STATISTICAL MACHINE 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.

Assignment 2. Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1) = p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$. Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns the strategy

$$h_m(x) = \begin{cases} y^j & \text{if } x = x^j \text{ for some } j \in \{1, \ldots, m\} \\ -1 & \text{otherwise} \end{cases}$$

a) Show that the empirical risk $R_{\mathcal{T}^m}(h_m) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h_m(x^i))$ equals 0 with probability 1 for any finite m.

b) Show that the expected risk $R(h_m) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h_m(x))$ equals 0.8 for any finite m.

Assignment 3. A professor created 1000 tests that he uses for examining students. Every student has an opportunity to come for one consultation, during which the professor gives the student 10 tests for self-study. At the day of exam, the professor randomly chooses one out of the 1000 tests. It may happen that the exam test is the same as one of the 10 tests a student got for self-study. Then, the student is just lucky and passes the

exam for sure. From the professor's viewpoint, this is an exam failure. Otherwise, the student succeeds based on his/her actual knowledge.

a) If a single student comes to the exam, what is the probability of exam failure, that is, the student passes just because of luck?

b) If there are N students coming to the exam, what is the probability that there will be an exam failure, that is, at least one student passes because of luck? Derive an upper bound on the probability of the exam failure. Hint: use the union bound (also called Boole's inequality).

Assignment 4. 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%?