is much slower than the exponential growth 2^N allowed by the upper bound.

Example 2.4 (Positive intervals). *Consider a binary classification in $\mathbb{R}$ with the set of positive intervals*

$$\mathcal{H} \triangleq \{h_{a,b} : \mathbb{R} \to \{\pm 1\} : x \mapsto \mathbb{1}\{x \in [a;b]\} - \mathbb{1}\{x \notin [a;b]\}\,|a < b \in \mathbb{R}\}. \tag{7}$$

As illustrated below, the thresholds $a < b$ define a classifier such that all points with $[a;b]$ are assigned label $+1$ while all points outside are assigned label -1.



Again, this is a situation for which we can compute the growth function exactly. Without loss of generality, we assume that all N datapoints are distinct and we introduce $x_0 \triangleq -\infty$ and $x_{N+1} \triangleq \infty$. We need to be a bit more careful when counting dichotomies:

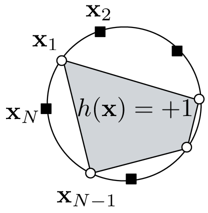
- *If $x_0 < a < b \leqslant x_1$, all classifiers h_{ab} induce an all--1 labeling;*
- *for any $0 \leqslant i < j \leqslant N$, all classifiers h_{ab} such that $x_i \leqslant a \leqslant x_{i+1} < x_j \leqslant b \leqslant x_{j+1}$ induce the same labelings;*
- *for any $0 \leqslant i \leqslant N$, all classifiers h_{ab} such that $x_i \leqslant a < b < x_{i+1}$ induce again an all--1 labeling.*

Consequently, the number of classifiers is $1 + \binom{N+1}{2}$ and $m_{\mathcal{H}}(N) = \frac{N^2}{2} + \frac{N}{2} + 1$, which grows again polynomially in N.

Example 2.5 (Convex sets). *Consider a binary classification in $\mathbb{R}^2$ with the set*

$$\mathcal{H} \triangleq \{h : \mathbb{R}^2 \to \{\pm 1\} | \{\mathbf{x} \in \mathbb{R}^2 : h(\mathbf{x}) = +1\} \text{ is convex}\}. \tag{8}$$

Consider a set of N distinct points distributed on the unit circle, as illustrated below.



Notice that irrespective of the labeling of the datapoints, the datapoints for which $h(\mathbf{x}_i) = +1$ define the vertices of a polytope, which is convex. Said differently, irrespective of the labeling there exists $h \in \mathcal{H}$ that generates the labeling. Therefore, by definition, $m_{\mathcal{H}} \geqslant 2^N$; since we also know that $m_{\mathcal{H}} \leqslant 2^N$, it must hold that $m_{\mathcal{H}} = 2^N$.

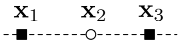
The three previous examples are not at all representative of a general situation because it is nearly impossible to compute the growth function exactly in most practical cases. As shown next, even for linear classifiers this can become a formidable task.

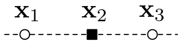
Example 2.6 (Linear classifiers). *Consider a binary classification in $\mathbb{R}^2$ with the set of linear classifiers*

$$\mathcal{H} \triangleq \{h : \mathbb{R}^2 \to \{\pm 1\} : \mathbf{x} \mapsto \text{sign}\left(\mathbf{w}^\intercal \mathbf{x} + b\right) | \mathbf{w} \in \mathbb{R}^2, b \in \mathbb{R}\} \tag{9}$$

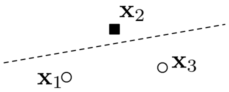
The challenge again is to identify the worst case dataset that generates the most dichotomies. We first note that $\{\mathbf{x} : \mathbf{w}^\intercal \mathbf{x} + b = 0\} = \{\mathbf{x} : -\mathbf{w}^\intercal \mathbf{x} + b = 0\}$, so that a single line actually defines two classifiers.

For $N = 3$, we need to distinguish two cases. If all three points are aligned, all dichotomies except those illustrated below are possible, we therefore obtain six dichotomies.



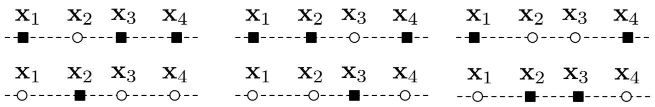


If the three points are not aligned, they form the vertices of a polytope and any hyperplane cutting the polytope will isolate one point. In addition, any hyperplane no cutting the polytope will assign the same label to all three points. Consequently, the number of dichotomies generated is $8 = 2^3$.

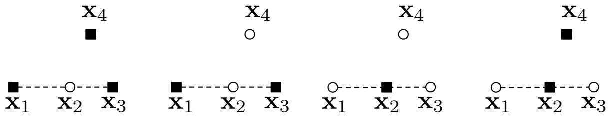


Consequently, $m_{\mathcal{H}}(3) = 8$.

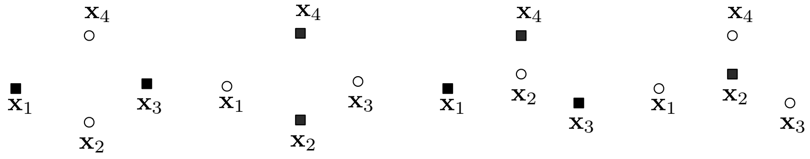
For $N = 4$*, we need to distinguish even more cases. If all four points are aligned, all dichotomies except those illustrated below are possible we therefore obtain 10 dichotomies.*


$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$

If three out of four points are aligned, the four points form a 3-vertex polytope, and one point, say $\mathbf{x}_2$*, is on the edge, say defined by* $\mathbf{x}_1$ *and* $\mathbf{x}_3$*. Any hyperplane cutting through the polytope cannot assign a label to* $\mathbf{x}_2$ *that is distinct of both* $\mathbf{x}_1$ *and* $\mathbf{x}_3$*. Consequently, the dichotomies illustrated below cannot be generated and we obtain 12 dichotomies.*


$\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

If no three out of four points are aligned, the four points could form a 4-vertex polytope, in which case a hyperplane cutting through the polytope cannot assign distinct labels to a vertex and all its neighbors. The four points could also form a 3-vertex polytope with a point in the interior, in which case a hyperplane cutting through the polytope cannot assign a label to the interior point distinct from all the vertices. Consequently, the dichotomies illustrated below cannot be generated and we obtain 14 dichotomies.



Consequently, $m_{\mathcal{H}}(4) = 14$.

This last example illustrates the essentially combinatorial nature of the calculation of the growth function. As we will soon seen, we will conveniently only care about the *scaling* of the growth function with N in particular whether it is polynomial or exponential.

3 Shattering and break point

We introduced in the previous section the notion of growth function, $m_{\mathcal{H}}(\mathrm{N})$, which characterizes the maximum number of labellings that can be obtained with a given hypothesis set $\mathcal{H}$ over all datasets $\{\mathbf{x}_i\}_{i=1}^{N}$ of size N. The behavior of the growth function as a function of N can be different depending on the structure of the hypotheses in $\mathcal{H}$, and we saw examples in which $m_{\mathcal{H}}(N)$ grows polynomially or exponentially in N.

The problem of computing $m_{\mathcal{H}}(N)$ is intractable, because it quickly becomes an intricate computational problem that depends not only on all possible configurations of points in the dataset but also on the constraints induced by the structure of hypotheses in $\mathcal{H}$. We will focus instead on determining the behavior of $m_{\mathcal{H}}(N)$ as a function of N, which will conveniently tell us a lot about generalization in a next lecture.

We start by introducing the notion of shattering and break points.

Definition 3.1 (Shattering). *If a hypothesis set $\mathcal{H}$ can generate all dichotomies on $\{\mathbf{x}_i\}_{i=1}^N$, we say that $\mathcal{H}$ shatters $\{\mathbf{x}_i\}_{i=1}^N$*

Definition 3.2 (Break point). *If not data set of size k can be shattered by $\mathcal{H}$, then k is a break point for $\mathcal{H}$*

Note that if k is a break point, any $\ell > k$ is also a break point.

Example 3.3. *For a binary linear classifier in $\mathbb{R}^2$, we saw that $m_{\mathcal{H}}(4) = 14 < 16$. In other words, no dataset of size 4 can be shattered by linear classifiers and $k = 4$ is a break point.*

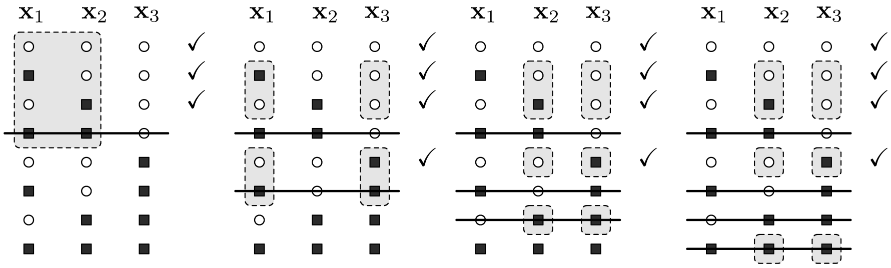
Although we gave up computing $m_{\mathcal{H}}(N)$ for linear classifiers in $\mathbb{R}^2$ for $N > 4$, it turns out that the existence of break point k is already enough for us to bound $m_{\mathcal{H}}(N)$ for every N. We will formalize this shortly in the next section, but we first illustrate this point with an example.

Example 3.4. *Consider a binary classification problem and assume that $k = 2$ is a break points for $\mathcal{H}$. How many dichotomies can we generate of set of size $N = 3$? Our assumption says that $\mathcal{H}$ cannot shatter a set of size 2, so that no $h \in \mathcal{H}$ can assign all four possible distinct labelings to any set of two points.*

Consider the table below, which illustrates all possible binary (○, ■) labelings on a set size 3.

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
■	○	○
○	■	○
■	■	○
○	○	■
■	○	■
○	■	■
■	■	■

As illustrated below, we proceed to eliminate labelings forbidden by our assumption that $k = 2$ is breakpoint starting from the top. You can check for yourself that any other order of labeling would result in us eliminating the same number of dichotomies.


$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

The first three rows correspond to labelings that do not violate our assumption. The fourth row has to be excluded because it would otherwise allow us to shatter a set of size 2, as illustrated by the gray region. The procedure continues and one can see that only 4 labelings are allowed out of the 8 possible.

The previous example shows that knowing a break point allows us to reason about the growth function without really knowing much about $\mathcal{H}$.

4 Bounding the growth function and Sauer's lemma

The crux of the approach to bound the growth function is to consider the following combinatorial quantity.

Definition 4.1. *Assume $\mathcal{H}$ has break point k. We define $B(N,k)$ as the maximum number of dichotomies of N points such that* no subset *of size k can be shattered by the dichotomies.*

Note that $B(N,k)$ is a purely combinatorial quantity, which depends on the fact that k is a break point for $\mathcal{H}$ but otherwise not on the specific nature of $\mathcal{H}$. By definition, if k is a break point for $\mathcal{H}$, then $m_{\mathcal{H}}(N) \leqslant B(N,k)$. What makes the definition of $B(N,k)$ useful is that we can bound it much more easily than $m_{\mathcal{H}}(N)$.

Lemma 4.2. *For $N>1$ $B(N,1)=1$, for $k>1$ $B(1,k)=2$, and*

$$\forall k>1 \quad B(N,k) \leqslant B(N-1,k)+B(N-1,k-1).$$

Lemma 4.3 (Sauer's lemma)**.**

$$B(N,k) \leqslant \sum_{i=0}^{k-1} \binom{N}{i} \tag{10}$$

Proof. See notes and [1, Section 2.1.2]. ■

Lemma 4.3 therefore implies that $B(N,k)$ is a polynomial in N of degree at most $k-1$. In addition, we can conclude that if $\mathcal{H}$ has a break point (no matter its value as long as it is finite) then $m_{\mathcal{H}}(N)$ is polynomial in N.

5 VC bound

We are finally ready to establish the result promised in the introduction, known in leaning theory as the VC bound. The goal of this section is to establish the following theorem.

Theorem 5.1. *Consider a potentially infinite hypothesis set $\mathcal{H}$. Then, for a dataset $\{(\mathbf{x}_i,y_i)\}_{i=1}^N$, we have*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right) \leqslant 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

This theorem should be compared with our previous generalization bound developed for $|\mathcal{H}|<\infty$ and for which we proved

$$\mathbb{P}\left(\max_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right) \leqslant 2\left|\mathcal{H}\right|e^{-2\epsilon^2 N}.$$

The major changes in Theorem 5.1 are 1. the max is replaced by sup; and 2. $|\mathcal{H}|$ is replaced by $m_{\mathcal{H}}(2N)$. In particular, note that Theorem 5.1 can handle *infinite* hypothesis classes. To obtain a PAC style bound, note that Theorem 5.1 implies that with probability at least $1-\delta$

$$R(h^*) \leqslant \widehat{R}_N(h^*)+\sqrt{\frac{8}{N}\left(\log m_{\mathcal{H}}(2N)+\log\frac{4}{\delta}\right)}. \tag{11}$$

The key difficulty behind proof is to somehow relate $\sup_{h\in\mathcal{H}}$ to $\max_{h\in\mathcal{H}'}$ with $\mathcal{H}'\subset\mathcal{H}$ and $|\mathcal{H}'|<\infty$. The first proof to achieve this was developed by Vapnik and Chervonenkis in 1971, hence the "VC" name attached to Theorem 5.1.

We start by developing some rough intuition for the approach. Note that growth function $m_{\mathcal{H}}$ replaces $|\mathcal{H}|$ in Theorem 5.1, capturing the fact that although there may be infinitely many hypotheses in $\mathcal{H}$, the hypotheses generate a finite number of unique dichotomies. Consequently, the set $\{\widehat{R}_N(h) : h\in\mathcal{H}\}$ has *finite* cardinality. Unfortunately, note that the term $R(h)$ that appears in Theorem 5.1 still potentially takes *infinitely* many different values.

The key insight behind the VC bound is that one can bound $\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)$ using events that only depend on empirical risks using a second *ghost* dataset of size N with empirical risk $\widehat{R}'_N(h)$, with the hope that we can squeeze $R(h)$ between $\widehat{R}'_N(h)$ and $\widehat{R}_N(h)$. Specifically, we will relate $\mathbb{P}\left(\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)$ to $\mathbb{P}\left(\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\epsilon'\right)$ with $\epsilon'=f(\epsilon)$. Since $\mathbb{P}\left(\left|\widehat{R}_N(h)-\widehat{R}'_N(h)\right|>\epsilon\right)$ only depends on the finite number of unique dichotomies, bounding it with a union bound is likely to be a more fruitful endeavor.

The idea behind the ghost dataset can be captured using the following lemma.

Lemma 5.2. *Assume that X, X' be i.i.d. random variables with* symmetric *distribution around their mean μ. Let $\mathcal{A}\triangleq\{|X-\mu|>\epsilon\}$ and let $\mathcal{B}\triangleq\{|X-X'|>\epsilon\}$. Then,*

$$\mathbb{P}(\mathcal{A})\leqslant 2\mathbb{P}(\mathcal{B}).$$

If $X\triangleq\widehat{R}_N(h)$ and $X'\triangleq\widehat{R}'_N(h)$ had symmetric distributions, we would obtain

$$\mathbb{P}\left(\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)\leqslant 2\mathbb{P}\left(\left|\widehat{R}_N(h)-\widehat{R}'_N(h)\right|>\epsilon\right).$$

This is not quite true, but we will prove the very similar result given next.

Lemma 5.3. *If $N\geqslant 4\epsilon^{-2}\ln 2$,*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)\leqslant 2\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\frac{\epsilon}{2}\right)$$

Lemma 5.4. *Let $\mathcal{S}\triangleq\{(\mathbf{x}_i,y_i)\}_{i=1}^{2N}$ be a dataset partitioned into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ each containing N points uniformly at random. Assume that $\widehat{R}_N(h)$ is computed on $\mathcal{S}_1$ while $\widehat{R}'_N(h)$ is computed on $\mathcal{S}_2$. Then*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\frac{\epsilon}{2}\right)\leqslant m_{\mathcal{H}}(2N)\sup_{\mathcal{S}_1,\mathcal{S}_2}\sup_{h\in\mathcal{H}}\mathbb{P}\left(\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\frac{\epsilon}{2}|\mathcal{S}\right).$$

Lemma 5.5. *For any $h\in\mathcal{H}$ and any set $\mathcal{S}$, we have*

$$\mathbb{P}\left(\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\frac{\epsilon}{2}|\mathcal{S}\right)\leqslant 2e^{-\frac{1}{8}\epsilon^2N}.$$

Combining the results of Lemmas 5.3- 5.5 establishes Theorem 5.1. Equation (11), which follows directly from Theorem 5.1, is usually expressed slighlty differently using the *VC dimension*.

Definition 5.6. *The VC dimension of a class $\mathcal{H}$, denoted $\mathrm{d}_{\mathrm{VC}}(\mathcal{H})$, is the *largest* n such that $m_{\mathcal{H}}(n) = 2^n$.*

By definition d_{VC} is one less that the *smallest* break point.

Example 5.7. *For linear classifiers in $\mathbb{R}^2$, we have already shown that $\mathrm{d}_{\mathrm{VC}} = 3$*

Lemma 4.3 can be refined to show the following.

Lemma 5.8. *If k is a break point for $\mathcal{H}$,*

$$m_{\mathcal{H}}(N) \leqslant \sum_{i=0}^{k-1} \binom{N}{i} \leqslant N^{k-1} + 1.$$

Consequently, one can show

$$R(h^*) \leqslant \widehat{R}_N(h^*) + \sqrt{\frac{8}{N} \log \frac{4((2N)^{d_{\mathrm{VC}}} + 1)}{\delta}} \overset{(d_{\mathrm{VC}} \geqslant 2)}{\leqslant} \widehat{R}_N(h^*) + \sqrt{\frac{8 d_{\mathrm{VC}}}{N} \log \frac{8N}{\delta}},$$

which clearly illustrates how the VC bound controls the ability to generalize.

In general, computing the VC dimension of a set of hypotheses $\mathcal{H}$ is difficult, but it is sometimes possible.

Proposition 5.9. *The VC dimension for linear classifiers in $\mathbb{R}^d$ is $d + 1$.*

The number of parameters of a linear classfier in $\mathbb{R}^d$ is $d + 1$, which might suggest that the number of parameters is what influences the VC dimension. This is actually misleading because more parameters does not necessarily means higher VC dimension.

Proposition 5.10. *The VC dimension for 1-NN classifiers is$\mathrm{d}_{\mathrm{VC}} = \infty$.*

Through fairly involved arguments, one can show that SVMs with large margin have a small VC dimension and that a small generalization error. This still is sort of true for soft margin, but the math is much more involved.

To conclude, the VC dimension is a measure of complexity of infinite sized hypothesis classes which allows us to study PAC learnability in a much broader setting. However, this is does not fully solve the learning problem because the VC dimension is hard to work with in practice and fails to explain why Deep Neural Networks perform so well. The VC dimension is therefore a useful theoretical tool to develop insight regarding generalization, but will see others that are more useful in practice.

References

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.