M2R MAF - MATHEMATICS OF MACHINE LEARNING
FINAL EXAM - FEBRUARY 22ND, 2017

Duration: 3h
No documents are authorized.

Exercise 1

Let $(\mathcal{X}, \mathcal{F})$ be any measurable space, and denote by $\mathcal{M}_1$ the set of all probability distributions on $(\mathcal{X}, \mathcal{F})$. For all $P, Q \in \mathcal{M}_1$, we define the *Hellinger distance* between P and Q by

$$H(P,Q) = \left(\frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 \mathrm{d}\mu(x) \right)^{1/2}, \tag{1}$$

where $\mu \in \mathcal{M}_1$ is such that $P \ll \mu$ and $Q \ll \mu$ (we say that μ dominates P and Q), and where $f = \dfrac{\mathrm{d}P}{\mathrm{d}\mu}$ and $g = \dfrac{\mathrm{d}Q}{\mathrm{d}\mu}$.

1. Let $P, Q \in \mathcal{M}_1$. Explain why there always exists $\mu \in \mathcal{M}_1$ that dominates P and Q, and show that the integral in (1) does not depend on the choice of μ. (Therefore, $H(P,Q)$ is well defined.)
2. Prove that H is a distance (or metric) on the set $\mathcal{M}_1$.
3. Show that
$$H(P,Q)^2 = 1 - \int_{\mathcal{X}} \sqrt{f(x)g(x)}\, \mathrm{d}\mu(x)\,.$$
4. Let $\sigma > 0$ and $a, b \in \mathbb{R}$. Compute $H\big(\mathcal{N}(a,\sigma^2), \mathcal{N}(b,\sigma^2)\big)^2$ as well as
$$\lim_{b \to a} \frac{H\big(\mathcal{N}(a,\sigma^2), \mathcal{N}(b,\sigma^2)\big)^2}{(b-a)^2}\,.$$
5. Show that $H\big(P^{\otimes n}, Q^{\otimes n}\big)^2 \leqslant nH(P,Q)^2$ for all $n \geqslant 1$.

Exercise 2

Let $\mathcal{D} \subseteq \mathbb{R}^d$ be any nonempty convex subset of $\mathbb{R}^d$ (the prediction space) and $\mathcal{Y}$ be any nonempty set (the observation space). Let $a < b$ and $\ell : \mathcal{D} \times \mathcal{Y} \to [a, b]$ be any loss function which is convex in its first argument. In the sequel, $K \geqslant 2$ denotes the number of experts. We consider the following online learning protocol.

At each round $t \in \mathbb{N}^*$,

- the expert advice $a_t = \big(a_{1,t}, \dots, a_{K,t}\big) \in \mathcal{D}^K$ are revealed to the statistician;
- the statistician makes her own prediction $\widehat{a}_t \in \mathcal{D}$ using the $a_{i,t}$ but also the past data (a_s, y_s), $1 \leqslant s \leqslant t-1$;
- The statistician observes $y_t \in \mathcal{Y}$ and incurs the loss $\ell\big(\widehat{a}_t, y_t\big)$.

Let $\eta_1 \geqslant \eta_2 \geqslant \eta_3 \geqslant \ldots > 0$ be any nonincreasing sequence of positive parameters. We consider the EWA algorithm, which predicts $\widehat{a}_t = \sum_{i=1}^K p_{i,t} a_{i,t}$ with weights

$$p_{i,t} = \frac{\exp\Big(-\eta_t \sum_{s=1}^{t-1} \ell\big(a_{i,s}, y_s\big)\Big)}{\sum_{j=1}^K \exp\Big(-\eta_t \sum_{s=1}^{t-1} \ell\big(a_{j,s}, y_s\big)\Big)} \,, \quad 1 \leqslant i \leqslant K \,.$$

At time t, the parameter η_t may be chosen as a function of the past data (a_s, y_s), $1 \leqslant s \leqslant t-1$. Moreoever, at time $t = 1$, $p_1 = \big(1/K, \dots, 1/K\big)$ by convention.

The goal of this exercise is to derive an upper bound on the regret

$$\mathrm{Reg}_T = \sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \,.$$

6. We set $L_{i,0} = 0$ and $L_{i,t} = \sum_{s=1}^t \ell\big(a_{i,s}, y_s\big)$ for all $t \geqslant 1$ and $i \in \{1, \dots, K\}$. We also define $W_t = \frac{1}{K} \sum_{i=1}^K e^{-\eta_t L_{i,t-1}}$ and $W'_{t+1} = \frac{1}{K} \sum_{i=1}^K e^{-\eta_t L_{i,t}}$ for all $t \geqslant 1$. Prove that
$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \geqslant - \min_{1 \leqslant i \leqslant K} L_{i,T} - \frac{\ln K}{\eta_{T+1}} \,.$$

7. Show that $W_{t+1} \leqslant \big(W'_{t+1}\big)^{\eta_{t+1}/\eta_t}$ and then that
$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \leqslant - \sum_{t=1}^T \sum_{i=1}^K p_{i,t} \ell\big(a_{i,t}, y_t\big) + \frac{(b-a)^2}{8} \sum_{t=1}^T \eta_t \,.$$

8. Prove that the EWA algorithm satisfies the following regret bound: for all $T \geqslant 1$ and all sequences of $a_t \in \mathcal{D}^K$ and $y_t \in \mathcal{Y}$,
$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) \leqslant \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) + \frac{\ln K}{\eta_{T+1}} + \frac{(b-a)^2}{8} \sum_{t=1}^T \eta_t \,.$$

9. Explain why the last inequality can be improved in order to imply that
$$\mathrm{Reg}_T \leqslant \frac{\ln K}{\eta_T} + \frac{(b-a)^2}{8} \sum_{t=1}^T \eta_t \,. \tag{2}$$

10. Show that the choice of $\eta_t = 2(b-a)^{-1}\sqrt{\ln(K)/t}$ leads to the regret bound $\mathrm{Reg}_T \leqslant (b-a)\sqrt{T \ln K}$. What is the advantage of taking a time-varying parameter η_t instead of a constant parameter η?

Exercise 3

Let $(X_i)_{1\leqslant i\leqslant n}$ be i.i.d. random variables with a density f^* belonging to the set $L^2([0,1])$ of square integrable functions on $[0,1]$. The goal of this exercise is to study an estimator of the density f^*. More precisely, we will analyze the performance of the estimator $\widehat{f}(x) = \sum_{k=0}^{\infty} \widehat{T}_k \phi_k(x)$ defined on the next page.

11. Cite another possible estimator of f^*, and give sufficient conditions for its consistency.

The scalar product of two functions $f, g \in L^2\big([0,1]\big)$ is denoted by $\langle f, g\rangle$. Let $\big\{\phi_k : k \in \mathbb{N}\big\}$ be the sequence of functions $\phi_k : [0,1] \to \mathbb{R}$ defined by $\phi_0(x) = 1$, and for all $k \in \mathbb{N}^*$ by

$$\phi_{2k-1}(x) = \sqrt{2}\sin(2\pi k x) \quad \text{and} \quad \phi_{2k}(x) = \sqrt{2}\cos(2\pi k x)\ .$$

We denote by $\ell^2(\mathbb{N})$ the set of all square summable sequences $(u_k)_{k\in\mathbb{N}}$. The usual scalar product of two sequences $u, v \in \ell^2(\mathbb{N})$ is denoted by $\langle u, v\rangle = \sum_{k\in\mathbb{N}} u_k v_k$.
Let $\theta_k^* = \langle f^*, \phi_k\rangle$, $k \in \mathbb{N}$, denote the Fourier coefficients of the unknown density function f^*. Furthermore, let

$$\widehat{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} \phi_k(X_i)$$

be the Fourier coefficients of the sample $\{X_1, \dots, X_n\}$.

Deviations. For every threshold $\lambda > 0$, defined $\mathcal{A}_k^\lambda = \big\{|\widehat{\theta}_k - \theta_k^*| \leqslant \lambda\big\}$ and $\mathcal{A}^\lambda = \big\{\max_{0\leqslant k\leqslant n-1} \big|\widehat{\theta}_k - \theta_k^*| \leqslant \lambda\big\}$.

12. Show that for all $k \in \mathbb{N}$, $\mathbb{E}\big[\widehat{\theta}_k\big] = \theta_k^*$.
13. Prove that for all $k \in \mathbb{N}$,

$$\mathbb{P}\big(\mathcal{A}_k^\lambda\big) \geqslant 1 - 2\exp\left(-\frac{n\lambda^2}{16}\right)\ .$$

14. Deduce that

$$\mathbb{P}\big(\mathcal{A}^\lambda\big) \geqslant 1 - 2n\exp\left(-\frac{n\lambda^2}{16}\right)\ .$$

15. For a given tolerance level $\delta > 0$, determine $\lambda > 0$ such that $\mathbb{P}(\mathcal{A}^\lambda) \geqslant 1 - \delta$.

Estimator. Let $\widehat{T} = \Big(\widehat{T}_k\Big)_{k\in\mathbb{N}}$ be the thresholded empirical Fourier coefficients:

$$\widehat{T}_k = \begin{cases} \widehat{\theta}_k & \text{if } |\widehat{\theta}_k| \geqslant 2\lambda \text{ and } k < n\ , \\ 0 & \text{otherwise}\ . \end{cases}$$

16. Prove that on the event $\mathcal{A}^\lambda$, for all $k \in \{0, \dots, n-1\}$,

$$\big(\widehat{T}_k - \theta_k^*\big)^2 \leqslant 9\min\big((\theta_k^*)^2, \lambda^2\big)\ .$$

17. Deduce that, on $\mathcal{A}^\lambda$,

$$\Big\|\widehat{T} - \theta^*\Big\|_2^2 := \sum_{k=0}^{\infty}\Big(\widehat{T}_k - \theta_k^*\Big)^2 \leqslant 9\min_{1\leqslant K\leqslant n-1}\left\{K\lambda^2 + \sum_{k\geqslant K}(\theta_k^*)^2\right\}\ .$$

Efficiency on Sobolev spaces. We now assume that f^* is continuously differentiable, and that it belongs to the Sobolev ball $\Sigma(1, L)$ defined for some $L > 0$ as:

$$\Sigma(1, L) := \left\{ g : [0, 1] \to \mathbb{R} : \int_0^1 g'(x)^2 \, dx \leqslant L, \; g(0) = g(1) \right\} .$$

We define the estimator $\widehat{f}$ of f^* as $\widehat{f}(x) = \sum_{k=0}^{\infty} \widehat{T}_k \phi_k(x)$.

18. [optional: you may simply assume this result.] Prove that the Fourier coefficients of f^* satisfy the inequality:

$$\sum_{k=0}^{\infty} k^2 (\theta_k^*)^2 \leqslant \frac{L}{4\pi^2} .$$

19. Prove that

$$\left\| \widehat{T} - \theta^* \right\|_2^2 \leqslant 9 \min_{1 \leqslant K \leqslant n-1} \left\{ K\lambda^2 + \frac{L}{4\pi^2 K^2} \right\} .$$

20. Find a constant $C > 0$ such that

$$\mathbb{P}\left(\|\widehat{f} - f^*\|_2^2 \leqslant C \left(\frac{\log(2n/\delta)}{n} \right)^{2/3} \right) \geqslant 1 - \delta .$$

21. Conclude.