STATISTICAL MACHINE LEARNING (WS2022/23)
SEMINAR 3

Assignment 1. 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 even without any knowledge. From the professor's viewpoint, this is an exam failure.

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 of the 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.

Assignment 2. Assume we are training a Convolution Neural Network (CNN) based classifier $h\colon \mathcal{X} \to \mathcal{Y}$ to predict a digit $y \in \mathcal{Y} = \{0, 1, \ldots, 9\}$ from an image $x \in \mathcal{X}$. We train the CNN by the Stochastic 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 different CNN classifiers. The goal is to select the best CNN out of $\mathcal{H}$ that has the minimal classification error

$$R(h) = \mathbb{E}_{(x,y)\sim p}(\llbracket y \neq h(x) \rrbracket) \,.$$

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

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

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 $\mathcal{H}$.

b) Let $\hat{h} \in \mathcal{H}$ be a CNN classifier selected in assignment a). Let us say we want to use the same validation set $\mathcal{V}^m$ not only to select $\hat{h}$ but also to estimate the true error $R(\hat{h})$. Compute the tolerance ε such that

$$R(\hat{h}) \leq R_{\mathcal{V}^m}(\hat{h}) + \varepsilon$$

holds with probability $\gamma = 0.95$ at least.

c) How many examples m we need to have in the validation set $\mathcal{V}^m$ to a guarantee that

$$R(\hat{h}) \leq R_{\mathcal{V}^m}(\hat{h}) + \varepsilon$$

holds with probability $\gamma \in (0,1)$ at least where $\varepsilon > 0$ is the maximal allowed tolerance of the estimate chosen by us. Solve the problem for general γ, ε and the size of hypothesis space $|\mathcal{H}|$. Then, compute the number of examples m for the tolerance $\varepsilon = 0.01$, confidence level $\gamma = 0.95$ and $|\mathcal{H}| = 100$.

Assignment 3. Assume two-class classification problem, $\mathcal{Y} = \{-1,+1\}$, with 0/1-loss $\ell(y,y') = [\![y \neq y']\!]$, when the input space $\mathcal{X} = [0,1]^n$ is n-dimensional hypercube. Let us partition $\mathcal{X}$ into K equally sized bins $\mathcal{X}_j$, $j \in \{1,\ldots,K\}$, each bin being itself a hypercube, such that

$$\mathcal{X} = \cup_{k=1}^{K} \mathcal{X}_k \quad \text{and} \quad \mathcal{X}_k \cap \mathcal{X}_j = \emptyset\,, \forall k \neq j\,,$$

where $K = D^n$, $D \in \{1,2,3,\ldots,\}$. A histogram classifier $h\colon \mathcal{X} \to \{-1,+1\}$, parametrized by a vector $\boldsymbol{w} \in \{-1,+1\}^K$, assigns input x into class w_k where k identifies hypercube $\mathcal{X}_k$ such that $\boldsymbol{x} \in \mathcal{X}_k$, that is,

$$h(\boldsymbol{x}) = \sum_{k=1}^{K} w_k [\![\boldsymbol{x} \in \mathcal{X}_k]\!]\,. \tag{1}$$

Let

$$\mathcal{H}_K = \left\{ h\colon \mathcal{X} \to \{-1,+1\} \mid h(\boldsymbol{x}) = \sum_{k=1}^{K} w_k [\![\boldsymbol{x} \in \mathcal{X}_k]\!]\,, \boldsymbol{w} \in \{-1,+1\}^K \right\}$$

denote a hypothesis space composed of all histogram classifiers that partition the input space into K hypercubes. Assume we have a training set $\mathcal{T}^m = \{(\boldsymbol{x}^1,y^1),\ldots,(\boldsymbol{x}^m,y^m)\} \in (\mathcal{X} \times \{+1,-1\})^m$ drawn i.i.d. from some unknown $p(\boldsymbol{x},y)$.

a) How does the number of strategies $|\mathcal{H}_K|$ depend on the number of bins K and on D?

b) Assume the number of bins K is fixed. The ERM based algorithm transforms learning of the histogram classifier into optimization problem

$$h_K \in \operatorname*{Arg\,min}_{h \in \mathcal{H}_K} R_{\mathcal{T}^m}(h) \quad \text{where} \quad R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{j=1}^{m} \ell(y^j, h(x^j))\,. \tag{2}$$

Design computationaly tractable algorithm which solves the problem (2).

c) Design an algorithm based on the structural risk minimization principle to learn the weights $\boldsymbol{w}$ and the number of bins K simultaneously using a single training set $\mathcal{T}^m$.