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

$$p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{1}{2\sigma^2} (x - \mu_y)^2 \right), \quad y \in Y,$$

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

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

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

b) Show what is the optimal prediction strategy in case when $\mu_+ = \mu_- \text{ and } \sigma_+ \neq \sigma_-.$

Assignment 2. Consider the following probabilistic model for real valued sequences $x = (x_1, \ldots, x_n), x_i \in \mathbb{R}$ of fixed length $n$. Each sequence is a combination of a leading part $i \leq k$ and a trailing part $i > k$. The boundary $k = 0, \ldots, n$ is random with uniform distribution. The values $x_i$, in the leading and trailing part are statistically independent and distributed with some probability density function $p_1(x)$ and $p_2(x)$ respectively. Altogether the distribution for pairs $(x, k)$ reads

$$p(x, k) = \frac{1}{n+1} \prod_{i=1}^{k} p_1(x_i) \prod_{j=k+1}^{n} p_2(x_j).$$

The densities $p_1$ and $p_2$ are known. Given a sequence $x$, we want to predict the boundary $k$.

a) Deduce the optimal predictor for the 0/1 loss, i.e $\ell(k, k') = \mathbb{I}[k \neq k']$.

b) Deduce the optimal predictor for the quadratic loss $\ell(k, k') = (k - k')^2$.

Assignment 3. We are given a prediction strategy $h: \mathcal{X} \rightarrow \mathcal{Y} = \{1, \ldots, Y\}$ assigning observations $x \in \mathcal{X}$ into one of $Y$ classes. Our task is to estimate the true risk $R(h) = \mathbb{E}_{(x, y) \sim \mathcal{P}} \ell(y, h(x))$ where $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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 true risk $R(h)$ 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 $\varepsilon$ and $\gamma$.

b) Assume the loss defined as $\ell(y, y') = \llbracket |y - y'| > 5 \rrbracket$. 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?

Assignment 4. Let us consider the family of linear classifiers $h \in \mathcal{H}$ defined by

$$y = h(x; w, b) = \text{sign}(x^T w - b),$$

where $x \in \mathbb{R}^n$ denotes a feature vector and $y = \pm 1$ denotes the binary class. The predictors are parametrised by the vector $w \in \mathbb{R}^n$ and the scalar $b \in \mathbb{R}$. Given training data $T = \{(x_i, y_i) | i = 1, 2, \ldots, m\}$, we want to find the predictor that minimises the empirical risk on the training data, i.e.

$$\mathbb{R}_T(h) = \frac{1}{|T|} \sum_{(x,y) \in T} \ell(y, h(x)) \rightarrow \min_{h \in \mathcal{H}},$$

for the 0/1 loss $\ell(y, y') = \llbracket y \neq y' \rrbracket$.

a) Consider the loss for a single example $(x, y) \in T$ as a function of the classifier parameters, i.e. $f(w, b) = \ell(y, h(x; w, b))$. What type of function is it? Can we minimise it by gradient descent? Conclude that the empirical risk $\mathbb{R}_T(h)$ can not be minimised by gradient descent w.r.t. $w$ and $b$.

b) Suppose, we know that there is a classifier $h^* \in \mathcal{H}$, with zero empirical risk on the training data. Give an algorithm that finds such a predictor.

c) Suppose now, no such predictor exists. How can we resolve the problem we encountered in a)?

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