6 The VC-Dimension

In the previous chapter, we decomposed the error of the $\text{ERM}_{\mathcal{H}}$ rule into approximation error and estimation error. The approximation error depends on the fit of our prior knowledge (as reflected by the choice of the hypothesis class $\mathcal{H}$) to the underlying unknown distribution. In contrast, the definition of PAC learnability requires that the estimation error would be bounded uniformly over all distributions.

Our current goal is to figure out which classes $\mathcal{H}$ are PAC learnable, and to characterize exactly the sample complexity of learning a given hypothesis class. So far we have seen that finite classes are learnable, but that the class of all functions (over an infinite size domain) is not. What makes one class learnable and the other unlearnable? Can infinite-size classes be learnable, and, if so, what determines their sample complexity?

We begin the chapter by showing that infinite classes can indeed be learnable, and thus, finiteness of the hypothesis class is not a necessary condition for learnability. We then present a remarkably crisp characterization of the family of learnable classes in the setup of binary valued classification with the zero-one loss. This characterization was first discovered by Vladimir Vapnik and Alexey Chervonenkis in 1970 and relies on a combinatorial notion called the Vapnik-Chervonenkis dimension (VC-dimension). We formally define the VC-dimension, provide several examples, and then state the fundamental theorem of statistical learning theory, which integrates the concepts of learnability, VC-dimension, the ERM rule, and uniform convergence.

6.1 Infinite-Size Classes Can Be Learnable

In Chapter 4 we saw that finite classes are learnable, and in fact the sample complexity of a hypothesis class is upper bounded by the log of its size. To show that the size of the hypothesis class is not the right characterization of its sample complexity, we first present a simple example of an infinite-size hypothesis class that is learnable.

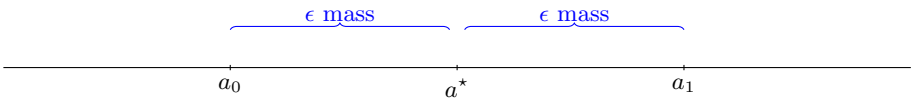
Example 6.1 Let $\mathcal{H}$ be the set of threshold functions over the real line, namely, $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0, 1\}$ is a function such that $h_a(x) = \mathbb{1}_{[x<a]}$. To remind the reader, $\mathbb{1}_{[x<a]}$ is 1 if $x < a$ and 0 otherwise. Clearly, $\mathcal{H}$ is of infinite

size. Nevertheless, the following lemma shows that $\mathcal{H}$ is learnable in the PAC model using the ERM algorithm.

Lemma 6.1 *Let $\mathcal{H}$ be the class of thresholds as defined earlier. Then, $\mathcal{H}$ is PAC learnable, using the ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$.*

Proof Let $a^\star$ be a threshold such that the hypothesis $h^\star(x) = \mathbb{1}_{[x<a^\star]}$ achieves $L_{\mathcal{D}}(h^\star) = 0$. Let $\mathcal{D}_x$ be the marginal distribution over the domain $\mathcal{X}$ and let $a_0 < a^\star < a_1$ be such that

$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^\star)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^\star, a_1)] = \epsilon.$$



(If $\mathcal{D}_x(-\infty, a^\star) \leq \epsilon$ we set $a_0 = -\infty$ and similarly for a_1). Given a training set S, let $b_0 = \max\{x : (x, 1) \in S\}$ and $b_1 = \min\{x : (x, 0) \in S\}$ (if no example in S is positive we set $b_0 = -\infty$ and if no example in S is negative we set $b_1 = \infty$). Let b_S be a threshold corresponding to an ERM hypothesis, h_S, which implies that $b_S \in (b_0, b_1)$. Therefore, a sufficient condition for $L_{\mathcal{D}}(h_S) \leq \epsilon$ is that both $b_0 \geq a_0$ and $b_1 \leq a_1$. In other words,

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0 \vee b_1 > a_1],$$

and using the union bound we can bound the preceding by

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1]. \tag{6.1}$$

The event $b_0 < a_0$ happens if and only if all examples in S are not in the interval (a_0, a^*), whose probability mass is defined to be ϵ, namely,

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x, y) \in S,\ x \notin (a_0, a^\star)] = (1-\epsilon)^m \leq e^{-\epsilon m}.$$

Since we assume $m > \log(2/\delta)/\epsilon$ it follows that the equation is at most $\delta/2$. In the same way it is easy to see that $\mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2$. Combining with Equation (6.1) we conclude our proof. □

6.2 The VC-Dimension

We see, therefore, that while finiteness of $\mathcal{H}$ is a sufficient condition for learnability, it is not a necessary condition. As we will show, a property called the VC-dimension of a hypothesis class gives the correct characterization of its learnability. To motivate the definition of the VC-dimension, let us recall the No-Free-Lunch theorem (Theorem 5.1) and its proof. There, we have shown that without

restricting the hypothesis class, for any learning algorithm, an adversary can construct a distribution for which the learning algorithm will perform poorly, while there is another learning algorithm that will succeed on the same distribution. To do so, the adversary used a finite set $C \subset \mathcal{X}$ and considered a family of distributions that are concentrated on elements of C. Each distribution was derived from a "true" target function from C to $\{0,1\}$. To make any algorithm fail, the adversary used the power of choosing a target function from the set of *all* possible functions from C to $\{0,1\}$.

When considering PAC learnability of a hypothesis class $\mathcal{H}$, the adversary is restricted to constructing distributions for which some hypothesis $h \in \mathcal{H}$ achieves a zero risk. Since we are considering distributions that are concentrated on elements of C, we should study how $\mathcal{H}$ behaves on C, which leads to the following definition.

DEFINITION 6.2 (Restriction of $\mathcal{H}$ to C) Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to C is the set of functions from C to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) : h \in \mathcal{H}\},$$

where we represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

If the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$, then we say that $\mathcal{H}$ *shatters* the set C. Formally:

DEFINITION 6.3 (Shattering) A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.

Example 6.2 Let $\mathcal{H}$ be the class of threshold functions over $\mathbb{R}$. Take a set $C = \{c_1\}$. Now, if we take $a = c_1 + 1$, then we have $h_a(c_1) = 1$, and if we take $a = c_1 - 1$, then we have $h_a(c_1) = 0$. Therefore, $\mathcal{H}_C$ is the set of all functions from C to $\{0,1\}$, and $\mathcal{H}$ shatters C. Now take a set $C = \{c_1, c_2\}$, where $c_1 \leq c_2$. No $h \in \mathcal{H}$ can account for the labeling $(0,1)$, because any threshold that assigns the label 0 to c_1 must assign the label 0 to c_2 as well. Therefore not all functions from C to $\{0,1\}$ are included in $\mathcal{H}_C$; hence C is not shattered by $\mathcal{H}$.

Getting back to the construction of an adversarial distribution as in the proof of the No-Free-Lunch theorem (Theorem 5.1), we see that whenever some set C is shattered by $\mathcal{H}$, the adversary is not restricted by $\mathcal{H}$, as they can construct a distribution over C based on *any* target function from C to $\{0,1\}$, while still maintaining the realizability assumption. This immediately yields:

COROLLARY 6.4 *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$. Let m be a training set size. Assume that there exists a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$. Then, for any learning algorithm, A, there exist a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $L_{\mathcal{D}}(h) = 0$ but with probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Corollary 6.4 tells us that if $\mathcal{H}$ shatters some set C of size $2m$ then we cannot learn $\mathcal{H}$ using m examples. Intuitively, if a set C is shattered by $\mathcal{H}$, and we receive a sample containing half the instances of C, the labels of these instances give us no information about the labels of the rest of the instances in C – every possible labeling of the rest of the instances can be explained by some hypothesis in $\mathcal{H}$. Philosophically,

If someone can explain every phenomenon, his explanations are worthless.

This leads us directly to the definition of the VC dimension.

DEFINITION 6.5 (VC-dimension) The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $\text{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.

A direct consequence of Corollary 6.4 is therefore:

THEOREM 6.6 *Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable.*

Proof Since $\mathcal{H}$ has an infinite VC-dimension, for any training set size m, there exists a shattered set of size $2m$, and the claim follows by Corollary 6.4. □

We shall see later in this chapter that the converse is also true: A finite VC-dimension guarantees learnability. Hence, the VC-dimension characterizes PAC learnability. But before delving into more theory, we first show several examples.

6.3 Examples

In this section we calculate the VC-dimension of several hypothesis classes. To show that $\text{VCdim}(\mathcal{H}) = d$ we need to show that

1. There exists a set C of size d that is shattered by $\mathcal{H}$.
2. Every set C of size $d + 1$ is not shattered by $\mathcal{H}$.

6.3.1 Threshold Functions

Let $\mathcal{H}$ be the class of threshold functions over $\mathbb{R}$. Recall Example 6.2, where we have shown that for an arbitrary set $C = \{c_1\}$, $\mathcal{H}$ shatters C; therefore $\text{VCdim}(\mathcal{H}) \geq 1$. We have also shown that for an arbitrary set $C = \{c_1, c_2\}$ where $c_1 \leq c_2$, $\mathcal{H}$ does not shatter C. We therefore conclude that $\text{VCdim}(\mathcal{H}) = 1$.

6.3.2 Intervals

Let $\mathcal{H}$ be the class of intervals over $\mathbb{R}$, namely, $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$, where $h_{a,b} : \mathbb{R} \to \{0, 1\}$ is a function such that $h_{a,b}(x) = \mathbb{1}_{[x \in (a,b)]}$. Take the set $C = \{1, 2\}$. Then, $\mathcal{H}$ shatters C (make sure you understand why) and therefore $\text{VCdim}(\mathcal{H}) \geq 2$. Now take an arbitrary set $C = \{c_1, c_2, c_3\}$ and assume without loss of generality that $c_1 \leq c_2 \leq c_3$. Then, the labeling $(1, 0, 1)$ cannot be obtained by an interval and therefore $\mathcal{H}$ does not shatter C. We therefore conclude that $\text{VCdim}(\mathcal{H}) = 2$.

6.3.3 Axis Aligned Rectangles

Let $\mathcal{H}$ be the class of axis aligned rectangles, formally:

$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

We shall show in the following that $\text{VCdim}(\mathcal{H}) = 4$. To prove this we need to find a set of 4 points that are shattered by $\mathcal{H}$, and show that no set of 5 points can be shattered by $\mathcal{H}$. Finding a set of 4 points that are shattered is easy (see Figure 6.1). Now, consider any set $C \subset \mathbb{R}^2$ of 5 points. In C, take a leftmost point (whose first coordinate is the smallest in C), a rightmost point (first coordinate is the largest), a lowest point (second coordinate is the smallest), and a highest point (second coordinate is the largest). Without loss of generality, denote $C = \{c_1, \ldots, c_5\}$ and let c_5 be the point that was not selected. Now, define the labeling $(1, 1, 1, 1, 0)$. It is impossible to obtain this labeling by an axis aligned rectangle. Indeed, such a rectangle must contain $c_1, \ldots, c_4$; but in this case the rectangle contains c_5 as well, because its coordinates are within the intervals defined by the selected points. So, C is not shattered by H, and therefore $\text{VCdim}(H) = 4$.

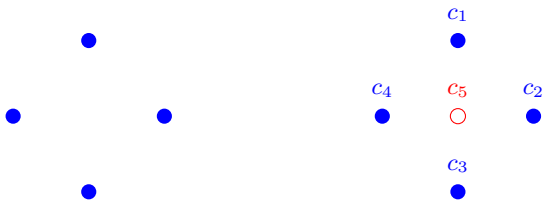


Figure 6.1 Left: 4 points that are shattered by axis aligned rectangles. Right: Any axis aligned rectangle cannot label c_5 by 0 and the rest of the points by 1.

6.3.4 Finite Classes

Let $\mathcal{H}$ be a finite class. Then, clearly, for any set C we have $|\mathcal{H}_C| \leq |\mathcal{H}|$ and thus C cannot be shattered if $|\mathcal{H}| < 2^{|C|}$. This implies that $\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$. This shows that the PAC learnability of finite classes follows from the more general statement of PAC learnability of classes with finite VC-dimension, which we shall see in the next section. Note, however, that the VC-dimension of a finite class $\mathcal{H}$ can be significantly smaller than $\log_2(|\mathcal{H}|)$. For example, let $\mathcal{X} = \{1, \ldots, k\}$, for some integer k, and consider the class of threshold functions (as defined in Example 6.2). Then, $|\mathcal{H}| = k$ but $\text{VCdim}(\mathcal{H}) = 1$. Since k can be arbitrarily large, the gap between $\log_2(|\mathcal{H}|)$ and $\text{VCdim}(\mathcal{H})$ can be arbitrarily large.

6.3.5 VC-Dimension and the Number of Parameters

In the previous examples, the VC-dimension happened to equal the number of parameters defining the hypothesis class. While this is often the case, it is not always true. Consider, for example, the domain $\mathcal{X} = \mathbb{R}$, and the hypothesis class $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$ where $h_\theta : \mathcal{X} \to \{0, 1\}$ is defined by $h_\theta(x) = \lceil 0.5 \sin(\theta x) \rceil$. It is possible to prove that $\text{VCdim}(\mathcal{H}) = \infty$, namely, for *every* d, one can find d points that are shattered by $\mathcal{H}$ (see Exercise 8).

6.4 The Fundamental Theorem of PAC learning

We have already shown that a class of infinite VC-dimension is not learnable. The converse statement is also true, leading to the fundamental theorem of statistical learning theory:

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the $0-1$ loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

The proof of the theorem is given in the next section.

Not only does the VC-dimension characterize PAC learnability; it even determines the sample complexity.

THEOREM 6.8 (The Fundamental Theorem of Statistical Learning – Quantitative Version) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the $0-1$ loss. Assume that $\text{VCdim}(\mathcal{H}) = d < \infty$. Then, there are absolute constants C_1, C_2 such that:*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m^{UC}_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq \ m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

3. *$\mathcal{H}$ is PAC learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

The proof of this theorem is given in Chapter 28.

Remark 6.3 We stated the fundamental theorem for binary classification tasks. A similar result holds for some other learning problems such as regression with the absolute loss or the squared loss. However, the theorem does not hold for all learning tasks. In particular, learnability is sometimes possible even though the uniform convergence property does not hold (we will see an example in Chapter 13, Exercise 2). Furthermore, in some situations, the ERM rule fails but learnability is possible with other learning rules.

6.5 Proof of Theorem 6.7

We have already seen that $1 \to 2$ in Chapter 4. The implications $2 \to 3$ and $3 \to 4$ are trivial and so is $2 \to 5$. The implications $4 \to 6$ and $5 \to 6$ follow from the No-Free-Lunch theorem. The difficult part is to show that $6 \to 1$. The proof is based on two main claims:

- If $\text{VCdim}(\mathcal{H}) = d$, then even though $\mathcal{H}$ might be infinite, when restricting it to a finite set $C \subset \mathcal{X}$, its "effective" size, $|\mathcal{H}_C|$, is only $O(|C|^d)$. That is, the size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$. This claim is often referred to as *Sauer's lemma*, but it has also been stated and proved independently by Shelah and by Perles. The formal statement is given in Section 6.5.1 later.
- In Section 4 we have shown that finite hypothesis classes enjoy the uniform convergence property. In Section 6.5.2 later we generalize this result and show that uniform convergence holds whenever the hypothesis class has a "small effective size." By "small effective size" we mean classes for which $|\mathcal{H}_C|$ grows polynomially with $|C|$.

6.5.1 Sauer's Lemma and the Growth Function

We defined the notion of *shattering*, by considering the restriction of $\mathcal{H}$ to a finite set of instances. The growth function measures the maximal "effective" size of $\mathcal{H}$ on a set of m examples. Formally:

DEFINITION 6.9 (Growth Function) Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_{\mathcal{H}}(m) \;=\; \max_{C \subset \mathcal{X} : |C| = m} \left| \mathcal{H}_C \right| .$$

In words, $\tau_H(m)$ is the number of different functions from a set C of size m to $\{0,1\}$ that can be obtained by restricting $\mathcal{H}$ to C.

Obviously, if $\text{VCdim}(\mathcal{H}) = d$ then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$. In such cases, $\mathcal{H}$ induces all possible functions from C to $\{0,1\}$. The following beautiful lemma, proposed independently by Sauer, Shelah, and Perles, shows that when m becomes larger than the VC-dimension, the growth function increases polynomially rather than exponentially with m.

LEMMA 6.10 (Sauer-Shelah-Perles) *Let $\mathcal{H}$ be a hypothesis class with $\text{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.*

Proof of Sauer's Lemma *

To prove the lemma it suffices to prove the following stronger claim: For any $C = \{c_1, \ldots, c_m\}$ we have

$$\forall \mathcal{H}, \quad |\mathcal{H}_C| \;\leq\; |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|. \tag{6.3}$$

The reason why Equation (6.3) is sufficient to prove the lemma is that if $\text{VCdim}(\mathcal{H}) \leq d$ then no set whose size is larger than d is shattered by $\mathcal{H}$ and therefore

$$|\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| \;\leq\; \sum_{i=0}^{d} \binom{m}{i}.$$

When $m > d+1$ the right-hand side of the preceding is at most $(em/d)^d$ (see Lemma A.5 in Appendix A).

We are left with proving Equation (6.3) and we do it using an inductive argument. For $m = 1$, no matter what $\mathcal{H}$ is, either both sides of Equation (6.3) equal 1 or both sides equal 2 (the empty set is always considered to be shattered by $\mathcal{H}$). Assume Equation (6.3) holds for sets of size $k < m$ and let us prove it for sets of size m. Fix $\mathcal{H}$ and $C = \{c_1, \ldots, c_m\}$. Denote $C' = \{c_2, \ldots, c_m\}$ and in addition, define the following two sets:

$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\},$$

and

$$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\}.$$

It is easy to verify that $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Additionally, since $Y_0 = \mathcal{H}_{C'}$, using the induction assumption (applied on $\mathcal{H}$ and C') we have that

$$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}|.$$

Next, define $\mathcal{H}' \subseteq \mathcal{H}$ to be

$$\begin{aligned}\mathcal{H}' &= \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } (1 - h'(c_1), h'(c_2), \ldots, h'(c_m)) \\ &= (h(c_1), h(c_2), \ldots, h(c_m))\},\end{aligned}$$

namely, $\mathcal{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1. Using this definition, it is clear that if $\mathcal{H}'$ shatters a set $B \subseteq C'$ then it also shatters the set $B \cup \{c_1\}$ and vice versa. Combining this with the fact that $Y_1 = \mathcal{H}'_{C'}$ and using the inductive assumption (now applied on $\mathcal{H}'$ and C') we obtain that

$$\begin{aligned}|Y_1| &= |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B\}| = |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \\ &= |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}' \text{ shatters } B\}| \leq |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}|.\end{aligned}$$

Overall, we have shown that

$$\begin{aligned}|\mathcal{H}_C| &= |Y_0| + |Y_1| \\ &\leq |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|,\end{aligned}$$

which concludes our proof.

6.5.2 Uniform Convergence for Classes of Small Effective Size

In this section we prove that if $\mathcal{H}$ has small effective size then it enjoys the uniform convergence property. Formally,

THEOREM 6.11 *Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$ we have*

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}.$$

Before proving the theorem, let us first conclude the proof of Theorem 6.7.

Proof of Theorem 6.7 It suffices to prove that if the VC-dimension is finite then the uniform convergence property holds. We will prove that

$$m^{\text{UC}}_{\mathcal{H}}(\epsilon, \delta) \leq 4\,\frac{16d}{(\delta\epsilon)^2}\,\log\left(\frac{16d}{(\delta\epsilon)^2}\right) + \frac{16\,d\log(2e/d)}{(\delta\epsilon)^2}.$$

From Sauer's lemma we have that for $m > d$, $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$. Combining this with Theorem 6.11 we obtain that with probability of at least $1-\delta$,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}}.$$

For simplicity assume that $\sqrt{d\log(2em/d)} \geq 4$; hence,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta}\,\sqrt{\frac{2d\log(2em/d)}{m}}.$$

To ensure that the preceding is at most ϵ we need that

$$m \geq \frac{2d\log(m)}{(\delta\epsilon)^2} + \frac{2\,d\log(2e/d)}{(\delta\epsilon)^2}.$$

Standard algebraic manipulations (see Lemma A.2 in Appendix A) show that a sufficient condition for the preceding to hold is that

$$m \geq 4\,\frac{2d}{(\delta\epsilon)^2}\,\log\left(\frac{2d}{(\delta\epsilon)^2}\right) + \frac{4\,d\log(2e/d)}{(\delta\epsilon)^2}.$$

□

Remark 6.4 The upper bound on $m_{\mathcal{H}}^{\mathrm{UC}}$ we derived in the proof Theorem 6.7 is not the tightest possible. A tighter analysis that yields the bounds given in Theorem 6.8 can be found in Chapter 28.

Proof of Theorem 6.11 *

We will start by showing that

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}. \tag{6.4}$$

Since the random variable $\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|$ is nonnegative, the proof of the theorem follows directly from the preceding using Markov's inequality (see Section B.1).

To bound the left-hand side of Equation (6.4) we first note that for every $h \in \mathcal{H}$, we can rewrite $L_{\mathcal{D}}(h) = \mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h)]$, where $S' = z'_1, \ldots, z'_m$ is an additional i.i.d. sample. Therefore,

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] = \underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\left|\underset{S'\sim\mathcal{D}^m}{\mathbb{E}} L_{S'}(h) - L_S(h)\right|\right].$$

A generalization of the triangle inequality yields

$$\left|\underset{S'\sim\mathcal{D}^m}{\mathbb{E}}[L_{S'}(h) - L_S(h)]\right| \leq \underset{S'\sim\mathcal{D}^m}{\mathbb{E}}|L_{S'}(h) - L_S(h)|,$$

and the fact that supermum of expectation is smaller than expectation of supremum yields

$$\sup_{h\in\mathcal{H}}\underset{S'\sim\mathcal{D}^m}{\mathbb{E}}|L_{S'}(h) - L_S(h)| \leq \underset{S'\sim\mathcal{D}^m}{\mathbb{E}}\sup_{h\in\mathcal{H}}|L_{S'}(h) - L_S(h)|.$$

Formally, the previous two inequalities follow from Jensen's inequality. Combining all we obtain

$$\begin{aligned}\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] &\leq \underset{S,S'\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}|L_{S'}(h) - L_S(h)|\right] \\ &= \underset{S,S'\sim\mathcal{D}^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}(\ell(h,z'_i) - \ell(h,z_i))\right|\right].\end{aligned} \tag{6.5}$$

The expectation on the right-hand side is over a choice of two i.i.d. samples $S = z_1, \ldots, z_m$ and $S' = z'_1, \ldots, z'_m$. Since all of these $2m$ vectors are chosen i.i.d., nothing will change if we replace the name of the random vector z_i with the name of the random vector z'_i. If we do it, instead of the term $(\ell(h, z'_i) - \ell(h, z_i))$ in Equation (6.5) we will have the term $-(\ell(h, z'_i) - \ell(h, z_i))$. It follows that for every $\boldsymbol{\sigma} \in \{\pm 1\}^m$ we have that Equation (6.5) equals

$$\mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right]$$

Since this holds for every $\boldsymbol{\sigma} \in \{\pm 1\}^m$, it also holds if we sample each component of $\boldsymbol{\sigma}$ uniformly at random from the uniform distribution over $\{\pm 1\}$, denoted $U_{\pm}$. Hence, Equation (6.5) also equals

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right],$$

and by the linearity of expectation it also equals

$$\mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right].$$

Next, fix S and S', and let C be the instances appearing in S and S'. Then, we can take the supremum only over $h \in \mathcal{H}_C$. Therefore,

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \left[\max_{h \in \mathcal{H}_C} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right]. \end{aligned}$$

Fix some $h \in \mathcal{H}_C$ and denote $\theta_h = \frac{1}{m} \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i))$. Since $\mathbb{E}[\theta_h] = 0$ and θ_h is an average of independent variables, each of which takes values in $[-1, 1]$, we have by Hoeffding's inequality that for every $\rho > 0$,

$$\mathbb{P}[|\theta_h| > \rho] \leq 2 \exp\left(-2\, m\, \rho^2\right).$$

Applying the union bound over $h \in \mathcal{H}_C$, we obtain that for any $\rho > 0$,

$$\mathbb{P}\left[\max_{h \in \mathcal{H}_C} |\theta_h| > \rho \right] \leq 2\, |\mathcal{H}_C| \exp\left(-2\, m\, \rho^2\right).$$

Finally, Lemma A.4 in Appendix A tells us that the preceding implies

$$\mathbb{E}\left[\max_{h \in \mathcal{H}_C} |\theta_h| \right] \leq \frac{4 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{2m}}.$$

Combining all with the definition of $\tau_{\mathcal{H}}$, we have shown that

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \right] \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}.$$

6.6 Summary

The fundamental theorem of learning theory characterizes PAC learnability of classes of binary classifiers using VC-dimension. The VC-dimension of a class is a combinatorial property that denotes the maximal sample size that can be shattered by the class. The fundamental theorem states that a class is PAC learnable if and only if its VC-dimension is finite and specifies the sample complexity required for PAC learning. The theorem also shows that if a problem is at all learnable, then uniform convergence holds and therefore the problem is learnable using the ERM rule.

6.7 Bibliographic remarks

The definition of VC-dimension and its relation to learnability and to uniform convergence is due to the seminal work of Vapnik & Chervonenkis (1971). The relation to the definition of PAC learnability is due to Blumer, Ehrenfeucht, Haussler & Warmuth (1989).

Several generalizations of the VC-dimension have been proposed. For example, the fat-shattering dimension characterizes learnability of some regression problems (Kearns, Schapire & Sellie 1994, Alon, Ben-David, Cesa-Bianchi & Haussler 1997, Bartlett, Long & Williamson 1994, Anthony & Bartlet 1999), and the Natarajan dimension characterizes learnability of some multiclass learning problems (Natarajan 1989). However, in general, there is no equivalence between learnability and uniform convergence. See (Shalev-Shwartz, Shamir, Srebro & Sridharan 2010, Daniely, Sabato, Ben-David & Shalev-Shwartz 2011).

Sauer's lemma has been proved by Sauer in response to a problem of Erdos (Sauer 1972). Shelah (with Perles) proved it as a useful lemma for Shelah's theory of stable models (Shelah 1972). Gil Kalai tells[1] us that at some later time, Benjy Weiss asked Perles about such a result in the context of ergodic theory, and Perles, who forgot that he had proved it once, proved it again. Vapnik and Chervonenkis proved the lemma in the context of statistical learning theory.

6.8 Exercises

1. Show the following monotonicity property of VC-dimension: For every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$ then $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$.
2. Given some finite domain set, $\mathcal{X}$, and a number $k \leq |\mathcal{X}|$, figure out the VC-dimension of each of the following classes (and prove your claims):
   1. $\mathcal{H}^{\mathcal{X}}_{=k} = \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k\}$. That is, the set of all functions that assign the value 1 to exactly k elements of $\mathcal{X}$.

[1] http://gilkalai.wordpress.com/2008/09/28/extremal-combinatorics-iii-some-basic-theorems

2. $\mathcal{H}_{at-most-k} = \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| \leq k \text{ or } |\{x : h(x) = 0\}| \leq k\}$.

3. Let $\mathcal{X}$ be the Boolean hypercube $\{0,1\}^n$. For a set $I \subseteq \{1, 2, \ldots, n\}$ we define a *parity function* h_I as follows. On a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$,

$$h_I(\mathbf{x}) = \left(\sum_{i \in I} x_i \right) \bmod 2 \, .$$

(That is, h_I computes parity of bits in I.) What is the VC-dimension of the class of all such parity functions, $\mathcal{H}_{n\text{-parity}} = \{h_I \ : \ I \subseteq \{1, 2, \ldots, n\}\}$?

4. We proved Sauer's lemma by proving that for every class $\mathcal{H}$ of finite VC-dimension d, and every subset A of the domain,

$$|\mathcal{H}_A| \leq |\{B \subseteq A \ : \ \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{|A|}{i} \, .$$

Show that there are cases in which the previous two inequalities are strict (namely, the $\leq$ can be replaced by $<$) and cases in which they can be replaced by equalities. Demonstrate all four combinations of $=$ and $<$.

5. **VC-dimension of axis aligned rectangles in $\mathbb{R}^d$:** Let $\mathcal{H}^d_{\text{rec}}$ be the class of axis aligned rectangles in $\mathbb{R}^d$. We have already seen that $\text{VCdim}(\mathcal{H}^2_{\text{rec}}) = 4$. Prove that in general, $\text{VCdim}(\mathcal{H}^d_{\text{rec}}) = 2d$.

6. **VC-dimension of Boolean conjunctions:** Let $\mathcal{H}^d_{con}$ be the class of Boolean conjunctions over the variables $x_1, \ldots, x_d$ ($d \geq 2$). We already know that this class is finite and thus (agnostic) PAC learnable. In this question we calculate $\text{VCdim}(\mathcal{H}^d_{con})$.

   1. Show that $|\mathcal{H}^d_{con}| \leq 3^d + 1$.
   2. Conclude that $\text{VCdim}(\mathcal{H}) \leq d \log 3$.
   3. Show that $\mathcal{H}^d_{con}$ shatters the set of unit vectors $\{\mathbf{e}_i : i \leq d\}$.
   4. **(**)** Show that $\text{VCdim}(\mathcal{H}^d_{con}) \leq d$.
      Hint: Assume by contradiction that there exists a set $C = \{c_1, \ldots, c_{d+1}\}$ that is shattered by $\mathcal{H}^d_{con}$. Let $h_1, \ldots, h_{d+1}$ be hypotheses in $\mathcal{H}^d_{con}$ that satisfy

$$\forall i, j \in [d+1], \ h_i(c_j) = \begin{cases} 0 & i = j \\ 1 & \text{otherwise} \end{cases}$$

      For each $i \in [d+1]$, h_i (or more accurately, the conjunction that corresponds to h_i) contains some literal ℓ_i which is false on c_i and true on c_j for each $j \neq i$. Use the Pigeonhole principle to show that there must be a pair $i < j \leq d+1$ such that ℓ_i and ℓ_j use the same x_k and use that fact to derive a contradiction to the requirements from the conjunctions h_i, h_j.
   5. Consider the class $\mathcal{H}^d_{mcon}$ of monotone Boolean conjunctions over $\{0,1\}^d$. Monotonicity here means that the conjunctions do not contain negations.

As in $\mathcal{H}^d_{con}$, the empty conjunction is interpreted as the all-positive hypothesis. We augment $\mathcal{H}^d_{mcon}$ with the all-negative hypothesis h^-. Show that $\text{VCdim}(\mathcal{H}^d_{mcon}) = d$.

7. We have shown that for a finite hypothesis class $\mathcal{H}$, $\text{VCdim}(\mathcal{H}) \leq \lfloor \log(|\mathcal{H}|) \rfloor$. However, this is just an upper bound. The VC-dimension of a class can be much lower than that:
   1. Find an example of a class $\mathcal{H}$ of functions over the real interval $\mathcal{X} = [0, 1]$ such that $\mathcal{H}$ is infinite while $\text{VCdim}(\mathcal{H}) = 1$.
   2. Give an example of a finite hypothesis class $\mathcal{H}$ over the domain $\mathcal{X} = [0, 1]$, where $\text{VCdim}(\mathcal{H}) = \lfloor \log_2(|\mathcal{H}|) \rfloor$.
8. **(*)** It is often the case that the VC-dimension of a hypothesis class equals (or can be bounded above by) the number of parameters one needs to set in order to define each hypothesis in the class. For instance, if $\mathcal{H}$ is the class of axis aligned rectangles in $\mathbb{R}^d$, then $\text{VCdim}(\mathcal{H}) = 2d$, which is equal to the number of parameters used to define a rectangle in $\mathbb{R}^d$. Here is an example that shows that this is not always the case. We will see that a hypothesis class might be very complex and even not learnable, although it has a small number of parameters.

   Consider the domain $\mathcal{X} = \mathbb{R}$, and the hypothesis class

   $$\mathcal{H} = \{x \mapsto \lceil \sin(\theta x) \rceil : \theta \in \mathbb{R}\}$$

   (here, we take $\lceil -1 \rceil = 0$). Prove that $\text{VCdim}(\mathcal{H}) = \infty$.

   Hint: There is more than one way to prove the required result. One option is by applying the following lemma: If $0.x_1x_2x_3\ldots$, is the binary expansion of $x \in (0, 1)$, then for any natural number m, $\lceil \sin(2^m \pi x) \rceil = (1 - x_m)$, provided that $\exists k \geq m$ s.t. $x_k = 1$.
9. Let $\mathcal{H}$ be the class of signed intervals, that is,
   $\mathcal{H} = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$ where

   $$h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}$$

   Calculate $\text{VCdim}(\mathcal{H})$.
10. Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0, 1\}$.
    1. Prove that if $\text{VCdim}(\mathcal{H}) \geq d$, for any d, then for some probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, for every sample size, m,

       $$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{d - m}{2d}$$

       Hint: Use Exercise 3 in Chapter 5.
    2. Prove that for every $\mathcal{H}$ that is PAC learnable, $\text{VCdim}(\mathcal{H}) < \infty$. (Note that this is the implication $3 \to 6$ in Theorem 6.7.)
11. **VC of union:** Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = \max_i \text{VCdim}(\mathcal{H}_i)$ and assume for simplicity that $d \geq 3$.

1. Prove that

$$\text{VCdim}\left(\cup_{i=1}^{r}\mathcal{H}_i\right) \leq 4d\log(2d) + 2\log(r) \ .$$

Hint: Take a set of k examples and assume that they are shattered by the union class. Therefore, the union class can produce all 2^k possible labelings on these examples. Use Sauer's lemma to show that the union class cannot produce more than rk^d labelings. Therefore, $2^k < rk^d$. Now use Lemma A.2.

2. **(*)** Prove that for $r = 2$ it holds that

$$\text{VCdim}\left(\mathcal{H}_1 \cup \mathcal{H}_2\right) \leq 2d + 1.$$

12. **Dudley classes:** In this question we discuss an algebraic framework for defining concept classes over $\mathbb{R}^n$ and show a connection between the VC dimension of such classes and their algebraic properties. Given a function $f : \mathbb{R}^n \to \mathbb{R}$ we define the corresponding function, $POS(f)(x) = \mathbb{1}_{[f(x)>0]}$. For a class $\mathcal{F}$ of real valued functions we define a corresponding class of functions $POS(\mathcal{F}) = \{POS(f) : f \in \mathcal{F}\}$. We say that a family, $\mathcal{F}$, of real valued functions is *linearly closed* if for all $f, g \in \mathcal{F}$ and $r \in \mathbb{R}$, $(f + rg) \in \mathcal{F}$ (where addition and scalar multiplication of functions are defined point wise, namely, for all $x \in \mathbb{R}^n$, $(f + rg)(x) = f(x) + rg(x)$). Note that if a family of functions is linearly closed then we can view it as a vector space over the reals. For a function $g : \mathbb{R}^n \to \mathbb{R}$ and a family of functions $\mathcal{F}$, let $\mathcal{F}+g \stackrel{\text{def}}{=} \{f+g : f \in \mathcal{F}\}$. Hypothesis classes that have a representation as $POS(\mathcal{F}+g)$ for some vector space of functions $\mathcal{F}$ and some function g are called *Dudley classes*.
    1. Show that for every $g : \mathbb{R}^n \to \mathbb{R}$ and every vector space of functions $\mathcal{F}$ as defined earlier, $\text{VCdim}(POS(\mathcal{F}+g)) = \text{VCdim}(POS(\mathcal{F}))$.
    2. **(**)** For every linearly closed family of real valued functions $\mathcal{F}$, the VC-dimension of the corresponding class $POS(\mathcal{F})$ equals the linear dimension of $\mathcal{F}$ (as a vector space). *Hint:* Let $f_1, \ldots, f_d$ be a basis for the vector space $\mathcal{F}$. Consider the mapping $x \mapsto (f_1(x), \ldots, f_d(x))$ (from $\mathbb{R}^n$ to $\mathbb{R}^d$). Note that this mapping induces a matching between functions over $\mathbb{R}^n$ of the form $POS(f)$ and homogeneous linear halfspaces in $\mathbb{R}^d$ (the VC-dimension of the class of homogeneous linear halfspaces is analyzed in Chapter 9).
    3. Show that each of the following classes can be represented as a Dudley class:
        1. The class HS_n of halfspaces over $\mathbb{R}^n$ (see Chapter 9).
        2. The class HHS_n of all homogeneous halfspaces over $\mathbb{R}^n$ (see Chapter 9).
        3. The class B_d of all functions defined by (open) balls in $\mathbb{R}^d$. Use the Dudley representation to figure out the VC-dimension of this class.
        4. Let P_n^d denote the class of functions defined by polynomial inequalities of degree $\leq d$, namely,

$$P_n^d = \{h_p : p \text{ is a polynomial of degree } \leq d \text{ in the variables } x_1, \ldots, x_n\},$$

where, for $\mathbf{x} = (x_1 \ldots, x_n)$, $h_p(\mathbf{x}) = \mathbb{1}_{[p(\mathbf{x}) \geq 0]}$ (the degree of a multivariable polynomial is the maximal sum of variable exponents over all of its terms. For example, the degree of $p(\mathbf{x}) = 3x_1^3 x_2^2 + 4x_3 x_7^2$ is 5).

1. Use the Dudley representation to figure out the VC-dimension of the class P_1^d – the class of all d-degree polynomials over $\mathbb{R}$.
2. Prove that the class of all polynomial classifiers over $\mathbb{R}$ has infinite VC-dimension.
3. Use the Dudley representation to figure out the VC-dimension of the class P_n^d (as a function of d and n).