INTRODUCTION TO MACHINE LEARNING

1 Bayes classifier

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Assume that (X, Y) is a couple of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathsf{X} \times \{-1, 1\}$ where X is a given state space, which means that the focus is set on a two-class classification problem. One aim of supervised classification is to define a function $h : \mathsf{X} \to \{-1, 1\}$, called *classifier*, such that $h(X)$ is the best prediction of Y in a given context. For instance, the risk of misclassification of h is

$$\mathsf{R}_{\text{miss}}(h) = \mathbb{E}\left[\mathbb{1}_{Y \neq h(X)}\right] = \mathbb{P}\left(Y \neq h(X)\right) .$$

Note that $\mathbb{E}[Y|X]$ is a random variable measurable with respect to the σ-algebra $\sigma(X)$. Therefore, there exists a function $\eta : \mathsf{X} \to [-1, 1]$ so that $\mathbb{E}[Y|X] = \eta(X)$ almost surely.

Lemma 1. *The classifier $h_\star$, defined for all $x \in \mathsf{X}$, by*

$$h_\star(x) = \begin{cases} 1 & \text{if } \eta(x) > 0\,, \\ -1 & \text{otherwise}\,, \end{cases}$$

is such that

$$h_\star = \underset{h:\mathsf{X}\to\{-1,1\}}{\text{argmin}} \; \mathsf{R}_{\text{miss}}(h)\,.$$

Proof. For all $u, v \in \{-1, 1\}$, $\mathbb{1}_{u \neq v} = \mathbb{1}_{uv=-1} = (1 - uv)/2$. Since Y and $h(X)$ take values in $\{-1, 1\}$, this implies

$$\mathsf{R}_{\text{miss}}(h) = \mathbb{P}(Y \neq h(X)) = \left(1 - \mathbb{E}[Y h(X)]\right)/2\,.$$

Now, using sucessively the tower property, the equality $|u| = u \times \text{sign}(u)$, and the tower property again,

$$\mathbb{E}[Y h(X)] = \mathbb{E}[\mathbb{E}[Y|X] h(X)] \leq \mathbb{E}[|\mathbb{E}[Y|X]| \underbrace{|h(X)|}_{=1}] = \mathbb{E}[\mathbb{E}[Y|X] \underbrace{\text{sign}(\mathbb{E}[Y|X])}_{h_\star(X)}] = \mathbb{E}[Y h_\star(X)]\,.$$

This yields $\mathsf{R}_{\text{miss}}(h) \geq \mathsf{R}_{\text{miss}}(h_\star)$, which concludes the proof. □

2 Empirical risk minimization

First, we do not assume that the joint law of (X, Y) belongs to any parametric or semiparametric family of models. Instead, we make some restrictions on the set of classifiers on which the optimisation occurs.

More precisely, we consider that the optimization of classifiers holds on a specific set $\mathcal{H}$ of classifiers (often called the *dictionary*), which may possibly not contain the Bayes classifier. Moreover, since in most cases, the classification risk R_{miss} cannot be computed nor minimized, it is instead estimated by the empirical classification risk defined as

$$\widehat{\mathsf{R}}^n_{\text{miss}}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq h(X_i)}\,,$$

where $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ are independent observations with the same distribution as (X, Y). The classification problem then boilds down to solving

$$\widehat{h}^n_{\mathcal{H}} \in \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h)\,.$$

In this context several practical and theoretical challenges arise from the minimization of the empirical classification risk. The choice of $\mathcal{H}$ is pivotal in designing an efficient classification procedure. Note that choosing $\mathcal{H}$ as all possible classifiers is meaningless, in this case, $\widehat{h}^n_{\mathcal{H}}$ is such that $\widehat{h}^n_{\mathcal{H}}(X_i) = Y_i$ for all $1 \leqslant i \leqslant n$ and $\widehat{h}^n_{\mathcal{H}}(x)$ is any element of $\{-1, 1\}$ for all $x \notin \{X_1, \ldots, X_n\}$. Although $\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h^n_{\mathcal{H}}) = 0$, is likely to be a poor approximation of $\mathsf{R}_{\mathrm{miss}}(h^n_{\mathcal{H}})$. To understand this, the excess misclassification risk may be decomposed as follows

$$\mathsf{R}_{\mathrm{miss}}\left(\widehat{h}^n_{\mathcal{H}}\right) - \mathsf{R}_{\mathrm{miss}}\left(h_\star\right) = \mathsf{R}_{\mathrm{miss}}\left(\widehat{h}^n_{\mathcal{H}}\right) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) + \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) - \mathsf{R}_{\mathrm{miss}}\left(h_\star\right) \geqslant 0\,.$$

The first term of the decomposition $\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right)$ is a **stochastic error** which is likely to grow when the size of $\mathcal{H}$ grows while $\min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) - \mathsf{R}_{\mathrm{miss}}\left(h_\star\right)$ is **deterministic** and likely to decrease as the size of $\mathcal{H}$ grows.

Lemma 2. *For all set $\mathcal{H}$ of classifiers and all $n \geqslant 1$,*

$$\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) \leqslant 2 \sup_{h \in \mathcal{H}} \left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}\left(h\right)\right|\,.$$

Proof. By definition of $\widehat{h}^n_{\mathcal{H}}$, for any $h \in \mathcal{H}$,

$$\begin{aligned}
\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) &= \mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) + \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right)\,, \\
&\leqslant \mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) + \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right)\,.
\end{aligned}$$

For all $\varepsilon > 0$ there exists $h_\varepsilon \in \mathcal{H}$ such that $\mathsf{R}_{\mathrm{miss}}(h_\varepsilon) < \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) + \varepsilon$ so that

$$\begin{aligned}
\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \min_{h \in \mathcal{H}} \mathsf{R}_{\mathrm{miss}}\left(h\right) &\leqslant \mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) + \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_\varepsilon) - \mathsf{R}_{\mathrm{miss}}(h_\varepsilon) + \varepsilon\,, \\
&\leqslant 2 \sup_{h \in \mathcal{H}} \left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}\left(h\right)\right| + \varepsilon\,,
\end{aligned}$$

which concludes the proof. □