STATISTICAL MACHINE LEARNING (WS2021)
SEMINAR 3

Assignment 1. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, $R(h)$ the true risk and let $h_{\mathcal{H}} \in \operatorname{Arg\,min}_{h \in \mathcal{H}} R(h)$ be the best predictor in the class $\mathcal{H}$. Assume that for $\mathcal{H}$ we have the uniform generalization bound

$$\mathbb{P}(\sup_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) \leq B(m, \mathcal{H}, \varepsilon) \ ,$$

where $B(m, \mathcal{H}, \varepsilon)$ depends on the number of training examples m, the hypothesis class $\mathcal{H}$ and the precision parameter $\varepsilon > 0$. For example, in the case of a finite hypothesis space, we have $B(m, \mathcal{H}, \varepsilon) = 2|\mathcal{H}| \exp(-\frac{2m\,\varepsilon^2}{(b-a)^2})$. Let h_m be a prediction strategy learned from the training examples $\mathcal{T}^m$ by the ERM algorithm

$$h_m \in \operatorname*{Arg\,min}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h) \ .$$

Show that in this case

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

holds with the probability $1 - B(m, \mathcal{H}, \varepsilon/2)$ at least.

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

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

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