STATISTICAL MACHINE LEARNING (WS2025/26)
SEMINAR: VC DIMENSION

Assignment 1. Let $\mathcal{H} = \{h_1, \dots, h_H\}$ be a finite hypothesis class. Prove that the Uniform Law of Large Numbers (ULLN) holds for $\mathcal{H}$. Specifically, show that there exists a function $m^{\mathcal{H}}_{\text{ul}} : (0,1) \times (0,1) \to \mathbb{N}$ such that for any i.i.d. sample $T_m = \big((x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \dots, m\big)$ of size $m \geq m^{\mathcal{H}}_{\text{ul}}(\varepsilon, \delta)$ drawn from the distribution $p(x,y)$, the following holds:

$$\mathbb{P}\bigg(\max_{h \in \mathcal{H}} \big|\hat{R}(T_m, h) - R(p,h)\big| > \varepsilon\bigg) \leq \delta \,,$$

where

$$R(p,h) = \mathbb{E}_{(x,y)\sim p}\big[\ell(y, h(x))\big] \quad \text{and} \quad \hat{R}(T_m, h) = \frac{1}{m}\sum_{i=1}^{m} \ell\big(y_i, h(x_i)\big)$$

denote the true and empirical risks, respectively. Assume the 0–1 loss function $\ell(y, y') = [\![y \neq y']\!]$.

Solution 1.

$$\begin{aligned}
\mathbb{P}\Big(\max_{h\in\mathcal{H}} \big|R(p,h) - \hat{R}(T_m,h)\big| \geq \varepsilon\Big) &\overset{(1)}{=} \mathbb{P}\left(\begin{array}{cc} \big|R(p,h^1) - \hat{R}(T_m,h^1)\big| \geq \varepsilon & \text{or} \\ \big|R(p,h^2) - \hat{R}(T_m,h^2)\big| \geq \varepsilon & \text{or} \\ \vdots & \\ \big|R(p,h^H) - \hat{R}(T_m,h^H)\big| \geq \varepsilon & \end{array}\right) \\
&\overset{(2)}{\leq} \sum_{h\in\mathcal{H}} \mathbb{P}\Big(\big|R(p,h) - \hat{R}(T_m,h)\big| \geq \varepsilon\Big) \\
&\overset{(3)}{\leq} 2\,|\mathcal{H}|\, e^{-2m\varepsilon^2}
\end{aligned}$$

(1) $a \geq \varepsilon \;\text{ or }\; b \geq \varepsilon \iff \max\{a,b\} \geq \varepsilon$

(2) Union bound: $\mathbb{P}\Big(A_1 \text{ or } A_2 \text{ or} \cdots \text{or } A_n\Big) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$

(3) Hoeffding inequality: $\mathbb{P}\Big(\big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon\Big) \leq 2\, e^{-2\,m\,\varepsilon^2}$

By setting $2\,|\mathcal{H}|\, e^{-2m\varepsilon^2} = \delta$ and solving for m, we get

$$m^{\mathcal{H}}_{\text{ul}}(\varepsilon, \delta) = \frac{1}{2\varepsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right) \qquad \Rightarrow \qquad \mathbb{P}\Big(\max_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon\Big) \leq \delta$$

Assignment 2. Let us consider the class of linear classifiers mapping $\boldsymbol{x} \in \mathbb{R}^d$ to $\{-1, +1\}$, that is

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

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

Solution 2. The proof has two steps:

(1) Show that $d_{\mathrm{VC}}(\mathcal{H}) \geq d+1$. We construct a set of $d+1$ input points and show that they can be shattered.
(2) Show that $d_{\mathrm{VC}}(\mathcal{H}) < d+2$. We show that no set of $d+2$ input points can be shattered.

1) Lower bound: $d_{\mathrm{VC}}(\mathcal{H}) \geq d+1$. We construct $d+1$ points in $\mathbb{R}^d$ that can be shattered by linear classifiers. Let the first d points be the standard basis vectors:

$$\boldsymbol{x}_i = [0, \ldots, \underbrace{1}_{i\text{-th coordinate}}, \ldots, 0], \quad \forall i \in 1, \ldots, d.$$

Let the last point be the origin,

$$\boldsymbol{x}_{d+1} = \mathbf{0}.$$

Fix an arbitrary sequence of labels $(y_1, y_2, \ldots, y_{d+1}) \in \{-1, +1\}^{d+1}$. We construct the parameters of the linear classifier as

$$\boldsymbol{w} = [y_1, y_2, \ldots, y_d], \qquad b = \tfrac{1}{2} y_{d+1}.$$

Now verify that the classifier

$$h(\boldsymbol{x}) = \operatorname{sign}\left(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b\right)$$

correctly predicts all $d+1$ labels. For $i \in 1, \ldots, d$,

$$h(\boldsymbol{x}_i) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) = \operatorname{sign}\left(y_i + \tfrac{1}{2} y_{d+1}\right) = y_i.$$

For the last point,

$$h(\boldsymbol{x}_{d+1}) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}_{d+1} \rangle + b) = \operatorname{sign}\left(\tfrac{1}{2} y_{d+1}\right) = y_{d+1}.$$

Thus, the constructed classifier realizes any labeling of these $d+1$ points, proving that the set is shattered. Therefore,

$$d_{\mathrm{VC}}(\mathcal{H}) \geq d+1.$$

2) Upper bound: $d_{\mathrm{VC}}(\mathcal{H}) < d+2$. We now show that no set of $d+2$ points in $\mathbb{R}^d$ can be shattered by $\mathcal{H}$. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+2}\} \subset \mathbb{R}^d$ be arbitrary. We lift these points into $\mathbb{R}^{d+1}$ by appending a constant coordinate:

$$\boldsymbol{z}_i = [\boldsymbol{x}_i, 1], \quad \forall i \in 1, \ldots, d+2.$$

We also represent affine classifiers as homogeneous ones:

$$\hat{h}(\boldsymbol{z}) = \operatorname{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{z} \rangle), \quad \text{where } \hat{\boldsymbol{w}} = [\boldsymbol{w}, b].$$

Clearly, $h(\boldsymbol{x}_i) = \hat{h}(\boldsymbol{z}_i)$ for all i. Since we have $d+2$ points in $\mathbb{R}^{d+1}$, they must be linearly dependent. Hence, there exist coefficients $(a_1, \ldots, a_{d+2})$, not all zero, such that

$$\sum_{i=1}^{d+2} a_i \boldsymbol{z}_i = \mathbf{0}.$$

Without loss of generality, assume $a_1 \neq 0$ and express

$$\boldsymbol{z}_1 = \sum_{i=2}^{d+2} a_i' \boldsymbol{z}_i, \quad \text{where } a_i' = -\frac{a_i}{a_1}.$$

Now define labels for the last $d+1$ points as

$$y_i = \operatorname{sign}(a_i'), \quad \forall i \in \{2, \ldots, d+2\}.$$

Assume that there exists a classifier $\hat{h}$ correctly predicting these labels, i.e.

$$\hat{h}(\boldsymbol{z}_i) = y_i \quad \iff \quad \mathrm{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{z}_i \rangle) = \mathrm{sign}(a'_i), \quad \forall i \in 2, \dots, d+2.$$

Then, for all $i \geq 2$, we have $a'_i \langle \hat{\mathbf{w}}, \mathbf{z}_i \rangle \geq 0$, and for at least one i, this product is strictly positive. Now, consider the prediction for $\boldsymbol{z}_1$:

$$\langle \hat{\boldsymbol{w}}, \boldsymbol{z}_1 \rangle = \sum_{i=2}^{d+2} a'_i \langle \hat{\boldsymbol{w}}, \boldsymbol{z}_i \rangle.$$

Because each term $a'_i \langle \hat{\boldsymbol{w}}, \boldsymbol{z}_i \rangle \geq 0$, the sum is non-negative and strictly positive unless all are zero. Hence,

$$\hat{h}(\boldsymbol{z}_1) = \mathrm{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{z}_1 \rangle) = +1.$$

But if we assign the label $y_1 = -1$, no classifier can realize this labeling, because $\boldsymbol{z}_1$ is forced into the positive class. Thus, the set of $d+2$ points cannot be shattered by linear classifiers, implying

$$d_{\mathrm{VC}}(\mathcal{H}) < d + 2.$$

Assignment 3. Consider a hypothesis space of classifiers

$$\mathcal{H} = \left\{ h(x; a) = \mathrm{sign}(\sin(ax)) \mid a \in \mathbb{R} \right\}.$$

That is, each $h \in \mathcal{H}$ is determined by a single parameter $a \in \mathbb{R}$ and it maps real valued input $x \in \mathbb{R}$ to a set of hidden labels $\{+1, -1\}$ based on the sign of the score $\sin(xa)$. Show that the VC dimension of $\mathcal{H}$ is infinite.

Hint: Show that for arbitrary set of labels $\{y^i \in \{+1, -1\} \mid i = 1, \dots, m\}$ *the inputs* $\{x^i = 10^{-i} \mid i = 1, \dots, m\}$ *can be predicted correctly by* $h(x; a)$ *with*

$$a = \pi \left(1 + \frac{1}{2} \sum_{i=1}^{m} (1 - y^i) 10^i \right)$$

Solution 3. The plan of the proof is to construct points $(x_1, x_2, \dots, x_m)$ such that for arbitrary labels $(y_1, y_2, \dots, y_m) \in \{-1, +1\}^m$, there will be a, which we also construct, such that all the points will be correctly classified by $h(x) = \sin(a\,x)$, i.e. the set of points is shattered. Thus, we conclude that the VC dimension of $\mathcal{H}$ is ∞.

(1) Set the m points to be

$$x_i = 10^{-i}, \quad i \in \{1, \dots, m\}.$$

and fix an arbitrary sequence of the labels $(y_1, y_2, \dots, y_m) \in \{-1, +1\}^m$.

(2) Let the classifier parameter be:

$$a = \pi + \pi \sum_{j=1}^{m} \hat{y}_i \, 10^j$$

where

$$\hat{y}_i = \frac{1 - y_i}{2} \in \{0, 1\}.$$

are auxiliary labels.

(3) We will use the indetity

$$\sin\Big(\pi(k+t)\Big) = (-1)^k \sin(\pi\, t)\,,$$

which is valid for every integer $k \in \mathbb{N}$.

(4) Let us rewrite the value of the discriminant function at x_i:

$$\begin{aligned}
a\,x_i &= \sin\left(\frac{\pi + \pi \sum_{j=1}^{m} \hat{y}_j 10^j}{10^i}\right) \\
&= \sin\left(\pi\left[10^{-i} + \sum_{j=1}^{i-1} \hat{y}_j\, 10^{j-i} + \hat{y}_i + \sum_{j=i+1}^{m} \hat{y}_j\, 10^{j-i}\right]\right) \\
&= \sin\left(\underbrace{\pi \sum_{j=i+1}^{m} \hat{y}_j\, 10^{j-i}}_{\text{even integer}} + \pi\left(10^{-i} + \sum_{j=1}^{i-1} \hat{y}_j\, 10^{j-i} + \hat{y}_i\right)\right) \\
&= \sin\left(\pi\Bigg(\underbrace{10^{-i} + \sum_{j=1}^{i-1} \hat{y}_j\, 10^{j-i}}_{\delta\in(0,1)} + \hat{y}_i\Bigg)\right) \\
&= \begin{cases} \sin(\pi\delta) & \text{if} \quad \hat{y}_i = +1 \\ \sin(\pi\delta + \pi) & \text{if} \quad \hat{y}_i = -1 \end{cases} \\
&= \operatorname{sign}(y_i)\,.
\end{aligned}$$