Machine Learning for EDS

TUTORIAL WEEK 2 : SOLUTIONS

2023/2024

Important: this file is meant for students of the course Machine Learning for EDS (2023/2024) and is not allowed to be distributed to others.

Problem 1

1. We are in a binary classification framework with 0-1 cost.

2. By the definition $\eta(X) = \mathbb{E}[Y|X]$ and the definition of an expectation, we have that

$$\eta(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) \cdot 1 + \mathbb{P}(Y = 0|X) \cdot 0 = \mathbb{P}(Y = 1|X)\,.$$

3. We know from the lectures that for this framework, a Bayes predictor and Bayes risk are of the form

$$f^* : x \mapsto \mathbb{1}_{\eta(x)>1/2}, \qquad \mathcal{R}_P^* = \mathbb{E}\Big[\min\big(\eta(X), 1-\eta(X)\big)\Big],$$

where $\eta(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$. Recall that any predictor f such that $\{f(X) = f^*(X)$ or $\eta(X) = \frac{1}{2}\}$ a.s. is a Bayes predictor.

4. We have:

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 2) = 1 - p,$$
$$\mathbb{P}(Y = 0|X = 1) = 1/2, \qquad \mathbb{P}(Y = 0|X = 2) = 1/6,$$
$$\mathbb{P}(Y = 1|X = 1) = 1/2, \qquad \mathbb{P}(Y = 1|X = 2) = 5/6.$$

We thus have:

$$\eta(X) = \frac{1}{2}\mathbb{1}_{X=1} + \frac{5}{6}\mathbb{1}_{X=2}, \qquad f^* : x \in \{1,2\} \mapsto \mathbb{1}_{x=2}.$$

Notice that $f^{**} : x \in \{1, 2\} \mapsto 1$ is also a Bayes predictor. Bayes risk is equal to:

$$\begin{aligned}
\mathcal{R}_P^* &= \mathbb{E}\Big[\min\big(\eta(X), 1-\eta(X)\big)\Big] \\
&= \min\big(\eta(1), 1-\eta(1)\big)\mathbb{P}(X = 1) + \min\big(\eta(2), 1-\eta(2)\big)\mathbb{P}(X = 2) \\
&= p\min\big(1/2, 1-1/2\big) + (1-p)\min\big(5/6, 1-5/6\big) \\
&= p/2 + (1-p)/6
\end{aligned}$$

$$= \frac{1}{6} + \frac{p}{3},$$

where we simply use the definition of an expectation and the form of $\eta(X)$ given above.

5. Bayes risk is an increasing function of p. It is minimal for $p = 0$ and equal to $1/6$, which is equal to the probability of a wrong prediction regarding the outcome of die 2. For $p = 1$, Bayes risk is maximum and is equal to the probability of a wrong prediction for die 1, which is as unpredictable as the outcome of a toss of a fair coin. It is intuitive that the Bayes risk increases in p, because a roll of die 1 is harder to predict than a roll of die 2, the Bayes' risk increases as the probability of choosing die 1 increases.

Problem 2

1. (a) For c the absolute value cost, the risk of a predictor $f : \mathcal{X} \to \mathbb{R}$ reads:

$$\mathcal{R}_P^c(f) = \mathbb{E}[|f(X) - Y|].$$

Fix $x \in \mathcal{X}$ and let us analyse the function $r_x : a \mapsto r_x(a) = \mathbb{E}\Big[|a - Y|\Big|X = x\Big]$ to find its minimum.

Extra explanation for why we start by analysing this function $r_x(a)$: For any x, the value of a that minimizes $r_x(a)$, say a_x^*, is the prediction (conditional on x) that would minimize the conditional expected loss. So conditional on x, a_x^* is the 'ideal prediction', in other words, the 'ideal predictor' f^* evaluated in x should take the value $f^*(x) = a_x^*$.

We have for any $M > a$:

$$\begin{aligned}
r_x(a) &= \int_{\mathbb{R}} |a-y| f_{Y|X=x}(y)dy \\
&= \int_{-\infty}^{a} (a-y) f_{Y|X=x}(y)dy + \int_{a}^{+\infty} (y-a) f_{Y|X=x}(y)dy \\
&= aF_{Y|X=x}(a) - \int_{-\infty}^{a} y f_{Y|X=x}(y)dy + \int_{a}^{+\infty} y f_{Y|X=x}(y)dy - a(1 - F_{Y|X=x}(a)) \\
&= a(2\ F_{Y|X=x}(a) - 1) - \int_{-\infty}^{a} y f_{Y|X=x}(y)dy + \int_{a}^{+\infty} y f_{Y|X=x}(y)dy \\
&= a(2\ F_{Y|X=x}(a) - 1) - \int_{-M}^{a} y f_{Y|X=x}(y)dy - \int_{-\infty}^{-M} y f_{Y|X=x}(y)dy \\
&\quad + \int_{a}^{M} y f_{Y|X=x}(y)dy + \int_{M}^{+\infty} y f_{Y|X=x}(y)dy\,,
\end{aligned}$$

where $F_{Y|X=x}(a) = \mathbb{P}(Y \leq a|X = x)$. Recall that if $\varphi(x) = \displaystyle\int_{A(x)}^{B(x)} g(x,t)dt$ (assuming all functions involved are differentiable):

$\varphi'(x) = \displaystyle\int_{A(x)}^{B(x)} \frac{\partial g(x,t)}{\partial x} dt + B'(x)g(x,B(x)) - A'(x)g(x,A(x))$. Furthermore, we know that the derivative of the cdf $F_{Y|X=x}(a)$ with respect to a is equal to the pdf $f_{Y|X=x}(a)$. It follows that:

$$\begin{aligned} r_x'(a) &= 2\ F_{Y|X=x}(a) - 1 + 2af_{Y|X=x}(a) - 1 \cdot af_{Y|X=x}(a) - 1 \cdot af_{Y|X=x}(a) \\ &= 2\ F_{Y|X=x}(a) - 1\,. \end{aligned}$$

Furthermore, by differentiating again, we have for any a:

$$r_x''(a) = 2f_{Y|X=x}(a)\,.$$

(b) The function r reaches a critical point ($r_x'(a) = 0$) at a_x^*:

$$r_x'(a_x^*) = 0 \iff 2F_{Y|X=x}(a_x^*) - 1 = 0 \iff F_{Y|X=x}(a_x^*) = \frac{1}{2}\,,$$

in other words:

$$\mathbb{P}(Y \leq a_x^* | X = x) = 1/2 = \mathbb{P}(Y \geq a_x^* | X = x),$$

which by definition is $a_x^* = \mathrm{med}(Y|X = x)$, which stands for the median of Y conditionally on $X = x$. Furthermore, $r_x''(a) = 2f_{Y|X=x}(a) \geq 0$, which shows that $r : a \mapsto r_x(a)$ is convex. Hence, a_x^* is a global minimum of $r_x : a \mapsto r_x(a)$.

(c) Letting $f^* : x \mapsto \mathrm{med}(Y|X = x)$, we have by construction of f^* that for any predictor $f : \mathcal{X} \to \mathbb{R}$:

$$\mathbb{E}\Big[|f^*(X) - Y|\Big|X\Big] \leq \mathbb{E}\Big[|f(X) - Y|\Big|X\Big].$$

Notice: we use here that $r_x(a) = \mathbb{E}[|a - Y||X = x]$ is minimized by $a_x^* = \mathrm{med}(Y|X = x)$ (which we saw in (b)). Instead of filling in a specific value of X, we can also condition on the random variable X, so then we have $r_X(a) = \mathbb{E}[|a - Y||X]$, which is still clearly minimized by $a_X^* = \mathrm{med}(Y|X)$ (conditional on X). The inequality above is an immediate consequence of this.

By taking expectations, we thus have that for any predictor f:

$$\mathbb{E}\Big[\mathbb{E}\Big[|f^*(X) - Y|\Big|X\Big]\Big] \leq \mathbb{E}\Big[\mathbb{E}\Big[|f(X) - Y|\Big|X\Big]\Big]$$

$$\begin{aligned} \iff \quad & \mathbb{E}\Big[|f^*(X) - Y|\Big] \le \mathbb{E}\Big[|f(X) - Y|\Big], \\ \iff \quad & \mathcal{R}_P(f^*) \le \mathcal{R}_P(f) \end{aligned}$$

by the law of iterated expectations and the definition of risk. Hence, for any $f \in \mathcal{F}$, $\mathcal{R}_P(f^*) \le \mathcal{R}_P(f)$, which yields that $\mathcal{R}_P(f^*) = \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$.

$$\boxed{f^* : x \mapsto \mathrm{med}(Y|X = x) \text{ is a Bayes predictor.}}$$

2. For c the cost $c(y, y') = c_-(y' - y)\mathbb{1}_{y<y'} + c_+(y - y')\mathbb{1}_{y>y'}$, the proof is very similar to that of point 1. Verify yourself that here

$$\boxed{f^* : x \mapsto F^{-1}_{Y|X=x}(q) \text{ with } q = \frac{c_-}{c_- + c_+} \text{ is a Bayes predictor.}}$$

where $F^{-1}_{Y|X=x}(q)$ denotes the inverse of the cdf of Y given $X = x$ at q, or equivalently, the q-quantile of the distribution of Y given $X = x$. In other words, $F^{-1}_{Y|X=x}(q)$ is equal to $a \in \mathbb{R}$ in case $\mathbb{P}(Y \le a|X = x) = q$ (and then automatically $\mathbb{P}(Y > a|X = x) = 1 - q$)

Problem 3

1. We are in a binary classification framework for which we know that Bayes predictors are of the form:

$$f^* : x \mapsto \mathbb{1}_{\eta(x)>1/2},$$

with $\eta : x \mapsto \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$.

2. Using that $\eta(x) = \mathbb{P}(Y = 1|X = x)$, we have that for almost any $x \in \mathbb{R}^d$:

$$\begin{aligned} \eta(x) > 1/2 \iff \quad & \mathbb{P}(Y = 1|X = x) > 1/2 \quad (\text{so } \mathbb{P}(Y = 0|X = x) \le 1/2) \\ \iff \quad & \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x) \\ \iff \quad & \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} > 1. \end{aligned}$$

3. Using Bayes' theorem (see mathematical reminder slides), we have for $i = 0, 1$ and almost any $x \in \mathbb{R}^d$ that

$$\mathbb{P}(Y = i|X = x) = \frac{f_{X|Y=i}(x)\mathbb{P}(Y = i)}{f_X(x)} = \frac{N_i(x)\mathbb{P}(Y = i)}{f_X(x)},$$

where N_i denotes the pdf of $\mathcal{N}(\mu_i, \Sigma_i)$. So $\mathbb{P}(Y = 1|X = x) = N_1(x)p/f_X(x)$ and $\mathbb{P}(Y = 0|X = x) = N_0(x)(1 - p)/f_X(x)$

4. Recall that the Bayes classifier for 0-1 cost is given by

$$f^* : x \mapsto \mathbb{1}_{\eta(x)>1/2},$$

so we only have to find an expression for the condition $\eta(x) > 1/2$ for this setting. We assume Σ_0 and Σ_1 are invertible. Recall that if Σ_i is invertible (i.e. positive definite in this case, all the eigenvalues of Σ_i are strictly positive), then the density of $\mathcal{N}(\mu_i, \Sigma_i)$ writes for any $x \in \mathbb{R}^d$:

$$N_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\Big(- \frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)\Big),$$

Hence, for almost any $x \in \mathbb{R}^d$, we have by points 2 and 3 that:

$$\begin{aligned}
\eta(x) > 1/2 \iff & \quad \frac{pN_1(x)}{(1-p)N_0(x)} > 1 \\
\iff & \quad \ln\Big(\frac{p}{1-p}\Big) + \ln\Big(\frac{N_1(x)}{N_0(x)}\Big) > 0 \\
\iff & \quad \ln\Big(\frac{p}{1-p}\Big) + \frac{1}{2}\ln\Big(\frac{|\Sigma_0|}{|\Sigma_1|}\Big) - \frac{1}{2}\Big((x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)\Big) > 0 \\
\iff & \quad x^\top A x + b^\top x + c > 0,
\end{aligned}$$

with:

$$\begin{aligned}
A &= \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}), \\
b &= \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0, \\
c &= \ln\Big(\frac{p}{1-p}\Big) + \frac{1}{2}\ln\Big(\frac{|\Sigma_0|}{|\Sigma_1|}\Big) + \frac{1}{2}\Big(\mu_0^\top \Sigma_0^{-1}\mu_0 - \mu_1^\top \Sigma_1^{-1}\mu_1\Big).
\end{aligned}$$

It follows that the Bayes classifier can be expressed as

$$f^* : x \in \mathbb{R}^d \mapsto \mathbb{1}_{x^\top Ax + b^\top x + c > 0},$$

5. (Less important) The decision boundary $x^\top Ax + b^\top x + c = 0$ is the equation of a quadric surface (see `https://en.wikipedia.org/wiki/Quadric`) in dimension d. In dimension $d = 2$ for instance, it can take the following forms: ellipse, hyperbola, parabola, circle, two intersecting straight lines, two parallel straight lines or a single straight line.
   In particular, if $\Sigma_0 = \Sigma_1$, then $A = 0_{d\times d}$ and the decision boundary is given by $b^\top x + c = 0$, which is the equation of an affine hyperplane (see `https://en.wikipedia.org/wiki/Hyperplane#Affine_hyperplanes`).

Problem 4

1. Let c the asymmetric 0-1 cost function: $c(y,y') = w_{y'}\mathbb{1}_{y\neq y'}$, for $y,y' \in \{0,1\}$, with $w_0, w_1 \geq 0$, $w_0 + w_1 > 0$, and consider

$$\begin{aligned} r_X(f) = \mathbb{E}\big[c\big(f(X),Y\big)\big|X\big] &= \mathbb{E}\big[w_Y\mathbb{1}_{f(X)\neq Y}\big|X\big] \\ &= \mathbb{E}\big[w_0\mathbb{1}_{f(X)\neq 0}\mathbb{1}_{Y=0} + w_1\mathbb{1}_{f(X)\neq 1}\mathbb{1}_{Y=1}\big|X\big] \\ &= w_0\mathbb{1}_{f(X)=1}\mathbb{E}\big[\mathbb{1}_{Y=0}\big|X\big] + w_1\mathbb{1}_{f(X)=0}\mathbb{E}\big[\mathbb{1}_{Y=1}\big|X\big] \\ &= w_0\mathbb{1}_{f(X)=1}(1-\eta(X)) + w_1\mathbb{1}_{f(X)=0}\eta(X) \\ r_X(f) &= w_0 f(X)(1-\eta(X)) + w_1(1-f(X))\eta(X) \end{aligned}$$

2. In the expression of $r_X(f)$ above, $f(X)$ can be either 0 or 1, so the first or second term disappears when an X is plugged in. Hence, for any predictor $f \in \mathcal{F}$:

$$r_X(f) \geq \min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big), \tag{0.1}$$

and we will show that an equality occurs for $f = f^*$:

$$r_X(f^*) = \min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big). \tag{0.2}$$

Indeed, we have for $f = f^*$ as defined in the question:

$$\begin{aligned} r_X(f^*) &= w_0 f^*(X)(1-\eta(X)) + w_1(1-f^*(X))\eta(X) \\ &= w_0\mathbb{1}_{\eta(X)>\frac{w_0}{w_0+w_1}}(1-\eta(X)) + w_1\mathbb{1}_{\eta(X)\leq\frac{w_0}{w_0+w_1}}\eta(X) \end{aligned}$$

Consider two cases: 1) $\eta(X) > \frac{w_0}{w_0+w_1}$, and 2) $\eta(X) \leq \frac{w_0}{w_0+w_1}$. In the first case, $f^*(X) = \mathbb{1}_{\eta(X)>\frac{w_0}{w_0+w_1}}$ takes value 1, so,

$$r_X(f^*) = w_0(1-\eta(X)) < w_0\Big(1 - \frac{w_0}{w_0+w_1}\Big) = \frac{w_0 w_1}{w_0+w_1} < w_1\eta(X),$$

so in that case indeed $r_X(f^*) = w_0(1-\eta(X)) = \min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big)$. Case 2) is similar (verify yourself!). So indeed

$$r_X(f^*) = \min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big).$$

3. Hence, by taking expectation on both sides of (0.1) and (0.2), and by using the law of iterated expectations, we get that for any predictor $f \in \mathcal{F}$:

$$\mathcal{R}_P^c(f) = \mathbb{E}[r_X(f)] \geq \mathbb{E}\Big[\min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big)\Big],$$

$$\mathcal{R}_P^c(f^*) = \mathbb{E}[r_X(f^*)] = \mathbb{E}\Big[\min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big)\Big].$$

We have shown that the risk of any predictor f is lower-bounded by $\mathbb{E}\Big[\min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big)\Big]$ and that this lower-bound is reached for f^*, thus:

$$\boxed{\begin{array}{l} f^* : x \mapsto \mathbb{1}_{\eta(X) > \frac{w_0}{w_0+w_1}} \text{ is a Bayes predictor, and Bayes risk reads} \\ \qquad\qquad \mathcal{R}_P^* = \mathbb{E}\Big[\min\Big(w_0(1-\eta(X)), w_1\eta(X)\Big)\Big]. \end{array}}$$

4. The definition if excess risk reads:

$$\ell(f^*, f) = \mathcal{R}_P^c(f) - \mathcal{R}_P^*$$

So if we have a Bayes classifier f^*, we can write

$$\ell(f^*, f) = \mathcal{R}_P^c(f) - \mathcal{R}_P^c(f^*) = \mathbb{E}\Big[r_X(f) - r_X(f^*)\Big]$$

5. By part 1 and simple linear algebra, we have

$$\begin{aligned} r_X(f) - r_X(f^*) &= [w_0 f(X)(1-\eta(X)) + w_1(1-f(X))\eta(X)] - [w_0 f^*(X)(1-\eta(X)) + w_1(1-f^*(X))\eta(X)] \\ &= (w_0+w_1)(f(X) - f^*(X))\left(\frac{w_0}{w_0+w_1} - \eta(X)\right). \end{aligned}$$

Then there are two possible cases: (i) $f(X) = f^*(X)$, in which case $r_X(f) - r_X(f^*) = 0$ or (ii) $f(X) \neq f^*(X)$, in which case there are two options:

- $f^*(X) = 0$ (which occurs if $\eta(X) \leq \frac{w_0}{w_0+w_1}$) and $f(X) = 1$. Then:

$$(f(X) - f^*(X))\left(\frac{w_0}{w_0+w_1} - \eta(X)\right) = \left(\frac{w_0}{w_0+w_1} - \eta(X)\right) = \left|\frac{w_0}{w_0+w_1} - \eta(X)\right|$$

- $f^*(X) = 1$ (which occurs if $\eta(X) > \frac{w_0}{w_0+w_1}$) and $f(X) = 0$. Then:

$$(f(X) - f^*(X))\left(\frac{w_0}{w_0+w_1} - \eta(X)\right) = -\left(\frac{w_0}{w_0+w_1} - \eta(X)\right) = \left|\frac{w_0}{w_0+w_1} - \eta(X)\right|$$

So combining these results, we have

$$r_X(f) - r_X(f^*) = (w_0+w_1)\mathbb{1}_{f(X)\neq f^*(X)}\left|\frac{w_0}{w_0+w_1} - \eta(X)\right|,$$

So using the definition of $r_X(f)$ of part 1 and the law of iterated expectations:

$$\begin{aligned} \mathbb{E}\left[r_X(f) - r_X(f^*)\right] &= \mathbb{E}\left[\mathbb{E}[c(f(X), Y)|X] - \mathbb{E}[c(f^*(X), Y)|X]\right] \\ &= \mathbb{E}[c(f(X), Y) - \mathbb{E}[c(f^*(X), Y)] = \mathcal{R}_P^c(f) - \mathcal{R}_P^* = \ell(f^*, f) \end{aligned}$$

Hence the expression of the excess risk is:

$$\ell(f^*, f) = (w_0 + w_1)\mathbb{E}\Big[\mathbb{1}_{f(X)\neq f^*(X)}\Big|\frac{w_0}{w_0+w_1} - \eta(X)\Big|\Big]$$

Bayes predictors are such that their excess risk is zero almost surely: $\ell(f^*, f) = 0$ a.s. A predictor $f \in \mathcal{F}$ is a Bayes predictor if and only if:

$$\begin{aligned} & & (w_0 + w_1)\mathbb{E}\Big[\underbrace{\mathbb{1}_{f(X)\neq f^*(X)}\Big|\frac{w_0}{w_0+w_1} - \eta(X)\Big|}_{\geq 0}\Big] = 0 \\ \iff & & \mathbb{1}_{f(X)\neq f^*(X)}\Big|\frac{w_0}{w_0+w_1} - \eta(X)\Big| = 0, \text{ a.s.} \\ \iff & & \Big\{f(X) = f^*(X) \text{ or } \eta(X) = \frac{w_0}{w_0+w_1}\Big\} \text{ a.s.} \end{aligned}$$

Notice that the first equivalence holds because the expectation of a non-negative random variable, can only be equal to zero if the random variable is equal to zero with probability one. Therefore:

A predictor $f \in \mathcal{F}$ is a Bayes predictor if and only if $f(X) = f^*(X)$, a.s., except perhaps on events such that $\eta(X) = \dfrac{w_0}{w_0+w_1}$.