Exercise 1.1

Since we are dealing with binary classification, we use the $0-1$ loss

$$l_{0-1}(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y) = \begin{cases} 1\,, & \text{if } \hat{y} \neq y \\ 0\,, & \text{if } \hat{y} = y \end{cases}$$

where $\mathbb{1}$ is the indicator function,
$\hat{y}$ the predicted class and y the true class.

The empirical risk (training error) of the predictor h_s° is

$$\begin{aligned} \hat{\mathcal{R}}_s(h_s^\circ) &= \frac{1}{n}\sum_{i=1}^{n} l_{0-1}(h_s^\circ(x_i), y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h_s^\circ(x_i) \neq y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq y_i) = 0 \end{aligned}$$

Indeed h_s° classifies every training input correctly since $h_s^\circ(x_i) = y_i$.

Assume $\mathcal{X} = \mathbb{R}$. Let $q(x) = \prod_{i\,:\,y_i=0}(x - x_i)$.

By construction, the roots are the x_i such that $y_i = 0$. In particular

$$\begin{aligned} q(x_i) &= 0\,, && \text{if } y_i = 0 \\ q(x_i) &\neq 0\,, && \text{if } y_i = 1 \\ q(x\,) &\neq 0\,, && \forall\, x \notin \{x_i\}_{i=1,\dots,n} \end{aligned}$$

Notice that

$$\begin{aligned} h_s^\circ(x_i) &= 0\,, && \text{if } y_i = 0 \\ h_s^\circ(x_i) &= 1\,, && \text{if } y_i = 1 \\ h_s^\circ(x\,) &= 1\,, && \forall\, x \notin \{x_i\}_{i=1,\dots,n} \end{aligned}$$

So to get the condition, we just need to change $\neq 0$ into > 0 .
To do so, we just square

$$p_s(x) = \big(q(x)\big)^2$$

Now p_s have the same zeroes of q , namely the x_i such that $y_i = 0$; and otherwise, which is when h_s° predicts 1, it is strictly positive.

Exercise 1.2

We prove a slightly more general statement, valid for bounded losses.

Let be

- $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$
- $h : \mathcal{X} \to \mathcal{Y}$ a fixed predictor / hypothesis
- $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a bounded loss (for example with values in $[0,1]$)

Then the theoretical risk is

$$\mathcal{R}_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h(x), y)]$$

and the empirical risk for an i.i.d sample $S = \{(x_i, y_i)\}_{i=1,\dots,n} \sim \mathcal{D}^n$ is

$$\hat{\mathcal{R}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$$

For each $i = 1, \dots, n$ define the random variable

$$Z_i = l(h(x_i), y_i)$$

It’s a random variable because it depends on the chosen sample.

By definition

$$\hat{\mathcal{R}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i) = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

and by linearity of expectation (on the samples $\sim\mathcal{D}^n$)

$$\mathbb{E}_S\left[\hat{\mathcal{R}}_S(h)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S[Z_i]$$

Since the loss l is bounded, the Z_i are integrable,
that is they have finite expectation.

Because the sample points (x_i, y_i) are drawn i.i.d from $\mathcal{D}$, each random variable $Z_i = l(h(x_i), y_i)$ has the same distribution as $l(h(x), y)$ for $(x, y) \sim \mathcal{D}$. Hence

$$\mathbb{E}_S[Z_i] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h(x), y)] = \mathcal{R}_{\mathcal{D}}(h)$$

We conclude

$$\mathbb{E}_S\left[\hat{\mathcal{R}}_S(h)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S[Z_i] = \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_{\mathcal{D}}(h) = \mathcal{R}_{\mathcal{D}}(h)$$

We can prove something stronger yet. Indeed, since $\{Z_i\}_{i=1,\dots,n}$ are i.i.d real-valued random variables with finite expectations,
we can apply the strong law of large numbers to get

$$\hat{\mathcal{R}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_S[Z_1] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h(x), y)] = \mathcal{R}_{\mathcal{D}}(h)$$

Thus the empirical risk converges to the theoretical risk for bigger and bigger samples.

Exercise 1.4

The Bayes classifier is the classifier h_B that assigns to a test input x the class c^* that maximizes the probability

$$\Pr\,(Y = c^* \mid X = x)$$

In other words, if $\mathcal{Y} = \{1, \dots, K\}$ then

$$\begin{aligned} h_B(x) &= \arg\max_{c\,\in\,\mathcal{Y}} \Pr\,(Y = c \mid X = x) \\ &= \arg\max_{c\,\in\,\mathcal{Y}} p(c \mid x) \end{aligned}$$

Let the pointwise risk of a classifier h at x be

$$\begin{aligned} \mathcal{R}(h\,; x) &= \mathbb{E}\,[\, l(h(x), Y) \mid X = x \,] \\ &= \sum_{c=1}^{K} l(h(x), Y)\; p(c \mid x) \end{aligned}$$

For the $0 - 1$ loss this becomes

$$\mathcal{R}(h\,; x) = \sum_{c\,\neq\,h(x)} p(c \mid x) = 1 - p(h(x) \mid x)$$

and for the Bayes classifier we use the previous expression of $h_B(x)$

$$\begin{aligned} \mathcal{R}(h_B\,; x) &= 1 - p(h_B(x) \mid x) \\ &= 1 - \max_{c\,\in\,\mathcal{Y}} p(c \mid x) \\ &\leq 1 - p(h(x) \mid x) \\ &= \mathcal{R}(h\,; x) \end{aligned}$$

and we get the inequality for every other classifier h .

Therefore the Bayes classifier minimizes the pointwise risk and thus also the overall risk (we integrate w.r.t the same marginal density)

$$\mathcal{R}(h_B) = \mathbb{E}_X[\,\mathcal{R}(h_B\,; X)\,] \leq \mathbb{E}_X[\,\mathcal{R}(h\,; X)\,] = \mathcal{R}(h)$$

Exercise 1.12

https://colab.research.google.com/drive/1eg2HWNxi7TXOld4uvGrM2OAjRY4DK773?usp=sharing

---

Exercise 1.15

https://colab.research.google.com/drive/1XkbKN5FkdXNZMMgc80cFD8OU9E_Ek8ig?usp=sharing

---