PC1. Foundations of Machine Learning. (MDC_51006_EP - 2025-2026)

Exercise 1 (Bayes Classifier) Prove that the optimal binary Bayes Classifier for the $0-1$ loss is given by

$$f^\star(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X) \geq \mathbb{P}(Y=-1|X) \\ -1 & \text{otherwise} \end{cases}$$

Exercise 2 (Training error optimism) **Assume that $\big((x_i, y_i)\big)_i$ is a sequence of i.i.d. random vectors**. Consider a parametric model f_β fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat\beta \in \mathbf{R}^d$ be the least squares estimate, i.e.

$$\hat\beta = \operatorname*{argmin}_\beta \frac{1}{N} \sum_{i=1}^N (y_i - f_\beta(x_i))^2$$

Suppose we have some test data $(\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_M, \widetilde{y}_M)$ drawn at random from the same population $\big((x_i, y_i)\big)_i$ as the training data.

1. If we define

$$R_{train}(\beta) = \frac{1}{N} \sum_{i=1}^N \left(y_i - f_\beta(x_i)\right)^2$$

and

$$R_{test}(\beta) = \frac{1}{M} \sum_{i=1}^M \left(\widetilde{y}_i - f_\beta(\widetilde{x}_i)\right)^2$$

Prove that

$$\mathbb{E}[R_{train}(\hat\beta)] \leq \mathbb{E}[R_{test}(\hat\beta)]$$

where the expectations are over all that is random in each expression.

2. Can we replace $R_{test}(\beta)$ by the risk $R(\beta) = \mathbb{E}\left[(\widetilde{y} - f_\beta(\widetilde{x}))^2\right]$ where $(\widetilde{x}, \widetilde{y})$ follows the population law?

3. Let $\beta^\star$ be the minimizer of the risk $R(\beta)$, prove that

$$\mathbb{E}\Big[R_{train}(\hat\beta)\Big] \leq R(\beta^\star) \leq \mathbb{E}\Big[R(\hat\beta)\Big]$$

Exercise 3 (Testing Error) 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

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

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.

In Exercise 1, we have shown that $h_\star$, 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

$$h_\star = \operatorname*{argmin}_{h : \mathcal{X} \to \{-1,1\}} L_{\text{miss}}(h) \ .$$

1. 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

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

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

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

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

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

2. Using Hoeffding's inequality, stating that if $X_i \in [a_i, b_i]$ then for all $t > 0$,

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

prove that when $\mathcal{H} = \{h_1, \ldots, h_M\}$ for a given $M \geqslant 1$, then, for all $\delta > 0$,

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