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

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

Solution.

$$\begin{aligned}\mathbb{E}\left[\ell^{0-1}(Y, f(X))\right] &= \mathbb{E}\left[\ell^{0-1}(Y, f(X))|Y=1, X\right]\mathbb{P}(Y=1|X) + \mathbb{E}\left[\ell^{0-1}(Y, f(X))|Y=-1, X\right]\mathbb{P}(Y=-1|X) \\ &= \mathbf{1}_{f(X)=-1}\mathbb{P}(Y=1|X) + \mathbf{1}_{f(X)=1}\mathbb{P}(Y=-1|X)\end{aligned}$$

which is indeed minimized by $f^\star$. □

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.

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

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.

Solution.

Let $\beta^\star = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\ \mathbb{E}[R_{train}(\beta)] = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\ \mathbb{E}\left[\left(y_0 - f_\beta(x_0)\right)^2\right]$. Note that $\beta^\star$ is a vector of $\mathbb{R}^d$ that is not random.

By construction, $R_{train}(\widehat\beta) \leq R_{train}(\beta^\star)$ and thus

$$\mathbb{E}\left[R_{train}(\widehat\beta)\right] \leq \mathbb{E}[R_{train}(\beta^\star)]\,. \tag{1}$$

Note that there is no contradiction with $\beta^\star = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\ \mathbb{E}[R_{train}(\beta)]$ because $\widehat\beta$ is a random vector that depends on and could be different for each $(x_1, y_1), \ldots, (x_N, y_N)$.

Denote $g(\beta) = \mathbb{E}[R_{test}(\beta)]$ for $\beta \in \mathbb{R}^d$. Since we also have $\mathbb{E}[R_{test}(\beta)] = \mathbb{E}\left[\left(y_0 - f_\beta(x_0)\right)^2\right] = \mathbb{E}[R_{train}(\beta)]$ for any $\beta \in \mathbb{R}^d$, we deduce that $\beta^\star = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\ \mathbb{E}[R_{test}(\beta)] = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\ g(\beta)$. But $\widehat\beta \in \sigma\big((x_1, y_1), \ldots, (x_N, y_N)\big)$ and is independant from $(\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_M, \widetilde{y}_M)$.

Therefore, for any $\beta \in \mathbb{R}$, we also have $g(\beta) = \mathbb{E}\left[R_{test}(\beta) \mid \sigma\big((x_1, y_1), \ldots, (x_N, y_N)\big)\right]$. Hence $g(\beta^\star) \leq \mathbb{E}\left[R_{test}(\beta) \mid \sigma\big((x_1, y_1), \ldots, (x_N, y_N)\big)\right]$ for any $\beta \in \mathbb{R}^d$.

Since $g(\widehat\beta) = \mathbb{E}\left[R_{test}(\widehat\beta) \mid \sigma\big((x_1, y_1), \ldots, (x_N, y_N)\big)\right]$, we have $g(\beta^\star) \leq g(\widehat\beta)$. Finally this induces $g(\beta^\star) = \mathbb{E}[g(\beta^\star)] \leq \mathbb{E}\left[g(\widehat\beta)\right]$ and this leads to

$$\mathbb{E}[R_{train}(\beta^\star)] = g(\beta^\star) = \mathbb{E}[R_{test}(\beta^\star)] \leq \mathbb{E}\left[R_{test}(\widehat\beta)\right] = \mathbb{E}\left[g(\widehat\beta)\right]. \tag{2}$$

As a consequence, we obtain

$$\mathbb{E}\left[R_{train}(\widehat\beta)\right] \leq \mathbb{E}[R_{train}(\beta^\star)] = \mathbb{E}[R_{test}(\beta^\star)] \leq \mathbb{E}\left[R_{test}(\widehat\beta)\right].$$

□

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?
Solution.

From the assumption, we establish for any $\beta \in \mathbb{R}^d$, $g(\beta) = R(\beta) = \mathbb{E}[R_{test}(\beta)] = \mathbb{E}\left[(\widetilde{y} - f_\beta(\widetilde{x}))^2\right] = \mathbb{E}\left[(y_1 - f_\beta(x_1))^2\right] = \mathbb{E}[R_{train}(\beta)]$. □

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

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

Solution.

From (2), $g(\beta^*) = R(\beta^*) \leq \mathbb{E}\left[g(\widehat{\beta})\right]$ and (1), $\mathbb{E}\left[R_{train}(\widehat{\beta})\right] \leq R(\beta^\star) = g(\beta^*)$, and this achieves the proof. □

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

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

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

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

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

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

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

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

Solution.

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

$$\begin{aligned} L_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) - \inf_{h \in \mathcal{H}} L_{\text{miss}}(h) &= L_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) - \widehat{L}^n_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) + \widehat{L}^n_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) - \inf_{h \in \mathcal{H}} L_{\text{miss}}(h) \ , \\ &\leqslant L_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) - \widehat{L}^n_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) + \widehat{L}^n_{\text{miss}}(h) - \inf_{h \in \mathcal{H}} L_{\text{miss}}(h) \ . \end{aligned}$$

For all $\varepsilon > 0$ there exists $h_\varepsilon \in \mathcal{H}$ such that $L_{\text{miss}}(h_\varepsilon) < \inf_{h \in \mathcal{H}} \ L_{\text{miss}}(h) + \varepsilon$ so that

$$\begin{aligned} L_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) - \inf_{h \in \mathcal{H}} L_{\text{miss}}(h) &\leqslant L_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) - \widehat{L}^n_{\text{miss}}(\widehat{h}^n_{\mathcal{H}}) + \widehat{L}^n_{\text{miss}}(h_\varepsilon) - L_{\text{miss}}(h_\varepsilon) + \varepsilon \ , \\ &\leqslant 2 \sup_{h \in \mathcal{H}} \left| \widehat{L}^n_{\text{miss}}(h) - L_{\text{miss}}(h) \right| + \varepsilon \ , \end{aligned}$$

which concludes the proof. □

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

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right| > t\right) \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) , \tag{3}$$

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

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

Solution.

By the previous question, for all $u > 0$,

$$\mathbb{P}\left(L_{\mathrm{miss}}(\widehat{h}_{\mathcal{H}}^n) > \min_{1 \leqslant j \leqslant M} L_{\mathrm{miss}}(h_j) + u\right) \leqslant \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left|\widehat{L}_{\mathrm{miss}}^n(h) - L_{\mathrm{miss}}(h)\right| > \frac{u}{2}\right) \leqslant \sum_{j=1}^{M} \mathbb{P}\left(\left|\widehat{L}_{\mathrm{miss}}^n(h_j) - L_{\mathrm{miss}}(h_j)\right| > \frac{u}{2}\right) .$$

By Hoeffding's inequality,

$$\mathbb{P}\left(L_{\mathrm{miss}}(\widehat{h}_{\mathcal{H}}^n) > \min_{1 \leqslant j \leqslant M} L_{\mathrm{miss}}(h_j) + u\right) \leqslant 2M\mathrm{e}^{-nu^2/2} ,$$

which concludes the proof by choosing

$$u = \sqrt{\frac{2}{n}\log\left(\frac{2M}{\delta}\right)} .$$

$\square$