Math of ML : Solutions 1 *

September 8, 2025

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

- *zero-one loss: $\mathcal{Y} = \{0, 1\}$ and $\ell(y, z) = 1_{y \neq z}$. You may express the answer in terms of the function $\eta(x) = \mathbf{P}(Y = 1|X = x)$.*
- *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 $\mathbb{R}$).*

Solution 1. *In each case, we minimize, for each value of x, the cond risk $\mathcal{R}(z|x) := \mathbf{E}[\ell(Y, z)|X = x]$.*

- *(0-1 loss) Let $\eta(x) = \mathbf{P}(Y = 1|X = x)$. Here*

$$\mathcal{R}(z|x) = \mathbf{P}(Y \neq z|X = x) = \begin{cases} \eta(x) & \text{if } z = 0 \\ 1 - \eta(x) & \text{if } z = 1. \end{cases}$$

This quantity is minimized by $z = 0$ if $\eta(x) < 1/2$, $z = 1$ if $\eta(x) > 1/2$ and $z \in \{0, 1\}$ if $\eta(x) = 1/2$. Thus Bayes predictors are exactly functions f^* of the form

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{if } \eta(x) < 1/2 \\ u(x) & \text{if } \eta(x) = 1/2 \end{cases}$$

for some arbitrary function $u : \mathcal{X} \to \{0, 1\}$. The corresponding Bayes risk is

$$\mathcal{R}^* = \mathbf{E}[\mathbf{P}(Y \neq f^*(X)|X = x)] = \mathbf{E}[\min\{\eta(X), 1 - \eta(X)\}].$$

since $\mathbf{P}(Y \neq f^*(X)|X = x) = \mathbf{P}(Y = 0|X = x) = 1 - \eta(x)$ if $\eta(x) \geq 1/2$ and $\mathbf{P}(Y \neq f^*(X)|X = x) = \mathbf{P}(Y = 1|X = x) = \eta(x)$ if $\eta(x) \leq 1/2$, which can be written in both cases as $\mathbf{P}(Y \neq f^*(X)|X = x) = \min\{\eta(X), 1 - \eta(X)\}$.

- *Here we need to assume that ρ has finite second moments. We perform a "bias-variance decomposition" : for any $z \in \mathbb{R}$,*

$$\begin{aligned} \mathcal{R}(z|x) &= \mathbf{E}[(Y - z)^2|X = x] \\ &= \mathbf{E}[(Y - \mathbf{E}[Y|X = x] + \mathbf{E}[Y|X = x] - z)^2|X = x] \\ &= \mathbf{E}[(Y - \mathbf{E}[Y|X = x])^2|X = x] + \underbrace{2\mathbf{E}[(Y - \mathbf{E}[Y|X = x])(\mathbf{E}[Y|X = x] - z)|X = x]}_{=0} \\ &\quad + (\mathbf{E}[Y|X = x] - z)^2 \end{aligned}$$

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

Thus z minimizes this quantity iff $z = \mathbf{E}[Y|X=x]$ since this cancels the last term, the only term that depends on z. The the Bayes predictor is $f^*(x) = \mathbf{E}[Y|X=x]$ and the Bayes risk is

$$\mathcal{R}^* = \mathbf{E}[(Y - \mathbf{E}[Y|X=x])^2].$$

- *Assume that ρ has finite first moments. We define the function $h_x(z) = \mathbf{E}[|Y-z| \mid X=x]$ and denoting ρ_x the density of $Y|X=x$, we have*

$$h_x(z) = \int_{-\infty}^{z} (z-y)\rho_x(y)\mathrm{d}y + \int_{z}^{+\infty} (y-z)\rho_x(y)\mathrm{d}y$$

and thus, by Leibniz integral rule[1], and our assumption that ρ_x is continuous, h_x is diffentiable in z with differential

$$h_x'(z) = \int_{-\infty}^{z} \rho_x(y)\mathrm{d}y - \int_{z}^{+\infty} \rho_x(y)\mathrm{d}y$$

and $h_x'(z) = 0$ iff $\int_{-\infty}^{z} \rho_x(y)\mathrm{d}y = \int_{z}^{+\infty} \rho_x(y)\mathrm{d}y$. Such a z (which can be non-unique) is called a median *of ρ_x and this property characterizes minimizers of h_x because h_x is convex and differentiable in z. Thus Bayes predictors are exactly the functions f^* such that for all $x \in \mathcal{X}$, $f^*(x)$ is a median of ρ_x. The Bayes risk is $\mathcal{R}^* = \mathbf{E}[|Y - m(X)|]$ (there is no simpler formula). [NB: a more general argument, that does not need the regularity assumption on ρ_x, could be made using the notion of subgradient].*

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?*

Solution 2. *In a supervised learning setting, conditioning on X, the prediction $f(X)$ is independent from Y (otherwise we'd be cheating!). Thus, denoting $f^*(x) = \mathbf{E}[Y|X=x]$ the Bayes predictor (which is deterministic), we have*

$$\begin{aligned}
\mathbf{E}[(f(X)-Y)^2|X=x] &= \mathbf{E}[(f(X)-f^*(X)+f^*(X)-Y)^2|X=x] \\
&= \mathbf{E}[(f(X)-f^*(X))^2|X=x] + 2\mathbf{E}[(f^*(X)-f^*(X))(f^*(X)-Y)|X=x] \\
&\quad + \mathbf{E}[(f^*(X)-Y)^2|X=x] \\
&= \mathbf{E}[(f(X)-f^*(X))^2|X=x] + \mathbf{E}[(f^*(X)-Y)^2|X=x] \\
&= 2\mathbf{E}[(f^*(X)-Y)^2|X=x]
\end{aligned}$$

where the cross-term vanishes by independence and in the last line we have used our assumption that the law of $f(X)|X=x$ is the same as $Y|X=x$. Thus f achieves the Bayes risk $\mathcal{R}^* = \mathbf{E}[(f^*(X)-Y)^2]$ iff $\mathbf{E}[(f^*(X)-Y)^2|X=x] = 0$ almost surely (because of the factor 2), i.e. when $Y = f^*(X)$ almost surely. More generally, this exercise tells us that it never helps to make a random prediction, from the point of view of the 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*

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

[1] https://en.wikipedia.org/wiki/Leibniz_integral_rule

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

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

Solution 3. *The following can be checked by computating partial derivatives.*
(i) For a general A, $f'(x) = \frac{1}{2}(Ax + A^\top x) - b$ and $f''(x) = A + A^\top$. For a symmetric A, $f'(x) = Ax - b$ and $f''(x) = A$.
(ii) $f'(w) = X^\top(Xw - y)$ and $f''(w) = X^\top X$.

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 (for some $\delta : \mathbb{R}_+ \to [0, 1]$)

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

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

(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 $a\mathbf{P}(X \geq a) \leq \mathbf{E}[X]$.] NB: This bound is very weak : we usually look for PAC bounds where $\delta(\epsilon)$ decreases exponentially fast.

Solution 4. *The definition of in-expectation and PAC bounds can be found in the lecture notes (Lecture 1, Section 7).*

(i) Let $Z = \mathcal{R}(\mathcal{A}(\mathcal{D})) - \mathcal{R}^*$ and we know $\mathbf{P}(Z > \epsilon) \leq \delta(\epsilon)$. Since Z is a nonnegative random variable, it holds (by Fubini-Tonelli)

$$\mathbf{E}[Z] = \mathbf{E}\Big[\int_0^\infty 1_{u<Z}\mathrm{d}u\Big] = \int_0^\infty \mathbf{E}[1_{u<Z}]\mathrm{d}u = \int_0^\infty \mathbf{P}(Z > u)\mathrm{d}u \leq \int_0^\infty \delta(\epsilon)\mathrm{d}\epsilon.$$

(ii) Again, let $Z = \mathcal{R}(\mathcal{A}(\mathcal{D})) - \mathcal{R}^*$. By Markov inequality, it holds for $\epsilon > 0$,

$$\mathbf{P}(Z \geq \epsilon) \leq \epsilon^{-1}\mathbf{E}[Z] \leq \alpha/\epsilon =: \delta(\epsilon).$$

Exercise 5 (Does there exist best algorithms?)**.** *We consider binary classification $\mathcal{Y} = \{0, 1\}$ with the* 0-1 *loss. 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.

Solution 5. *We may just choose $\mathcal{A}_\rho(\mathcal{D}) = f^*$ where f^* is the Bayes predictor (which is shown to exist in exercise 1). I.e. $\mathcal{A}$ does not depend on the training set and always output the same function. (This says that learning is non-trivial only if one considers* a class *of distributions 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* $\mathcal{X} = \{1, \dots, k\}$*. 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.*

- *(⋆) 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 samples.*

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

Solution 6. *See the proof of Thm.2.1 in [Bach, 2022, Sec.2.5].*

References

Francis Bach. Learning theory from first principles. *Draft of a book, version of Sept.*, 6:2022, 2022.