STATISTICAL MACHINE LEARNING (WS2023/24)
SEMINAR 1

Assignment 1. Assume a prediction problem with a scalar observation $\mathcal{X} = \mathbb{R}$, two classes $\mathcal{Y} = \{-1, +1\}$ and 0/1-loss $\ell(y, y') = [\![y \neq y']\!]$[1]. The observations of both classes are generated from normal distributions, i.e.

$$p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{1}{2\sigma^2}(x - \mu_y)^2 \Big) , \qquad y \in \mathcal{Y} ,$$

where $p(y)$ is the prior distribution of the hidden state, $\sigma_+, \sigma_- \in \mathbb{R}_+$ are the standard deviations and $\mu_+, \mu_- \in \mathbb{R}$ are the mean values.

a) Assume $\mu_- < \mu_+$ and $\sigma_+ = \sigma_-$. Show that under this assumption the optimal prediction strategy is the thresholding rule

$$h(x) = \begin{cases} -1 & \text{if} \quad x < \theta , \\ +1 & \text{if} \quad x \geq \theta , \end{cases}$$

parametrized by the scalar $\theta \in \mathbb{R}$. Write an explicit formula for computing θ.

b) Deduce the optimal prediction strategy for the case $\mu_+ = \mu_-$ and $\sigma_+ \neq \sigma_-$.

Assignment 2. Let $\mathcal{S}^l = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l)$ be a test set i.i.d drawn from some $p(x, y)$ and let $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function. The test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$ is an unbiased estimator of the generalization error $R(h) = \mathbb{E}_{(x,y)\sim p} \ell(y, h(x))$.

a) What does it mean that the test error is an unbiased estimator of the generalization error?

b) Prove that it holds true.

(*) Can you deduce something about the variance of the test error?

Assignment 3. We are given a prediction strategy $h \colon \mathcal{X} \to \mathcal{Y} = \{1, \ldots, Y\}$ assigning observations $x \in \mathcal{X}$ into one of Y classes. Our task is to estimate the generalization error $R(h) = \mathbb{E}_{(x,y)\sim p} \ell(y, h(x))$ where $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a chosen loss function. To this end, we collect a test set $\mathcal{S}^l = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l)$ i.i.d. drawn from the distribution $p(x, y)$, compute the test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$ and use it to construct the confidence interval such that

$$R(h) \in (R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon) \quad \text{holds with probability } 1 - \delta \in (0, 1) \text{ at least} . \qquad (1)$$

The number of test examples $l \in \mathbb{N}$, the precision parameter $\varepsilon > 0$ and the error level $\delta \in (0, 1)$ are three interdependent variables, i.e., fixing two of the variables allows to compute the third one.

a) Use the Hoeffding's inequality to derive a formula to compute ε as a function of l and δ such that (1) holds.

[1] $[\![e]\!]$ denotes the Iverson bracket with value 1 if the expression in the brackets is true and 0 otherwise.

b) Use the Hoeffding's inequality to derive a formula to compute l as a function of ε and δ such that (1) holds.

c) Instantiate the formulas derived in a) and b) for the following loss functions:
(1) $\ell(y, y') = [\![y \neq y']\!]$
(2) $\ell(y, y') = |y - y'|$
(3) $\ell(y, y') = [\![|y - y'| \geq K]\!]$ where $K < Y$.

d) Assume that we use the loss $\ell(y, y') = [\![y \neq y']\!]$. Plot the precision ε as a function of the number of examples $l \in \{10, 100, \ldots, 100000\}$ for $\delta \in \{0.1, 0.05, 0.01\}$.

e) Assume that we use the loss $\ell(y, y') = [\![y \neq y']\!]$. What is the minimal number of examples l we need to use to have a guarantee that the test error will approximate the generalization error $\pm 1\%$ with probability 95% at least?