Introduction to Statistical Learning

Final exam (3 pages)

Duration : 2h00 - Lecture notes allowed

Exercise 1 - Consider a learning problem with objective (to be minimized) $L(g) = \mathbb{E}\big(\ell(Z,g)\big)$ where Z is the observation vector with distribution P, g is the decision rule and ℓ is a positive loss function. Now let $Z_1,\dots,Z_n$ be an IID sample from the distribution P and $\widehat{L}_n(g) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i,g)$ the empirical risk of the rule g.

We introduce the following elements :

— $\mathcal{G}_1$, $\mathcal{G}_2$, ... is a sequence of function classes,

— g_1^*, g_2^*, ... is the corresponding sequence of optimal decision rules in the sense that $g_k^* = \arg\min_{\mathcal{G}_k} L(g)$ for $k \geq 1$,

— $\widehat{g}_n^{(1)}, \widehat{g}_n^{(2)}, \dots$ is the corresponding sequence of ERM for the empirical risk $\widehat{L}_n$,

— $\widehat{C}_1$, $\widehat{C}_2$, ... is a sequence of nonnegative random variables.

We now define the following strategy with output :

$$\widehat{g}_n = \widehat{g}_n^{(\widehat{k})}$$

where

$$\widehat{k} = \arg\min_{k\geq 1}\big(\widehat{L}_n\big(\widehat{g}_n^{(k)}\big) + \widehat{C}_k\big) \ .$$

1. Explain briefly the rationale behind this strategy.
2. We assume that there exists some $\gamma > 0$ such that, for any $k \geq 1$ and $n \geq 1$, the random variable $\widehat{C}_k$ satisfies the following inequalities :

$$\mathbb{P}(\widehat{C}_k \leq (L - \widehat{L}_n)(\widehat{g}_n^{(k)})) \leq \frac{\gamma}{n^2k^2} \ ,$$

and

$$\mathbb{P}(\widehat{C}_k \leq (\widehat{L}_n - L)(g_k^*)) \leq \frac{\gamma}{n^2k^2} \ .$$

Show that, in this case, there is some $\delta(\gamma,n) > 0$ such that, with probability at least $(1-\delta(\gamma,n))$, the following inequality holds :

$$L(\widehat{g}) - L^* \leq \inf_{k\geq 1}\big(L_k^* - L^* + 2\widehat{C}_k\big) \ .$$

where $L_k^* = L(g_k^*) = \inf_{g\in\mathcal{G}_k} L(g)$ denotes the optimal error in $\mathcal{G}_k$ and $L^* = \inf L$.

3. Give some examples of machine learning algorithms whose behavior could be understood with this framework. Please provide short but precise answers.

Exercise 2 - Consider the setup of preference learning where we observe an IID sample of triples $(X_1, X_1', Y_1), \ldots, (X_n, X_n', Y_n)$. The probabilistic model assumes that, for each i, the triple (X_i, X_i', Y_i) is such that X_i, X_i' are IID random vectors over $\mathbb{R}^d$ and Y_i is a random variable over $\{-1, 0, +1\}$. We define the ranking error of a preference rule $g : \mathbb{R}^d \to \{-1, 0, +1\}$ as :

$$L(g) = \mathbb{P}\{Y \neq 0,\ Y \cdot (g(X') - g(X)) \leq 0\}$$

ans the empirical margin ranking error as :

$$\widehat{L}_{n,\rho}(g) = \frac{1}{n} \sum_{i=1}^{n} m_\rho(Y_i \cdot (g(X_i') - g(X_i)))$$

where the margin loss is defined, for any $\rho > 0$, by

$$m_\rho(t) = \mathbb{I}\{t \leq 0\} + \mathbb{I}\{0 \leq t \leq \rho\} \left(1 - \frac{t}{\rho}\right) .$$

Now consider a class $\mathcal{G}$ of preference rules and define :

$$\widetilde{\mathcal{G}} = \{(x, x', y) \mapsto y(g(x') - g(x))\ :\ g \in \mathcal{G}\} .$$

1. Provide an upper bound of the empirical Rademacher average of $\widetilde{\mathcal{G}}$ in terms of the empirical Rademacher average of $\mathcal{G}$.
2. Show that m_ρ is Lipschitz and provide its Lipschitz constant.
3. Which inequality relates the empirical Rademacher average of the loss class $m_\rho \circ \widetilde{\mathcal{G}}$ to the empirical Rademacher average of $\widetilde{\mathcal{G}}$? Provide a proof of this inequality.
4. Show that, for any $\delta \in (0, 1)$, we have, with probability at least $1 - \delta$: for any $g \in \mathcal{G}$

$$\mathbb{E}(m_\rho(y(g(x') - g(x)))) \leq \widehat{L}_{n,\rho}(g) + c_1 \hat{R}_n(m_\rho \circ \tilde{\mathcal{G}}) + c_2(n, \delta)$$

   for some c_1 and $c_2(n, \delta)$ that will have to be given explicitly.
5. Deduce from the previous question a margin error bound for $L(g)$ that holds with large probability for any $g \in \mathcal{G}$ and which involves the empirical ranking error of g over the sample and the complexity of $\mathcal{G}$.
6. Specify the previous result to the case of a linear class of functions where $\mathcal{G} = \mathcal{F}_M = \{x \mapsto w^T x\ :\ w \in \mathbb{R}^d, \|w\|_2 \leq M\}$.
7. What kind of algorithm can be justified by the inequalities obtained in the two previous questions.
8. Propose an algorithm based on convex risk minimization for the preference learning problem. Formulate it precisely in pseudocode and explain what theoretical justification could be provided.

Exercise 3 - Consider the setup of supervised binary classification with the usual notations (labels are in $\{-1,+1\}$). We consider a convex risk minimization strategy to derive a soft classifier (real-valued decision rule) with an exponential loss function based on the following functional : for any $h : \mathbb{R}^d \to \mathbb{R}$,

$$\widehat{A}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-Y_i \cdot h(X_i)\right)$$

1. Justify briefly the use of such a functional for the supervised binary classification problem.
2. Consider convex weights $\pi(i)$ over the sample points : for any $i = 1, \ldots, n$, $\pi(i) \geq 0$ and $\sum_{i=1}^{n} \pi(i) = 1$. We introduce the following functional :
   for any binary $\{-1,1\}$-valued classifier g,
   $$\epsilon(g) = \sum_{i=1}^{n} \pi(i)\mathbb{I}\{Y_i \cdot g(X_i) = -1\}$$
   Provide an expression of $\pi(i)$ such that : for any fixed h, for $\alpha \in \mathbb{R}$, minimizing
   $$g \mapsto \left.\frac{\partial A_n(h+\alpha g)}{\partial \alpha}\right|_{\alpha=0}$$
   is equivalent to minimizing $\epsilon(g)$.
3. We propose to build decision rules h of the form $h_T = \sum_{t=1}^{T} \alpha_t g_t$ where the α_t's are real-valued coefficients and g_t's are simple classifiers taking their values in $\{-1,1\}$. Propose an algorithm relying on an iterative principle to determine the updates of (α_t, g_t).
4. Give the explicit expression of α_t at every iteration of the algorithm. *Hint :* We may consider the zero of the function $\alpha \mapsto \frac{\partial \widehat{A}_n(h_{t-1}+\alpha g_t)}{\partial \alpha}$.
5. What are the parameters of the algorithm and how to calibrate them ?