STATISTICAL MACHINE LEARNING (WS2024/25)
HOMEWORK: HISTOGRAM CLASSIFIER

Assignment (4 points) *Assume a classification problem, $\mathcal{Y} = \{1, 2, \ldots, Y\}$, with a loss $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, when the input space $\mathcal{X} = [x_{\min}, x_{\max}] \subset \mathbb{R}$ is an interval between $x_{\min}$ and $x_{\max}$. Let the input space $\mathcal{X}$ be partitioned into K equally sized intervals $\mathcal{X}_j$, $j \in \{1, 2, \ldots, K\}$, such that*

$$\mathcal{X} = \cup_{k=1}^{K} \mathcal{X}_k \quad \textit{and} \quad \mathcal{X}_k \cap \mathcal{X}_j = \emptyset, \forall j \neq k \, .$$

A histogram classifier $h \colon \mathcal{X} \to \mathcal{Y}$, parameterized by a vector $\boldsymbol{w} = (w_1, \ldots, w_K) \in \mathcal{Y}^K$, assigns an input x into class w_k where k represents the interval $\mathcal{X}_k$ containing the input, $x \in \mathcal{X}_k$. That is, the histogram classifier reads

$$h(x) = \sum_{k=1}^{K} w_k [\![x \in \mathcal{X}_k]\!] \, . \tag{1}$$

Let $\mathcal{H}_K$ be the set of all histogram classifiers (1) with K bins. Assume we have a training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ drawn i.i.d. from some unknown $p(x, y)$.

a) *Implement the ERM algorithm which learns the histogram classifier with K bins by solving:*

$$h_m \in \operatorname*{Arg\,min}_{h \in \mathcal{H}_K} R_{\mathcal{T}^m}(h) \quad \textit{where} \quad R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{j=1}^{m} \ell(y^j, h(x^j)) \, . \tag{2}$$

b) *For a given error level $\delta \in (0, 1)$, compute an upper bound on the expected $R(h_m) = \mathbb{E}_{(x,y)\sim p} \ell(y, h_m(x))$ which holds with probability $1 - \delta$ at least.*

c) *Let $h_{\mathcal{H}} \in \arg\min_{h \in \mathcal{H}_K} R(h)$ be the best histogram classifier in $\mathcal{H}_K$. For a given error level $\delta \in (0, 1)$, compute an upper bound on the estimation error $R(h_m) - R(h_{\mathcal{H}})$ which holds with probability $1 - \delta$ at least.*