Math of ML : Exercises 1*

September 8, 2025

References

Exercise 1 (Bayes predictor)**.** *Let $\rho \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. What are the Bayes predictors f^* and the Bayes risk $\mathcal{R}^*$ in the following cases? You may make tail assumptions on ρ if needed.*

- *zero-one loss: $\mathcal{Y} = \{0, 1\}$ and $\ell(y, z) = 1_{y \neq z}$*
- *square loss: $\mathcal{Y} = \mathbb{R}$ and $\ell(y, z) = (y - z)^2$*
- *absolute loss: $\mathcal{Y} = \mathbb{R}$ and $\ell(y, z) = |y - z|$ (to avoid the use of "subgradients", you may assume that the law of $Y|X = x$ has a continuous density on $\mathbb{R}$).*

Exercise 2 (Random prediction)**.** *We consider now a random prediction rule where we predict from the probability distribution of y given $x = x'$; and we assume that the loss is the square loss $\ell(y, z) = (y - z)^2$ and $\mathcal{Y} = \mathbb{R}$. When is this achieving the Bayes risk?*

The purpose of the next exercise is to prepare yourself for the lecture on least-squares.

Exercise 3 (Differential calculus)**.** *(i) Let $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$. Compute the first and second derivatives of the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by*

$$f(x) = \frac{1}{2} x^\top A x - b^\top x$$

when A is symmetric and when A is not symmetric.

(ii) Let $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$. Compute the first and second derivatives of the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(w) = \frac{1}{2} \|y - Xw\|^2.$$

Exercise 4 (Relations between in-expectation and PAC bounds)**.** *Let $\mathcal{A}$ be a learning algorithm (i.e. a function $(\mathcal{X} \times \mathcal{Y})^n \to (\mathcal{X} \to \mathcal{Y})$) and let $\mathcal{D} = (x_i, y_i)_{i=1}^n$ denotes the random training set (the sample size n is fixed).*

(i) Assume that $\mathcal{A}$ satisfies the PAC bound

$$\mathbf{P}(\mathcal{R}(\mathcal{A}(\mathcal{D})) - \mathcal{R}^* \leq \epsilon) \geq 1 - \delta(\epsilon).$$

Prove that $\mathcal{A}$ satisfies an "in-expectation" bound.

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

(ii) Suppose $\mathcal{A}$ satisfies an expectation bound

$$\mathbf{E}(\mathcal{R}(\mathcal{A}(\mathcal{D})) - \mathcal{R}^*) \leq \alpha.$$

Using Markov inequality, prove a PAC bound (i.e. of the form in (i)). [Reminder: Markov's inequality states that if X is a nonnegative random variable and $a > 0$, then $aP(X \geq a) \leq E[X]$.] NB: This bound is weak : we usually look for PAC bounds where $\delta(\epsilon)$ decreases exponentially fast.

Exercise 5 (Does there exist best algorithms?)**.** *We consider binary classification $\mathcal{Y} = \{0, 1\}$. We say that a learning algorithm $\mathcal{A}$* is better than *$\mathcal{B}$ with respect to some probability distribution ρ if*

$$\mathcal{R}_\rho(\mathcal{A}(\mathcal{D})) \leq \mathcal{R}_\rho(\mathcal{B}(\mathcal{D}))$$

for all training samples $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^n$.

Prove that for every distribution $\rho \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, there exists a learning algorithm $\mathcal{A}_\rho$ that is better than any other algorithm with respect to ρ.

The previous exercise was a caution against "very strong" statements that can sometimes be found in the literature: they must be scrutinised to see whether they do not hide something obvious.

Let us now recall the statement of the first no free-lunch theorem. The next exercise is a guided proof.

Theorem 1 (No-free-lunch)**.** *Consider the binary classification with* 0-1 *loss, with $\mathcal{X}$ having at least k elements and $\mathcal{Y} = \{0, 1\}$. For any $n \in \mathbb{N}^*$ and learning algorithm $\mathcal{A}$,*

$$\sup_{\rho \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \mathbb{E}\big[\mathcal{R}(\mathcal{A}(\mathcal{D}_n(\rho)))\big] - \mathcal{R}^*_\rho \geq \frac{1}{2}(1 - 1/k)^n.$$

Exercise 6 (Proof of the no-free-lunch)**.** *Without loss of generality, take $\mathcal{X} = \{1, \ldots, k\}$. Let $r \in \{0, 1\}^k$ and let $\rho \in \mathcal{P}(\mathcal{X} \times \{0, 1\})$ be such that $\mathbf{P}(X = j, Y = r_j) = \frac{1}{k}$.*

- *What is the Bayes risk $\mathcal{R}^*_\rho$?*

Now consider the expected risk $S(r) = \mathbf{E}[\mathcal{R}_\rho(\mathcal{A}(\mathcal{D}_n(\rho)))]$, and choose r randomly, such that each coordinate of r is an unbiaised Bernoulli variable.

- *($\star$) Show that*

$$\mathbf{E}_r[S[r]] \geq \frac{1}{2}\big(1 - 1/k\big)^n.$$

Hint: observe that $(1 - 1/k)^n$ is the probability that the random test sample does not coincide with any of the n training sample.

- *Conclude the proof of the no-free lunch theorem.*