TD 6 – Vapnik-Chervonenkis Dimension

Mathematics of data

04/12/24

Exercise 1. Let S be a set of classifiers $\mathbb{R}^d \to \{0,1\}$. For any $k \geq 1$, denote

$$C(S,k) := \sup_{x_1,\ldots,x_k \in \mathbb{R}} \operatorname{Card}\{(f(x_1),\ldots,f(x_k)) : f \in S\}.$$

We say that S is a Vapnik-Chervonenkis (VC) class if $V(S) := \sup\{k \geq 1 : C(S,k) = 2^k\} < +\infty$. If this is the case, $V(S)$ is called the Vapnik-Chervonenkis dimension of S.

1. Let S be a finite set. Is S a VC class? Upper bound its VC dimension.

For any collection of measurable subsets of $\mathcal{X}$, define the model

$$S_{\mathcal{A}} = \{\mathbf{1}_A : A \in \mathcal{A}\}.$$

For each case below, say whether $S_{\mathcal{A}}$ is a VC class. If this is the case, determine its VC dimension.

2. $\mathcal{X} = \mathbb{R}$ and $\mathcal{A}$ is the set of half-lines of the form $(-\infty, a]$ with $a \in \mathbb{R}$.
3. $\mathcal{X} = \mathbb{R}$ and $\mathcal{A}$ is the set of half-lines of $\mathbb{R}$.
4. $\mathcal{X} = \mathbb{R}$ and $\mathcal{A}$ is the set of intervals of $\mathbb{R}$.
5. $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{A} = \{(-\infty, a_1] \times \cdots \times (-\infty, a_d] : a_1, \ldots, a_d \in \mathbb{R}\}$.
6. $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{A} = \{[a_1, b_1] \times \cdots \times [a_d, b_d] : a_1, \ldots, a_d, b_1, \ldots, b_d \in \mathbb{R}\}$.
7. $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{A}$ is the set of convex subsets of $\mathbb{R}^d$.

Solutions

Exercise 1.

1. $C(S,k) \leq \operatorname{Card} S$ for any k, so $V(S) \leq \log_2(\operatorname{Card} S)$.
2. $C(S_{\mathcal{A}},k) = k+1$, so $V(S_{\mathcal{A}}) = 1$.
3. $C(S_{\mathcal{A}},k) = 2k$, so $V(S_{\mathcal{A}}) = 2$.
4. $C(S_{\mathcal{A}},k) = 1 + k(k+1)/2$, so $V(S_{\mathcal{A}}) = 2$.
5. For $i = 1, \ldots, d$, denote $e_i = (0, \ldots, 1, \ldots, 0)$ with 1 at coordinate i. Then, any combination of labels can be assigned to $(e_1, \ldots, e_d)$ by an element of $S_{\mathcal{A}}$. Indeed, if we want to assign 1 to e_i for $i \in I$ for some subset $I \subset \{1, \ldots, d\}$, it suffices to set $a_i = 1$ if $i \in I$ and $a_i = 0$ otherwise. Therefore, $V(S_{\mathcal{A}}) \geq d$. Now let $x_1, \ldots, x_n \in \mathbb{R}^d$ such that $n \geq d+1$. There exists at least one index $1 \leq j \leq n$ such that

$$\text{for all } i = 1, \ldots, d, \quad (x_j)_i \leq \max_{k \neq j} (x_k)_i.$$

   Then, no element of $S_{\mathcal{A}}$ can assign 0 to x_j and 1 to all the other points x_k, $k \neq j$. This proves that $V(S_{\mathcal{A}}) = d$.

6. With the notations above, any combination of labels can be assigned to $(e_1, \ldots, e_d, -e_1, \ldots, -e_d)$. Indeed, if we want to assign 1 to e_i for $i \in I \subset \{1, \ldots, d\}$ and to $-e_j$ for $j \in J \subset \{1, \ldots, d\}$, it suffices to set $a_j = -1$ if $j \in J$ and 0 otherwise, and $b_i = 1$ if $i \in I$ and 0 otherwise. Thereforer, $V(S_{\mathcal{A}}) \geq 2d$. Now if $x_1, \ldots, x_n \in \mathbb{R}^d$ with $n \geq 2d+1$, there exists at least one index $1 \leq j \leq n$ such that

$$\text{for all } i = 1, \ldots, d, \quad (x_j)_i \leq \max_{k \neq j} (x_k)_i$$

   and

$$\text{for all } i = 1, \ldots, d, \quad (x_j)_i \geq \min_{k \neq j} (x_k)_i.$$

   Then, no element of $S_{\mathcal{A}}$ can assign 0 to x_j and 1 to all the other points x_k, $k \neq j$. This proves that $V(S_{\mathcal{A}}) = 2d$.

7. $S_{\mathcal{A}}$ is not a VC class (take distinct $x_1, \ldots, x_n$ on a 2D circle, and choose A to be the convex hull of the points you mant to map to 1.)