PC1. Ecole Polytechnique. MAP 569. Machine Learning II.

Exercise 1 (**Hoeffding's inequality**) 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 $t > 0$,

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

1. Assume that $\mathbb{E}[X_i] = 0$ for all $1 \leqslant i \leqslant n$. Prove that it is enough to prove that for all $t > 0$,

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

2. Prove that for all $s, t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leqslant \mathrm{e}^{-st}\prod_{i=1}^n \mathbb{E}\left[\mathrm{e}^{sX_i}\right].$$

3. Define for all $1 \leqslant i \leqslant n$, $\phi_i : s \mapsto \log\left(\mathbb{E}\left[\mathrm{e}^{sX_i}\right]\right)$. Prove that for all $s > 0$,

$$\phi_i''(s) \leqslant \left(\frac{b_i - a_i}{2}\right)^2.$$

4. Prove that this upper bound implies for all $s, t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leqslant \mathrm{e}^{-st}\mathrm{e}^{s^2\sum_{i=1}^n \frac{(b_i-a_i)^2}{8}}$$

   and conclude.

Exercise 2 (**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

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

2. 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 boilds down to solving

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

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

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

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

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

Exercise 3 (Cross-validation) Consider the training data set: $\mathcal{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$ where $\mathbf{X}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. Assume that we construct the regressor function $\hat{f}$ by linear regression: we first define

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{j=1}^n (Y_j - \mathbf{X}_j^T\boldsymbol{\beta})^2$$

and we set $\hat{\mathbf{Y}} = \begin{pmatrix}\hat{Y}_1\\ \vdots \\ \hat{Y}_n\end{pmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}}$ where $\mathbf{X}$ is the $n\times p$ matrix, $\mathbf{X} = \begin{pmatrix}\mathbf{X}_1^T\\ \vdots \\ \mathbf{X}_n^T\end{pmatrix}$ and $\hat{Y}_i = \hat{f}(\mathbf{X}_i) = \mathbf{X}_i^T\hat{\boldsymbol{\beta}}$ for every $i \in \{1, \ldots, n\}$.

1. Assume $\mathrm{rank}(\mathbf{X}) = p$. Prove that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ where $\mathbf{Y} = \begin{pmatrix}Y_1\\ \vdots \\ Y_n\end{pmatrix}$.

2. For every $i \in \{1, \ldots, n\}$, we leave the i-th data out of the training set, that is, we define

$$\hat{\boldsymbol{\beta}}_{-i} = \operatorname{argmin} \sum_{j=1, j\neq i}^n (Y_j - \mathbf{X}_j^T\boldsymbol{\beta})^2$$

and we set $\hat{Y}_{-i} = \hat{f}^{-i}(\mathbf{X}_i) = \mathbf{X}_i^T\hat{\boldsymbol{\beta}}_{-i}$. Define the vector $\tilde{\mathbf{Y}} = \begin{pmatrix}\tilde{Y}_1\\ \vdots \\ \tilde{Y}_n\end{pmatrix}$ where $\tilde{Y}_k = Y_k$ for $k \neq i$ and $\tilde{Y}_i = \hat{Y}_{-i}$. Show that $\hat{\boldsymbol{\beta}}_{-i}$ is obtained by linear regression of the vector $\tilde{\mathbf{Y}}$ with respect to $\mathbf{X}$. Deduce the expression of $\hat{\boldsymbol{\beta}}_{-i}$ in terms of $\tilde{\mathbf{Y}}$ and $\mathbf{X}$.

3. Defining the hat matrix $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = [H_{k\ell}]_{1\leq k,\ell\leq n}$, deduce that

$$\hat{Y}_{-i} = \hat{Y}_i - H_{ii}Y_i + H_{ii}\hat{Y}_{-i}.$$

4. Show that the Leave-One-Out cross-validation error is:

$$\mathcal{R}_n^{LOO}(\hat{f}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_{-i})^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2.$$