STATISTICAL MACHINE LEARNING (WS2020)
SEMINAR 2

Assignment 1. Assume a prediction problem with a scalar observation $\mathcal{X} = \mathbb{R}$, two classes $\mathcal{Y} = \{-1, +1\}$ and 0/1-loss $\ell(y, y') = [\![y \neq y']\!]$. The observations of both classes are generated according to the Normal distribution, 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) Show what is the optimal prediction strategy in case when $\mu_+ = \mu_-$ and $\sigma_+ \neq \sigma_-$.

Assignment 2. Consider a prediction problem $h\colon \mathcal{X} \to \mathcal{Y}$ where observation $x \in \mathcal{X}$ and hidden state $y \in \mathcal{Y} \subseteq \mathbb{R}$ are realizations of random variables distributed according to a known joint distribution $p(x, y)$. Deduce the optimal inference rule minimizing the expected risk assuming that the loss function is:

a) quadratic $\ell(y, y') = |y - y'|^2$.

b) absolute deviation $\ell(y, y') = |y - y'|$.

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 expected risk $R^\ell(h) = \mathbb{E}_{(x,y)\sim p}\ell(y, h(x))$ where $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some application specific loss function. To this end, we collect a set of examples $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ drawn i.i.d. from the distribution $p(x, y)$ and compute the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l}\ell(y^i, h(x^i))\,.$$

What is the minimal number of test examples l we need to collect in order to have a guarantee that the expected risk $R^\ell(h)$ is inside the interval $(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon)$ with probability $\gamma \in (0, 1)$ for some predefined $\varepsilon > 0$?

a) Use Hoeffding's inequality to derive a formula to compute l as a function of ε and γ.

b) Assume the loss defined as $\ell(y, y') = [\![|y - y'| > 5]\!]$. Evaluate l for $\varepsilon = 0.01$ and $\gamma \in \{0.90, 0.95, 0.99\}$. Give an interpretation of the expectation of the loss.

c) Solve the problem b) in case that the loss is the mean absolute error, $\ell(y, y') = |y - y'|$. Evaluate l for $\varepsilon = 1$, $Y = 100$ and $\gamma \in \{0.90, 0.95, 0.99\}$.

d) How do the formulas depend on the particular loss function?

Remark: $[\![A]\!]$ *is the Iverson bracket which evaluates to 1 if the logical statement* A *is true and to* 0 *otherwise.*

Assignment 4. 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.