Problem 1. *Given a finite domain set $\mathcal{X}$ and an integer $k \le |\mathcal{X}|$, determine the* VC-dimension *of each of the following hypothesis classes, and prove your claims.*

(1)

$$\mathcal{H}_{=k}^{\mathcal{X}} = \left\{ h \in \{0,1\}^{\mathcal{X}} \;\middle|\; \left|\{x \in \mathcal{X} : h(x) = 1\}\right| = k \right\}.$$

That is, the set of all functions that assign the value 1 *to* exactly k *elements of* $\mathcal{X}$*.*

(2)

$$\mathcal{H}_{\textit{at-most-k}} = \left\{ h \in \{0,1\}^{\mathcal{X}} \;\middle|\; \left|\{x \in \mathcal{X} : h(x) = 1\}\right| \le k \textit{ or } \left|\{x \in \mathcal{X} : h(x) = 0\}\right| \le k \right\}.$$

Solution

Let $\mathcal{X}$ be a finite domain set.

1. The class $\mathcal{H}_{=k}^{\mathcal{X}}$

$$\mathcal{H}_{=k}^{\mathcal{X}} = \{h : \mathcal{X} \to \{0,1\} \mid |\{x \in \mathcal{X} : h(x) = 1\}| = k\}.$$

Lower bound

Let $C \subseteq \mathcal{X}$ with $|C| = k$. Any labeling of C can be extended to $\mathcal{X}$ by assigning additional 1 labels outside C so that exactly k points are labeled 1. Hence, C is shattered and

$$\mathrm{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) \ge k.$$

Upper bound

Any set of size $k+1$ admits the labeling where all points are labeled 1, which cannot be realized by a hypothesis that assigns exactly k ones. Thus,

$$\mathrm{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) \le k.$$

Conclusion

$$\boxed{\mathrm{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) = k.}$$

2. The class $\mathcal{H}_{\text{at-most-}k}$

$$\mathcal{H}_{\text{at-most-}k} = \{h : \mathcal{X} \to \{0,1\} \mid |\{x : h(x) = 1\}| \le k \text{ or } |\{x : h(x) = 0\}| \le k\}.$$

Lower bound

Let $C \subseteq \mathcal{X}$ with $|C| = 2k$. For any labeling of C:

- if the number of 1's is at most k, the first condition holds;
- if the number of 1's is at least $k+1$, then the number of 0's is at most $k-1$, so the second condition holds.

Thus C is shattered and

$$\mathrm{VCdim}(\mathcal{H}_{\text{at-most-}k}) \ge 2k.$$

Upper bound

Let C have $2k+1$ points and label $k+1$ points as 1 and $k+1$ points as 0. This labeling violates both defining conditions, hence is unrealizable. Therefore,

$$\mathrm{VCdim}(\mathcal{H}_{\text{at-most-}k}) \leq 2k.$$

Conclusion

$$\boxed{\mathrm{VCdim}(\mathcal{H}_{\text{at-most-}k}) = 2k.}$$

Problem 2. ***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

$$\mathrm{VCdim}(\mathcal{H}^2_{\text{rec}}) = 4.$$

Prove that, in general,

$$\mathrm{VCdim}(\mathcal{H}^d_{\text{rec}}) = 2d.$$

Solution

Lower bound

For each coordinate $i \in \{1, \ldots, d\}$ define

$$p_i^- = (0, \ldots, 0, -1, 0, \ldots, 0), \qquad p_i^+ = (0, \ldots, 0, 1, 0, \ldots, 0).$$

Let $C = \{p_1^-, p_1^+, \ldots, p_d^-, p_d^+\}$. Each coordinate boundary can be adjusted independently, so every labeling of C is realizable. Hence

$$\mathrm{VCdim}(\mathcal{H}^d_{\text{rec}}) \geq 2d.$$

Upper bound

Let $C \subset \mathbb{R}^d$ with $|C| = 2d+1$. For each coordinate i, choose points $x^{(i)}_{\min}, x^{(i)}_{\max} \in C$ such that

$$x^{(i)}_{\min,i} = \min_{x \in C} x_i, \qquad x^{(i)}_{\max,i} = \max_{x \in C} x_i.$$

At most $2d$ points are extreme, so there exists $x^* \in C$ that is not extreme in any coordinate. Then for all i,

$$x^{(i)}_{\min,i} < x^*_i < x^{(i)}_{\max,i}.$$

Any rectangle containing all extreme points must contain x^*, so C cannot be shattered. Thus

$$\mathrm{VCdim}(\mathcal{H}^d_{\text{rec}}) \leq 2d.$$

Conclusion

$$\boxed{\text{VCdim}(\mathcal{H}_{\text{rec}}^d) = 2d.}$$

Problem 3. *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

$$\mathcal{H} = \{x \mapsto [\sin(\theta x)] : \theta \in \mathbb{R}\},$$

(where we take $[-1] = 0$).

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,

$$[\sin(2^m \pi x)] = 1 - x_m,$$

provided that there exists $k \geq m$ such that $x_k = 1$.

Solution

We show that for every $m \in \mathbb{N}$ there exists a set of m points in $\mathbb{R}$ that is shattered by $\mathcal{H}$. This implies that the VC-dimension of $\mathcal{H}$ is infinite.

Key Lemma (Binary Expansion Trick)

Let $x \in (0,1)$ have binary expansion

$$x = 0.x_1x_2x_3\ldots$$

and assume that x is not a dyadic rational (i.e., its binary expansion does not terminate). Then for every $m \in \mathbb{N}$,

$$[\sin(2^m \pi x)] = 1 - x_m.$$

Proof of the Lemma

Write

$$2^m x = n + r,$$

where $n = \lfloor 2^m x \rfloor \in \mathbb{Z}$ and $r \in (0,1)$ is the fractional part. Multiplication by 2^m shifts the binary point m places, so the parity of n is determined by the m-th binary digit x_m. Hence,

$$(-1)^n = (-1)^{x_m}.$$

Now,

$$\sin(2^m \pi x) = \sin(\pi(n + r)) = (-1)^n \sin(\pi r).$$

Since $r \in (0, 1)$, we have $\sin(\pi r) > 0$, and therefore

$$[\sin(2^m \pi x)] = 1 \iff (-1)^n > 0 \iff x_m = 0.$$

This proves the lemma.

Shattering Arbitrarily Large Sets

Fix $m \in \mathbb{N}$. Choose m distinct non-dyadic points $x^{(1)}, \ldots, x^{(m)} \in (0, 1)$ whose binary expansions encode arbitrary labels. Given any labeling $(y_1, \ldots, y_m) \in \{0, 1\}^m$, choose a point

$$z \in (0, 1) \quad \text{with binary expansion} \quad z = 0.(1 - y_1)(1 - y_2) \cdots (1 - y_m)1000 \cdots$$

and set $\theta = 2^m \pi$.

By the lemma,

$$h_\theta(z) = [\sin(2^m \pi z)] = 1 - z_m = y_m,$$

and similarly each desired label can be realized. Thus every labeling of a set of size m is realizable by some hypothesis in $\mathcal{H}$.

Since m was arbitrary, we conclude that

$$\boxed{\mathrm{VCdim}(\mathcal{H}) = \infty.}$$