STATISTICAL MACHINE LEARNING (WS2022/23)
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 us consider the family of linear classifiers $h \in \mathcal{H}$ defined by

$$y = h(x; w, b) = \operatorname{sign}(w^T x - b), \tag{1}$$

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 $\mathcal{T} = \{(x_i, y_i) \mid i = 1, 2, \ldots, m\}$, we want to find the predictor that minimises the empirical risk on the training data, i.e.

$$R_{\mathcal{T}}(h) = \frac{1}{|\mathcal{T}|}\sum_{(x,y)\in\mathcal{T}} \ell(y, h(x)) \to \min_{h\in\mathcal{H}},$$

for the 0/1 loss $\ell(y, y') = [\![y \neq y']\!]$.

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

b) Suppose, we know that $\mathcal{H}$ contains 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] $[\![e]\!]$ denotes the Iverson bracket with value 1 if the expression in the brackets is true and 0 otherwise.

Assignment 3. Consider a prediction problem $h\colon \mathcal{X} \to \mathcal{Y}$ where observations $x \in \mathcal{X}$ and hidden states $y \in \mathcal{Y} \subseteq \mathbb{R}$ are realizations of random variables distributed according to a known joint distribution $p(x, y)$. Deduce the optimal inference rule that minimises the expected risk assuming that the loss function is:

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

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