STATISTICAL MACHINE LEARNING (WS2017)
SEMINAR 2

Assignment 1. Consider a prediction problem when the set of input observations is $\mathcal{X} = \mathbb{R}$, the set of hidden states is $\mathcal{Y} = \{+1, -1\}$, the loss function is $\ell(y, y') = 0$ if $y = y'$ and $\ell(y, y') = 1$ if $y \neq y$, and the joint distribution reads

$$p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_y)^2}{2\sigma^2}}$$

where $p(y)$, $y \in \mathcal{Y}$, are the prior probabilities and $\mu_{+1} \in \mathbb{R}\,, \mu_{-1} \in \mathbb{R}\,, \sigma \in \mathbb{R}_{++}$, are some parameters such that $\mu_{+1} > \mu_{-1}$.

a) Show that optimal prediction rule minimizing the expected risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x))$ is of the form

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

where θ is a constant. Derive an explicit formula to compute θ from the parameters.

b) Show that the expected risk of the thresholding rule (1) reads

$$R(h) = \int_{-\infty}^{\theta} p(x, +1) \mathrm{d}x + \int_{\theta}^{\infty} p(x, -1) \mathrm{d}x \;.$$

Assignment 2. Consider the task of age estimation based on visual cues. Let us denote the visual features by $x \in \mathcal{X}$ and the unknown age by $y \in \mathbb{N}$. The statistical relation between the two random variables is known and given by their joint distribution $p(x, y)$.

a) Deduce the optimal prediction rule for the loss function $\ell(y, y') = |y - y'|^2$.

b) Same for the loss function $\ell(y, y') = |y - y'|$.

Assignment 3. We are given a prediction rule $h\colon \mathcal{X} \to \{-1, +1\}$. The task is to estimate the probability of misclassification $R(h) = \mathbb{E}_{(x,y)\sim p}([\![y \neq h(x)]\!])$ by computing the (empirical) test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^{l} [\![y^j \neq h(x^j)]\!]$$

where $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x, y)$.

What is the minimal number of test examples l we need to collect in order to have a guarantee that the probability of misclassification $R(h)$ is in the interval $(R_{\mathcal{S}^l}(h) - 0.01, R_{\mathcal{S}^l}(h) + 0.01)$ with probability $90\%, 95\%$ and 99% ?

Assignment 4. Let $\mathcal{H}$ be a hypothesis space containing all linear prediction rules assigning input observation $x \in \mathcal{X}$ to two hidden states $y \in \mathcal{Y} = \{+1, -1\}$ so that

$$h(x) = \begin{cases} +1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b \geq 0\,, \\ -1 & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b < 0\,, \end{cases}$$

where $\boldsymbol{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$, are parameters and $\boldsymbol{\phi} \colon \mathcal{X} \to \mathbb{R}^n$ is a map which returns a n-dimensional feature vector for each $x \in \mathcal{X}$.

Let $R(h) = \mathbb{E}_{(x,y)\sim p}([\![y \neq h(x)]\!])$ denote the expectation of $0/1$-loss function w.r.t some distribution $p(x, y)$, let $R^* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R(h)$ be the best attainable risk and $h_\mathcal{H} \in \operatorname{Arg\,min}_{h\in\mathcal{H}} R(h)$ the best (if not unique then one of the best) predictor in $\mathcal{H}$.

Try to find examples of triplets $\mathcal{X}$, $p(x, y)$ and $\boldsymbol{\phi}(x)$ such that using the hypothesis space containing all linear predictors implies zero approximation error $R(h_\mathcal{H}) - R^*$.