STATISTICAL MACHINE LEARNING (WS2022)
SEMINAR 4

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

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

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

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

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 the observation $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$ and the hidden state $y \in \mathcal{Y} = \{+1, -1\}$ be generated by a multivariate normal distribution

$$p(\boldsymbol{x}, y) = p(y)\, \frac{1}{(2\pi)^{\frac{n}{2}} \det(\boldsymbol{C}_y)^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^T \boldsymbol{C}_y^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_y)}$$

where $\boldsymbol{\mu}_y \in \mathbb{R}^n$, $y \in \mathcal{Y}$, are mean vectors, $\boldsymbol{C}_y \in \mathbb{R}^{n\times n}$, $y \in \mathcal{Y}$, are covariance matrices and $p(y)$ is a prior probability. Assume that the model parameters are unknown and we want to learn a strategy $h \in \mathcal{X} \to \mathcal{Y}$ which minimizes the probability of misclassification. To this end we use a learning algorithm $A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ which returns a strategy h from the class $\mathcal{H} = \{h(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \mid \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$ containing all linear classifers.

a) What is the approximation error in case that $\boldsymbol{C}_+ = \boldsymbol{C}_-$?

b) Is the approximation error going to increase or decrease if $\boldsymbol{C}_+ \neq \boldsymbol{C}_-$?

c) Give example(s) of distribution $p(x, y)$ such that the approximation error is zero when using the class $\mathcal{H}$.

Assignment 5. 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 $\varepsilon > 0$,

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

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 $\mathcal{H}$.

Assignment 6. Let $h_m \in \operatorname{Arg\,min}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$ be a predictor learned by ERM on training examples $\mathcal{T}^m$ and let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a finite hypothesis space. Let $h_{\mathcal{H}} \in \operatorname{Arg\,min}_{h \in \mathcal{H}} R(h)$ be the best predictor in $\mathcal{H}$. On the lecture we have derived an upper bound on the probability that the estimation error $R(h_m) - R(h_{\mathcal{H}})$ is equal or above $\varepsilon > 0$, namely that

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \Big) \leq 2|\mathcal{H}| e^{-\frac{m\, \varepsilon^2}{2(\ell_{max} - \ell_{min})^2}} . \tag{1}$$

a) Use the bound (1) to prove that for arbitrary $\varepsilon, \delta \in (0, 1)$, the inequality

$$R(h_m) - R(h_{\mathcal{H}}) \leq \varepsilon$$

holds with a probability $1 - \delta$ at least, provided the number of training examples m is at least

$$\frac{2(\log 2|\mathcal{H}| - \log \delta)}{\varepsilon^2} (\ell_{max} - \ell_{min})^2 .$$

b) Describe how would you use the result derived in a) for model selection?