Machine Learning Theory Exam

June 10, 2020

Question 1

Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a strictly increasing convex function that satisfies $\psi(0) = 0$. The $\psi-$ Orlicz norm of a random variable X is defined as

$$\|X\|_\psi := \inf \left\{ t > 0 \,\middle|\, \mathbb{E}\left[\psi\left(t^{-1}|X|\right)\right] \leq 1 \right\} \tag{1}$$

where $\|X\|_\psi$ is infinite if there is no finite t for which the expectation $\mathbb{E}\left[\psi\left(t^{-1}|X|\right)\right]$ exists. For the functions $u \mapsto u^q$ for some $q \in [1, \infty]$, then the Orlicz norm is simply the usual ℓ_q -norm $\|X\|_q = \left(\mathbb{E}\left[|X|^q\right]\right)^{1/q}$. Here, we consider the Orlicz norms $\|\cdot\|_{\psi_q}$ defined by the convex functions $\psi_q(u) = \exp\left(u^q\right) - 1$, for $q \geq 1$.
(1) If $\|X\|_{\psi_q} < +\infty$, show that there exist positive constants c_1, c_2 such that

$$\mathbb{P}[|X| > t] \leq c_1 \exp\left(-c_2 t^q\right) \qquad \text{for all } t > 0 \tag{2}$$

(2) Suppose that a random variable Z satisfies the tail bound (2). Show that $\|X\|_{\psi_q}$ is finite.

Question 2

Derive the Lagrange dual of the optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \phi\left(x_i\right) \\ \text{subject to} & Ax = b \end{array} \tag{3}$$

with variable $x \in \mathbf{R}^n$, where

$$\phi(u) = \frac{|u|}{c - |u|} = -1 + \frac{c}{c - |u|}, \quad \operatorname{dom} \phi = (-c, c) \tag{4}$$

c is a positive parameter.

Question 3

Let P be a distribution over (X, Y) pairs where $X \in \mathcal{X}$ and $Y \in \{+1, -1\}$ and let $\mathcal{H} \subset \mathcal{X} \to \{+1, -1\}$ be a finite hypothesis class and let ℓ denote the zero-one loss $\ell(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$. As usual let $R(h) = \mathbb{E}\ell(h(X), Y)$ denote the risk, and let $h^\star = \min_{h \in \mathcal{H}} R(h)$. Given n samples let $\hat{h}_n$ denote the empirical risk minimizer.

(1) Prove that with probability at least $1 - \delta$,

$$R\left(\hat{h}_n\right) - R\left(h^\star\right) \leq c_1 \sqrt{\frac{R\left(h^\star\right) \log(|\mathcal{H}|/\delta)}{n}} + c_2 \frac{\log(|\mathcal{H}|/\delta)}{n} \tag{5}$$

where c_1 and c_2 are constants.

(2) Given a family of hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2 \ldots, \subset \mathcal{H}_L$, of sizes $N_1 \leq N_2 \leq \ldots \leq N_L < \infty$, a loss function bounded on [0,1] and a sample of size n, design an algorithm that guarantees

$$R(\hat{h}) \leq \min_{i \in [L]} \min_{h^\star \in \mathcal{H}_i} \left\{ R\left(h^\star\right) + c_1 \sqrt{\frac{R\left(h^\star\right) \log\left(L N_i/\delta\right)}{n}} + c_2 \frac{\log\left(L N_i/\delta\right)}{n} \right\} \tag{6}$$

for $n \geq 2$. Your algorithm may use ERM (so need not be efficient) and your constants may vary.

You may find it useful to use the following (empirical) bernstein inequality.

Theorem 1 (Bernstein's inequality). Let $X_1, \ldots, X_n$ be iid real-valued random variables with mean zero and such that $|X_i| \leq M$ for all i. Then for all $t > 0$

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] + Mt/3}\right)$$

Theorem 2 (Empirical Berstein's Inequality) Let $X_1, \ldots, X_n$ be i.i.d. random variables from a distribution P supported on [0,1] and define the sample variance $V_n = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} \left(X_i - X_j\right)^2$. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$

$$\mathbb{E}X - \frac{1}{n} \sum_{i=1}^n X_i \leq \sqrt{\frac{2 V_n \log(2/\delta)}{n}} + \frac{7 \log(2/\delta)}{3(n-1)}$$

Question 4

Let $n \in \mathbb{N}^+$ and $(A_i)_{i=1}^m$ be a partition of $[n]$ so that $\cup_{i=1}^m A_i = [n]$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Suppose that $\delta \in (0, 1)$ and $X_1, X_2, \ldots, X_n$ is a sequence of independent random variables with mean μ and variance σ^2. The median-of-means estimator $\hat{\mu}_M$ of μ is the median of $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m$, where $\hat{\mu}_i = \sum_{t \in A_i} X_t / |A_i|$ is the mean of the data in the i th block.

(a) Show that if $m = \left\lfloor \min\left\{\frac{n}{2}, 8 \log\left(\frac{e^{1/8}}{\delta}\right)\right\}\right\rfloor$ and A_i are chosen as equally sized as possible, then

$$\mathbb{P}\left(\hat{\mu}_M + \sqrt{\frac{192\sigma^2}{n} \log\left(\frac{e^{1/8}}{\delta}\right)} \leq \mu\right) \leq \delta$$

Feel free to replace the constant 192 with any other positive constant.
(b) Use the median-of-means estimator to design an upper confidence bound algorithm such that for all $\nu \in \mathcal{E}_{\mathrm{V}}^{k}\left(\sigma^2\right)$

$$R_n \leq C \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{\sigma^2 \log(n)}{\Delta_i}\right)$$

where $C > 0$ is a universal constant. $\mathcal{E}_{\mathrm{V}}^{k}\left(\sigma^2\right)$ denotes the set of instances of k-arm bandits: $\{(P_i)_i : \mathbb{V}_{X\sim P_i}[X] \leq \sigma^2 \text{ for all } i\}$