STURCTURED MODEL LEARNING (WS2021/22)
SEMINAR 2

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 $p(x, y_1, \ldots, y_n)$.

a) Assume that the sequence length is $n = 10$ and that we compute the test error from $l = 1000$ examples. Use the Hoeffding inequality to bound the probability that $R(h)$ will be in the interval $(R_{\mathcal{S}^l}(h) - 1, R_{\mathcal{S}^l}(h) + 1)$?

b) 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 δ at least? Write l as a function of ε, n and δ.

Hint: Apply the Hoeffding inequality on slide 4 of lecture 2.

Assignment 2. Assume we are training a Convolution Neural Network (CNN) based classifier $h\colon \mathcal{X} \to \mathcal{Y}$ to predcit a digit $y \in \mathcal{Y} = \{0, 1, \ldots, 9\}$ from an image $x \in \mathcal{X}$. We train the CNN by the Stchastic Gradient Descent (SGD) algorithm using 100 epochs. After each epoch we save the current weights so that at the end of training we have a set $\mathcal{H} = \{h_t\colon \mathcal{X} \to \mathcal{Y} \mid i = 1, \ldots, 100\}$ containing 100 CNN classifiers. The goal is to select the best CNN out of $\mathcal{H}$ that has the minimal classification error

$$R(h) = \mathbb{E}_{(x,y)\sim p}([\![y \neq h(x)]\!]) ,$$

where the expectation is w.r.t. an unknown distribution $p(x, y)$ generating the data. Because $p(x, y)$ is unknown, we approximate $R(h)$ by the empirical risk

$$R_{\mathcal{V}^m}(h) = \frac{1}{m}\sum_{i=1}^{m}[\![y^j \neq h(x^j)]\!] ,$$

computed from a validation set $\mathcal{V}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ containing m examples i.i.d. drawn from $p(x, y)$.

a) Define a method based on the Empirical Risk Minimization which uses $\mathcal{V}^m$ to select the best CNN out of a finite hypothesis class $\mathcal{H}$.

b) What is the minimal number of validation examples m we need to collect in order to have a guarantee that $R(h)$ is in the interval $(R_{\mathcal{V}^m}(h) - 0.01, R_{\mathcal{V}^m}(h) + 0.01)$ for every $h \in \mathcal{H}$ with probability at least 95%?

Hint: Apply the uniform generalization bound for finite hypothsis space from slide 14 of lecture 2.

Assignment 3. 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 the estimation error is at most ε, i.e.

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

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

Hint: Use the inequality at the very bottom of slide 13 of lecture 2.