STATISTICAL MACHINE LEARNING (WS2024)
VC DIMENSION

Assignment 1. What is the VC dimension of the hypothesis space of thresholding classifiers $\mathcal{H} = \{h(x) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$?

Assignment 2. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a finite hypothesis space. Show that the VC dimension of $\mathcal{H}$ is not greater than $\log_2(|\mathcal{H}|)$, where $|\mathcal{H}|$ is the number of hypothesis in $\mathcal{H}$.

Assignment 3. Let us consider the space of all linear classifiers mapping $\boldsymbol{x} \in \mathbb{R}^d$ to $\{-1, +1\}$, that is

$$\mathcal{H} = \left\{h(\boldsymbol{x}; \boldsymbol{w}, b) = \text{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \mid (\boldsymbol{w}, b) \in (\mathbb{R}^d \times \mathbb{R})\right\} .$$

Show that the VC dimension of $\mathcal{H}$ is $d + 1$.

Hint: The proof has two steps:

(1) *Show that the VC dimension is at least* $n + 1$ *by constructing* $n + 1$ *points that are shatted by* $\mathcal{H}$.
(2) *Show that the VC dimension is less than* $n + 2$ *by proving that* $n + 2$ *points cannot be shattered by* $\mathcal{H}$.

Assignment 4. Let $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$ and $\mathcal{T}^m = ((x^1, y^1), \ldots, (x^m, y^m)) \in (\mathcal{X} \times \mathcal{Y})^m$ a training set drawn from i.i.d. random variables with distribution $p(x, y)$. Then, the following inequality holds for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 4 \left(\frac{2\, e\, m}{d} \right)^d e^{-\frac{m\, \varepsilon^2}{8}} ,$$

where $R^{0/1}(h) = \mathbb{E}_{(x,y)\sim p}([\![y \neq h(x)]\!])$ and $R^{0/1}_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^m [\![y^i \neq h(x^i)]\!]$. Show that this implies the ULLN for the class of strategies $\mathcal{H}$.