Day 2. Classification. EMINES 2023.

Exercise 1 (Excess of risk for a finite class of classifiers) 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 $\mathcal{X} \times \{-1, 1\}$ where $\mathcal{X}$ is a given state space. One aim of supervised classification is to define a function $h : \mathcal{X} \to \{-1, 1\}$, called *classifier*, such that $h(X)$ is the best prediction of Y in a given context. For instance, the probability of misclassification of h is

$$L_{\mathrm{miss}}(h) = \mathbb{P}\left(Y \neq h(X)\right) .$$

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

1. Prove that the classifier $h_\star$, defined for all $x \in \mathcal{X}$, by

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

   is such that

   $$h_\star \in \underset{h:\mathcal{X}\to\{-1,1\}}{\mathrm{argmin}}\ L_{\mathrm{miss}}(h)\,.$$

   Note that this optimal classifier corresponds to $h_\star = 1$ if $\mathbb{P}(Y = 1|X) \geq \mathbb{P}(Y = -1|X)$ and $h_\star = -1$ if $\mathbb{P}(Y = -1|X) > \mathbb{P}(Y = 1|X)$, which is sometimes more intuitive...

2. Recall the mixture of Gaussians example (in Day 1 exercise on Classification). Consider a random variable Y which takes the value 1 with probability π_1 and -1 with probability π_{-1}. If $Y = 1$, we have $X|_{Y=1} \sim N(\mu_1, \Sigma)$ and if $Y = -1$, $X|_{Y=-1} \sim N(\mu_{-1}, \Sigma)$. Assume we observe $(X_1, \ldots, X_n)$. Use the previous question and the EM algorithm to classify which of these X_i have been obtained from $Y_i = 1$ and which have been obtained by $Y_i = -1$.

3. In practice, the minimization of L_{miss} 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 L_{miss} cannot be computed nor minimized, it is instead estimated by the empirical classification risk defined as

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

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

   $$\widehat{h}^n_{\mathcal{H}} \in \underset{h\in\mathcal{H}}{\mathrm{argmin}}\ \widehat{L}^n_{\mathrm{miss}}(h)\,.$$

   Prove that for all set $\mathcal{H}$ of classifiers and all $n \geqslant 1$,

   $$L_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) - \inf_{h\in\mathcal{H}} L_{\mathrm{miss}}(h) \leqslant 2\sup_{h\in\mathcal{H}} \left|\widehat{L}^n_{\mathrm{miss}}(h) - L_{\mathrm{miss}}(h)\right| .$$

4. Using Hoeffding's inequality, prove that when $\mathcal{H} = \{h_1, \ldots, h_M\}$ for a given $M \geqslant 1$, then, for all $\delta > 0$,

   $$\mathbb{P}\left(L_{\mathrm{miss}}(\widehat{h}^n_{\mathcal{H}}) \leqslant \min_{1\leqslant j\leqslant M} L_{\mathrm{miss}}(h_j) + \sqrt{\frac{2}{n}\log\left(\frac{2M}{\delta}\right)}\right) \geqslant 1 - \delta\,.$$

Recall that the Hoeffing's inequality yields: Let $(X_i)_{1\leqslant i\leqslant n}$ be n independent random variables such that for all $1 \leqslant i \leqslant n$, $\mathbb{P}(a_i \leqslant X_i \leqslant b_i) = 1$ where a_i, b_i are real numbers such that $a_i < b_i$. The aim of this exercise is to prove the following inequality. For all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbb{E}\left[X_i\right]\right| > t\right) \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$