Dr. Ulrich Terstiege,
Dr. Tizian Wenzel

Summer semester 2025

Mathematical and Statistical Foundations of Machine Learning
Exercise sheet 1

Submission deadline on Wednesday 07.05.2025 at 9 am via moodle

Notation

For $p, q \in \mathbb{Z}$ such that $p \leq q$, we denote: $[\![p, q]\!] = \{p, p+1, \ldots, q-1, q\}$

Exercise 1 (Union bound). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, i.e., Ω a set, $\mathcal{F} \subset \mathcal{P}(\Omega)$ a σ-algebra and $\mathbb{P}$ a probability measure on $(\Omega, \mathcal{F})$. Show that for any sequence $(A_i)_{i\in\mathbb{N}}$ of events, the following holds:*

$$\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i).$$

Exercise 2 (Bernoulli random variable). *Let $X \in \{0, 1\}$ be a Bernoulli random variable with parameter $p \in [0, 1]$, i.e., $\mathbb{P}(X = 1) = p$. Calculate the expectation and variance of X.*

Exercise 3 (Binomial random variable). *Let $Y \in \{0, ..., N\}$ be a Binomial random variable with parameters $N \geq 2$ and $p \in [0, 1]$, i.e.:*

$$\mathbb{P}(Y = k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \text{for all } \; k = 0, ..., N.$$

Calculate the expectation and variance of Y.

Exercise 4 (Useful inequality).
Show that for every $x \in \mathbb{R}$:

$$1 - x \leq \exp(-x).$$

Exercise 5 (Axis-aligned rectangles). *An axis-aligned rectangle classifier in $\mathbb{R}^d$ is a classifier that assigns the value 1 to a point if and only if it is inside a given rectangle. Formally, given real numbers $a_i \leq b_i$, for $i \in [\![1, d]\!]$, we define an axis-aligned rectangle:*

$$R = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$$

and an associated classifier h_R by

$$h_R(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} 1_{[a_i, b_i]}(x_i) = \begin{cases} 1 & \text{if } \forall i \in [\![1, d]\!]: \quad a_i \leq x_i \leq b_i \\ 0 & \text{otherwise.} \end{cases}$$

The class of all axis-aligned rectangles in $\mathbb{R}^d$, denoted $\mathcal{H}^d_{rec}$ is defined as the collection of all classifiers associated to axis-aligned rectangles. i.e.: $\mathcal{H}^d_{rec} = \{h_{[a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_d,b_d]} : a_i \leq b_i, i \in [\![1, d]\!]\}$. Note that this is an infinite size hypothesis class.

The aim of this detailed problem is to show that the concept class $\mathcal{H}^d_{rec}$ is PAC-learnable (the definition of PAC-learnability is not needed to solve this problem).

1) *Given an unknown axis-aligned rectangle R and labelled training data $S = \{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathbb{R}^d$, and $y_i = h_R(x_i) \in \{0, 1\}$, let $\mathcal{A}_S^d$ be the algorithm that returns the smallest rectangle R_S enclosing all positive examples in the training set. Show that h_{R_S} is an Empirical Risk Minimiser.*

2) *In this question, we treat the simple case $d = 1$. Let R be an unknown compact interval $[a, b]$. Let $\varepsilon > 0$ and $\delta > 0$.*

   (a) *Let $S = \{(x_i, y_i)\}_{i=1}^m$ denote a sample generated according to a probability distribution $\mathcal{D}$ on $\mathbb{R}$ and let R_S be the compact interval returned by the algorithm $\mathcal{A}_S^1$ (i.e. the smallest compact interval enclosing all positive examples in S). We denote $\mathcal{R}(h_{R_S})$ the generalised error of the hypothesis h_{R_S}. Given a random variable X with distribution $\mathcal{D}$, show that $\mathcal{R}(h_{R_S}) = \mathbb{P}(X \in R - R_S)$.*

   (b) *Show that if $\mathbb{P}(X \in R) \leq \varepsilon$, then $\mathbb{P}_{S\sim\mathcal{D}^m}(\mathcal{R}(h_{R_S}) > \varepsilon) = 0$.*

   (c) *We assume now that $\mathbb{P}(X \in R) > \varepsilon$. We define:*

   $$s = \inf\{u \in [a, b]; \mathbb{P}(X \in [a, u]) \geq \varepsilon/2\} \text{ and } t = \sup\{u \in [a, b]; \mathbb{P}(X \in [u, b]) \geq \varepsilon/2\}$$

   and the intervals $r_1 = [a, s]$, $r_1^* = [a, s)$, $r_2 = [t, b]$ and $r_2^* = (t, b]$. Verify that:

   $$\forall i \in \{1, 2\}: \quad \mathbb{P}(X \in r_i^*) \leq \varepsilon/2 \quad \text{and} \quad \mathbb{P}(X \in r_i) \geq \varepsilon/2$$

   (d) *Show that if R_S intersects both r_1 and r_2 then*

   $$\mathcal{R}(h_{R_S}) = \mathbb{P}(X \in R - R_S) \leq \varepsilon.$$

   (e) *Justify that, for $i \in \{1, 2\}$, $\mathbb{P}_{S\sim\mathcal{D}^m}(R_S \cap r_i = \varnothing) \leq (1 - \varepsilon/2)^m$.*

   (f) *Using Exercise 4, deduce that $\mathbb{P}_{S\sim\mathcal{D}^m}(\mathcal{R}(h_{R_S}) > \varepsilon) \leq 2\exp(-\frac{\varepsilon m}{2})$.*

   (g) *Conclude that if $m \geq m_1(\varepsilon, \delta) = \frac{2}{\varepsilon}\log(\frac{2}{\delta})$, then $\mathbb{P}_{S\sim\mathcal{D}^m}(\mathcal{R}(h_{R_S}) > \varepsilon) \leq \delta$. In other words, if $\mathcal{A}_S^1$ receives a training set of size $m \geq m_1(\varepsilon, \delta)$ then, with probability of at least $1-\delta$ it returns a hypothesis with error of at most ϵ.*

Exercise 6 (Bonus question on axis aligned rectangles)**.** *We now treat the case $d = 2$. Let R be an unknown axis-aligned rectangle $[a_1, b_1] \times [a_2, b_2]$ and $\varepsilon > 0$. As in exercise , let R_S be the axis-aligned rectangle returned by the algorithm $\mathcal{A}_S^2$, for S a labelled training data generated according to a probability distribution $\mathcal{D}$ on $\mathbb{R}^2$. X denotes a random variable with distribution $\mathcal{D}$.*

(a) *Show that if $\mathbb{P}(X \in R) \leq \varepsilon$, then $\mathbb{P}_{S\sim\mathcal{D}^m}(\mathcal{R}(h_{R_S}) > \varepsilon) = 0$.*

(b) *We assume now that $\mathbb{P}(X \in R) > \varepsilon$. We define:*

$$\begin{cases} s_1 = \inf\{u \in [a_1, b_1]; \mathbb{P}(X \in [a_1, u] \times [a_2, b_2]) \geq \varepsilon/4\}; \\ t_1 = \sup\{u \in [a_1, b_1]; \mathbb{P}(X \in [u, b_1] \times [a_2, b_2]) \geq \varepsilon/4\}; \\ s_2 = \inf\{u \in [a_2, b_2]; \mathbb{P}(X \in [a_1, b_1] \times [a_2, u]) \geq \varepsilon/4\}; \\ t_2 = \sup\{u \in [a_2, b_2]; \mathbb{P}(X \in [a_1, b_1] \times [u, b_2]) \geq \varepsilon/4\} \end{cases}$$

and the rectangles:

$$r_1 = [a_1, s_1] \times [a_2, b_2], r_2 = [t, b_1] \times [a_2, b_2], r_3 = [a_1, b_1] \times [a_2, s_2], r_4 = [a_1, b_1] \times [t_2, b_2]$$

Show that if R_S intersects all rectangles $r_i, i \in [\![1, 4]\!]$ then $\mathbb{P}(X \in R - R_S) \leq \varepsilon$.

(c) *Justify that, for $i \in [\![1, 4]\!]$, $\mathbb{P}_{S\sim\mathcal{D}^m}(R_S \cap r_i = \varnothing) \leq (1 - \varepsilon/4)^m$.*

(d) *Prove that $\mathbb{P}_{S\sim\mathcal{D}^m}(\mathcal{R}(h_{R_S}) > \varepsilon) \leq 4\exp(-\frac{\varepsilon m}{4})$.*

(e) *Conclude that if $\mathcal{A}_S^2$ receives a training set of size $m \geq m_2(\varepsilon, \delta) = \frac{4}{\varepsilon}\log(\frac{4}{\delta})$ then, with probability of at least $1 - \delta$ it returns a hypothesis with error of at most ϵ.*