Machine Learning - Exam

December 16th, 2019

Exercise 1 is on 4 points. Exercise 2 is on 3 points. Exercice 3 is on 16 points. All exercises are independent. In exercise 3, if you do not find the answer to a question, you may admit the corresponding result in order to answer to the following questions. The maximal mark is 20 points. Take great care of the redaction: it must be clear and precise.

1. **PAC Learnable classes**
   Let d be a positive integer, and let $D(0,r) = \left\{x \in \mathbb{R}^d : \|x\| \leq r\right\}$ denote the disk of center 0 and radius r. We consider the hypothesis class $\mathcal{H} = \left\{\mathbb{1}_{D(0,r)} : r > 0\right\}$. Give two proofs that $\mathcal{H}$ is PAC-learnable (assuming realizability):

   - a direct proof, showing that the sample complexity is bounded by $1 + \log(1/\delta)/\epsilon$;
   - and a proof involving the fundamental theorem of PAC learning theory.

2. **0-1 loss and local minima.**
   We consider a binary classification task with $\mathcal{X} = \mathbb{R}^2$. For the value m and the hypothesis class $\mathcal{H} = \left\{h_w : w \in \mathbb{R}^2\right\}$ of your choice, construct a training sample $S = \big((X_1, Y_1), \ldots, (X_m, Y_m)\big) \in \big(\mathcal{X} \times \{-1, +1\}\big)^m$ such that there exists $w \in \mathbb{R}^2$ and $\epsilon > 0$ such that

   - for every $w' \in \mathbb{R}^2$ such that $\|w' - w\| \leq \epsilon, L_S(w) \leq L_S(w')$,
   - there exists $w^* \in \mathbb{R}^2$ such that $L_S(w^*) < L_S(w)$,

   where $L_S(w) = \sum_{k=1}^m \mathbb{1}\left\{h_w(X_k) \neq Y_k\right\}$ is the training error of hypothesis h_w.

3. **Problem**

Preliminaries.

Let X be a random variable such that $\mathbb{P}(0 \leq X \leq 1) = 1$, let $\mu = \mathbb{E}[X]$ and let $\phi : \lambda \mapsto \log \mathbb{E}\big[\exp(\lambda X)\big]$.

1. Show that ϕ is defined and infinitely differentiable on $\mathbb{R}$.
2. Show that $\phi(0) = 0$.
3. Show that $\phi'(0) = \mu$.
4. Show that for all $\lambda \in \mathbb{R}, \phi''(\lambda) \leq 1/4$.
5. Show Hoeffding's lemma: $\phi(\lambda) \leq \mu\lambda + \lambda^2/8$.
6. Show that Hoeffding's lemma entails Hoeffding's inequality: if $X_1, \dots, X_n$ are independent variables with the same distribution as X, then for all $\epsilon > 0$

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} > \mu + \epsilon\right) \leq \exp\big(-2n\epsilon^2\big) .$$

Prediction with expert advice.

We consider a setting where, at each round $t \in \mathbb{N}_+$, a value $y_t \in \mathcal{Y}$ is observed, where $\mathcal{Y}$ is an arbitrary set. The goal of the learner is to provide a prediction $\hat{p}_t \in \mathcal{X}$, where $\mathcal{X}$ is a convex set. The accuracy of a prediction is measured by a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ such that $\ell(\cdot, y)$ is convex for every $y \in \mathcal{Y}$.

The prediction $\hat{p}_t$ is allowed to depend on the advice of N "experts", which provide at time t the predictions $f_{1,t}, \dots, f_{N,t} \in \mathcal{X}$. More precisely, the prediction $\hat{p}_t$ must be a function of the predictions given so far $\big\{f_{j,s} : 1 \leq j \leq N, 1 \leq s \leq t\big\}$ and of the past observations $\big\{y_s : 1 \leq s < t\big\}$.

The *cumulated loss* of the learner at horizon $n \in \mathbb{N}_+$ is defined as

$$\hat{L}_n = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) ,$$

while the cumulated loss of expert $j \in \{1, \dots, n\}$ is defined as $L_{j,n} = \sum_{t=1}^{n} \ell(f_{j,t}, y_t)$.

The goal of the learner is to do almost as well as the *best expert in hindsight*: defining the learner's *regret* as

$$R_n = \hat{L}_n - \min\big\{L_{1,n}, \dots, L_{N,n}\big\} ,$$

one wishes to find a strategy such that R_n grows sub-linearly with n.

7. In this question only, we assume that $\mathcal{Y} = \{0, 1\}$, that the $(y_t)_t$ are independent random variables with Bernoulli distribution of parameter $\mu \in [0, 1]$, that $\mathcal{X} = [0, 1]$, $N = 3$ and that for each $j \in \{1, 2, 3\}$ and for all $t \geq 1$, $f_{j,t} = (j - 1)/2$. Propose a strategy such that R_n/n goes to 0 almost surely. Justify your answer.
8. In this question only, we assume that, for each expert $j \in \{1, \dots, N\}$, the sequence of losses $\big(\ell(f_{j,t}, y_t)\big)_t$ are independent and identically distributed. In that case, propose a strategy such that R_n/n goes almost-surely to 0 as $n \to \infty$. Justify your answer.
9. In this question, and in all the following, we no longer assume that the expert's losses obey any assumption; we want to find a strategy such that $R_n = o(n)$ for *every* sequence $(y_1, y_2, \dots)$. Is it the case of the strategy that you proposed in the previous question?

The Exponential Weights algorithm.

The Exponential Weights strategy of parameter $\eta > 0$ is defined as follows:

$$\hat{p}_t = \sum_{j=1}^{n} \frac{w_{j,t}}{W_t}\, f_{j,t} ,$$

where for all $j \in \{1, \ldots, N\}, w_{j,1} = 1$, $W_1 = N$ and for $t \geq 2$:

$$w_{j,t} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_{j,s}, y_s)\right) \quad \text{and } W_t = \sum_{j=1}^{N} w_{j,t} \ .$$

For simplicity, for all $t \in \{1, \ldots, n\}$ and all $j \in \{1, \ldots, N\}$ we denote $\alpha_{j,t} = \frac{w_{j,t}}{W_t}$ and $\ell_t(j) = \ell(f_{j,t}, y_t)$.

10. Show that

$$R_n \leq \sum_{t=1}^{n} \sum_{j=1}^{N} \alpha_{j,t} \ell_t(j) - \min_{1 \leq j \leq N} \sum_{t=1}^{n} \ell_t(j) \ .$$

11. Show that for all $j \in \{1, \ldots, N\}$, $W_{n+1} \geq \exp(-\eta L_{j,n})$ and hence that

$$\log \frac{W_{n+1}}{W_1} \geq -\eta L_{j,n} - \log(N) \ .$$

12. For all $t \in \{1, \ldots, n\}$, show that

$$\log \frac{W_{t+1}}{W_t} = \log\left(\sum_{j=1}^{N} \alpha_{j,t} \exp\big(-\eta \ell_t(j)\big)\right) \leq -\eta \sum_{j=1}^{N} \alpha_{j,t} \ell_t(j) + \frac{\eta^2}{8} \ .$$

13. Conclude that $R_n \leq \frac{\log(N)}{\eta} + \frac{n\eta}{8}$.
14. What is the value of the parameter η that minimizes the previous bound?
15. In this last question only, we assume that the loss function has range $[a, b]$ (and not $[0, 1]$ as before). What regret bound can be obtained in that case?
16. Discuss the optimality of the previous bound.