Mathematical Tools in Machine Learning **(19.12.19)**

Solutions to Exercise Session 4

Instructor: Fadoua Balabdaoui *Assistant:* Loris Michel

Problem 11.1: Leave-one-out cross validation (LOOCV) can fail

Let S be an i.i.d. sample. Let h be the output of the described learning algorithm. Note that (independently of the identity of S), $L_{\mathcal{D}}(h) = 1/2$ (since h is a constant function).
Let us calculate the estimate $L_V(h)$. Assume that the parity of S is 1. Fix some fold $\{(x, y)\} \subseteq S$. We distinguish between two cases:

- The parity of $S \setminus \{x\}$ is 1. It follows that $y = 0$. When being trained using $S \setminus \{x\}$, the algorithm outputs the constant predictor $h(x) = 1$. Hence, the leave-one-out estimate using this fold is 1.
- The parity of $S \setminus \{x\}$ is 0. It follows that $y = 1$. When being trained using $S \setminus \{x\}$, the algorithm outputs the constant predictor $h(x) = 0$. Hence, the leave-one-out estimate using this fold is 1.

Averaging over the folds, the estimate of the error of h is 1. Consequently, the difference between the estimate and the true error is 1/2. The case in which the parity of S is 0 is analyzed analogously

Problem 12.2: Convexity-Lipschitzness-Smoothness

Convexity: Note that the function $g : \mathbb{R} \to \mathbb{R}$, defined by $g(a) = \log(1 + \exp(a))$ is convex. To see this, note that g^a is non-negative. The convexity of l (or more accurately, of $l(\cdot, z)$ for all z) follows now from Claim 12.4 (in the book).
Lipschitzness: The function $g(a) = \log(1 + \exp(a))$ is 1-Lipschitz, since $|g'(a)| = \frac{\exp(a)}{1+\exp(a)} = \frac{1}{\exp(-a)+1} \leq 1$. Hence by Claim 12.7 (in the book) l is B-Lipschitz.
Smoothness: We claim that $g(a) = \log(1 + \exp(a))$ is 1/4-smooth. To see this, note that

$$\begin{aligned} g''(a) &= \frac{\exp(-a)}{(\exp(-a)+1)^2} \\ &= \left(\exp(a)(\exp(-a)+1)^2\right)^{-1} \\ &= \frac{1}{2+\exp(a)+\exp(-a)} \\ \leq 1/4. \end{aligned}$$

Combine this with the mean value theorem, to conclude that g' is 1/4-Lipschitz. Using Claim 12.9 (in the book), we conclude that l is $B^2/4$-smooth.
Boundness: The norm of each hypothesis is bounded by B according to the assumptions. All in all, we conclude that the learning problem of linear regression is Convex-Smooth-Bounded with parameters $B^2/4$, B, and Convex-Lipschitz-Bounded with parameters B, B.

Problem 12.4 (optional): Turing Machines

a) Fix a Turing machine T. If T halts on the input 0, then for every $h \in [0, 1]$,

$$l(h, T) = \langle (h, 1 - h), (1, 0) \rangle.$$

If T halts on the input 1, then for every $h \in [0, 1]$,

$$l(h, T) = \langle (h, 1 - h), (0, 2) \rangle.$$

In both cases, l is linear, and hence convex over $\mathcal{H}$.

b) The idea is to reduce the halting problem to the learning problem . More accurately, the following decision problem can be easily reduced to the learning problem described in the question: Given a Turing machine M, does M halt given the input M? The proof that the halting problem is not decidable implies that this decision problem is not decidable as well. Hence, there is mo computable algorithm that learns the problem described in the question.

Problem 13.1: From bounded expected risk to agnostic PAC learning

We assume A is a (proper) algorithm that guarantees the following: If $m \geq m_{\mathcal{H}}(\cdot)$ then for every distribution $\mathcal{D}$ it holds that

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Since $A(S) \in \mathcal{H}$, the random variable $\theta = L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}$ is non-negative. Therefore, Markov's inequality implies that

$$\mathbb{P}[\theta \geq \mathbb{E}[\theta]/\delta] \leq \frac{\mathbb{E}[\theta]}{\mathbb{E}[\theta]/\delta} = \delta.$$

In other words, with probability of at least $1-\delta$ we have $\theta \leq \mathbb{E}[\theta]/\delta$. But, if $m \geq m_{\mathcal{H}}(\cdot, \delta)$ then we know that $\mathbb{E}[\theta] \leq \epsilon\delta$. This yields $\theta \leq \epsilon$, which concludes our proof.