Feb 2, 2023

DEFINITION 6.5 (VC-dimension) The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, 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.

6.3 Examples

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

Toy Example:

$\{\mathbb{1}_{(a,b)} : a,b \in \mathbb{R}, a<b\}$

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 $\mathrm{VCdim}(\mathcal{H}) \geq 2$. 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 $\mathrm{VCdim}(\mathcal{H}) = 2$.

Exercise: Let $\mathcal{F}$ be a linear space of functions $f : \mathcal{X} \to \mathbb{R}$ and $\mathcal{H}$, the space of binary functions on $\mathcal{X} \times \mathbb{R}$ defined by

$$\mathcal{H} := \{ g(x,t) = \mathbb{1}_{\{f(x) \geq t\}} : f \in \mathcal{F} \}$$

prove that, if $\mathcal{F}$ is finite-dimensional, then $\mathrm{VCdim}(\mathcal{H}) \leq \dim(\mathcal{F}) + 1$

proof: Let $\dim(\mathcal{F}) = d$, by the definition of VC-dimension, we need to establish that any arbitrary points of size $d+2$ can't be shattered by $\mathcal{H}$.

$\{(x_1,t_1), (x_2,t_2), \cdots, (x_d,t_d), (x_{d+1},t_{d+1}), (x_{d+2},t_{d+2})\}$: a set of size $d+2$

Consider the subset of $\mathbb{R}^{d+2}$

$$\mathcal{P} = \left\{ \begin{bmatrix} f(x_1) - t_1 \\ \vdots \\ f(x_{d+2}) - t_{d+2} \end{bmatrix} : f \in \mathcal{F} \right\}$$

Since $\mathcal{F}$ is a linear space of dimension d, then it has a basis, say $\{\phi_1, \cdots, \phi_d\}$. For each vector in $\mathcal{P}$, we have

$$\begin{bmatrix} f(x_1) - t_1 \\ f(x_2) - t_2 \\ \vdots \\ f(x_{d+2}) - t_{d+2} \end{bmatrix} = \sum_{i=1}^{d} \alpha_i \begin{bmatrix} \phi_i(x_1) \\ \vdots \\ \phi_i(x_{d+2}) \end{bmatrix} + (-1) \begin{bmatrix} t_1 \\ \vdots \\ t_{d+2} \end{bmatrix}$$

$\Rightarrow$ $\mathrm{span}(\mathcal{P})$ is a subspace of $\mathbb{R}^{d+2}$ w/ dimension at most $d+1$

$$\sum_{i=1}^{d+2} v_i (f(x_i) - t_i) = 0 \quad \text{for some nonzero vector } v = \begin{bmatrix} v_1 \\ \vdots \\ v_{d+2} \end{bmatrix} \in \mathbb{R}^{d+2}$$

WLOG, we can assume there is at least one v_i that is strictly positive, otherwise v can be replaced by $-v$. Thus

$$\sum_{i: v_i > 0} v_i (f(x_i) - t_i) = - \sum_{i: v_i \leq 0} v_i (f(x_i) - t_i) \qquad (*)$$

Construct a binary vector $a = (a_1, \cdots, a_{d+2})$ s.t.

$$a_i = \begin{cases} 1, & \text{if } v_i > 0; \\ 0, & \text{if } v_i \leq 0. \end{cases}$$

there must exist $i \in \{1, \cdots, d+2\}$ s.t. $g(x_i,t_i) \neq a_i$, otherwise $\begin{matrix} \text{LHS } (*) > 0 \\ \text{RHS } (*) \leq 0 \end{matrix}$, which is a contradiction

$\Rightarrow$ $\mathrm{VC\text{-}dim}(\mathcal{H}) \leq \dim(\mathcal{F}) + 1$