STRUCTURED MODEL LEARNING (SML2019)
SEMINAR 3.

Assignment 1. Let $\mathcal{X}$ be a set of input observations and $\mathcal{Y} = \mathcal{A}^n$ a set of sequences of length n defined over a finite alphabet $\mathcal{A}$. Let $h\colon \mathcal{X} \to \mathcal{Y}$ be a prediction rule that for each $x \in \mathcal{X}$ returns a sequence $h(x) = (h_1(x), \ldots, h_n(x))$. Assume that we want to measure the prediction accuracy of $h(x)$ by the expected Hamming distance $R(h) = \mathbb{E}_{(x,y_1,\ldots,y_n)\sim p}(\sum_{i=1}^n [\![h_i(x) \neq y_i]\!])$ where $p(x, y_1, \ldots, y_n)$ is a p.d.f. defined over $\mathcal{X} \times \mathcal{Y}$. As the distribution $p(x, y_1, \ldots, y_n)$ is unknown we estimate $R(h)$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{j=1}^{l} \sum_{i=1}^{n} [\![y_i^j \neq h_i(x^j)]\!]$$

where $\mathcal{S}^l = \{(x^i, y_1^i, \ldots, y_n^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x, y_1, \ldots, y_n)$. What is the minimal number of the test examples l which we need to collect in order to guarantee that $R(h)$ is in the interval $[R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon]$ with probability $1 - \delta$ at least where $\delta \in (0, 1)$? Write l as a function of ε, n and δ.

Hint: Use the Hoeffding's inequality

$$\mathbb{P}_{\mathcal{S}^l \sim p^l}\left(\Big| R(h) - R_{\mathcal{S}^l}(h) \Big| \geq \varepsilon \right) \leq 2 \exp\left(\frac{-2l\,\varepsilon^2}{(\ell_{\max} - \ell_{\max})^2} \right) \tag{1}$$

Assignment 2. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{Y}}$ be a finite hypothesis space, $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [\ell_{min}, \ell_{max}]$ a loss function, $R(h) = \mathbb{E}_{(x,y)\sim p}\big(\ell(y, h(x))\big)$ the expected risk of a hypothesis $h \in \mathcal{H}$, $R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^m \ell(y^i, h(x^i))$ the empirical risk of $h \in \mathcal{H}$ computed from examples $\mathcal{T}^m = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m\}$ drawn i.i.d. from $p(x, y)$. Prove that

$$\mathbb{P}_{\mathcal{T}^m \sim p^m}\left(\max_{h \in \mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \geq \varepsilon \right) \leq 2|\mathcal{H}| \exp\left(\frac{-2m\,\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2} \right)$$

holds for any $\varepsilon > 0$.

Hint:

- Start from the Hoeffding's inequality (1) which claims the same for the case when $\mathcal{H}$ contains just a single hypothesis.
- Note that for a sequence of random variables $A_1, \ldots, A_n$ it holds

  $$\mathbb{P}\big(\max_{i=1,\ldots,n} A_i \geq \varepsilon \big) = \mathbb{P}\Big((A_1 \geq \varepsilon) \vee (A_2 \geq \varepsilon) \vee \cdots \vee (A_n \geq \varepsilon) \Big)$$

- Exploit the identity $\mathbb{P}(A \vee B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \wedge B)$

Assignment 3. Assume we want to train a convolutinal neural network $h\colon \mathcal{X} \to \mathcal{Y}$ which minimizes the probability of classification error when predicting a label $y \in \mathcal{Y}$ from an image $x \in \mathcal{X}$. Assume $\mathcal{H} = \{h_t\colon \mathcal{X} \to \mathcal{Y} \mid t = 1,\ldots,T\}$ are CNNs obtained after $1, 2, \ldots, T$ training epochs when one epoch corresponds to running SGD though the entire training set. The final CNN h^* is selected out of $\mathcal{H}$ by minimizing the validation error

$$R_{\text{val}}(h) = \frac{1}{v}\sum_{i=1}^{v} [\![h(x^i) \neq y^i]\!]$$

where $\{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1,\ldots,v\}$ are i.i.d. drawn validation examples. What is the minimal number of validation examples v which guarantees that the expected classification error is within the interval $(R_{\text{val}}(h^*) - 0.01, R_{\text{val}}(h^*) + 0.01)$ with probability 95% at least ? Write the number of examples as a function of T and evaluate it for $T = 100$.

Assignment 4. Let $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$ be a set of functions $g\colon \mathcal{Z} \to [a,b]$ where $a, b \in \mathbb{R}$ and $a < b$. The *empirical Rademacher complexity* of $\mathcal{G}$ w.r.t. to the sample $\mathcal{U}^m = \{z^1,\ldots,z^m\} \in \mathcal{Z}^m$ is

$$\hat{\mathfrak{R}}_m(\mathcal{G},\mathcal{U}^m) = \mathbb{E}_{\sigma\sim\text{Unif}\{-1,+1\}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\, g(z_i)\right].$$

Prove that $\hat{\mathfrak{R}}_m(\mathcal{G},\mathcal{U}^m)$ is always non-negative and that it is 0 if $\mathcal{G}$ contains just a single function.

Assignment 5. Assume a class of binary classifiers $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$. Let $\mathcal{Z} = \mathcal{X} \times \{+1,-1\}$ and $\mathcal{G} = \{[\![h(x) \neq y]\!] \mid h \in \mathcal{H}\}$ be a class of functions $g(z) = [\![y \neq h(x)]\!]$, i.e. composition of the 0/1-loss $[\![y \neq y']\!]$ and the hypothesis $h \in \mathcal{H}$. Let $\mathcal{U}^m = \{z^i \in \mathcal{Z} \mid i = 1,\ldots,m\} = \{(x^i, y^i) \in \mathcal{X} \times \{+1,-1\} \mid i = 1,\ldots,m\}$ be a sample of points from $\mathcal{X} \times \{-1,+1\}$ and $\mathcal{V}^m = \{x^i \in \mathcal{X} \mid i = 1,\ldots,m\}$ be a projection of $\mathcal{U}^m$ on the domain $\mathcal{X}$. The *empirical Rademacher complexity* of $\mathcal{G}$ w.r.t. to the sample $\mathcal{U}^m$ is

$$\hat{\mathfrak{R}}_m(\mathcal{G},\mathcal{U}^m) = \mathbb{E}_{\sigma\sim\text{Unif}\{-1,+1\}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\, g(z_i)\right].$$

Similarly, the *empirical Rademacher complexity* of $\mathcal{H}$ w.r.t. to the sample $\mathcal{V}^m$ is

$$\hat{\mathfrak{R}}_m(\mathcal{H},\mathcal{V}^m) = \mathbb{E}_{\sigma\sim\text{Unif}\{-1,+1\}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\, h(x_i)\right].$$

Prove that $\hat{\mathfrak{R}}_m(\mathcal{G},\mathcal{U}^m) = \frac{1}{2}\hat{\mathfrak{R}}_m(\mathcal{H},\mathcal{V}^m)$.