Mathematical Tools in Machine Learning **(03.10.19)**

Solutions to Exercise Session 1

Instructor: Fadoua Balabdaoui *Assistant:* Loris Michel

Problem 2.2: Axis-aligned rectangles

(a) First, by definition of the learning algorithm A, every positive instance of the training set is correctly labeled. By minimality of the rectangle, all negative instances are as well correctly labeled. So A corresponds to ERM (empirical risk minimization).
(b) Let $\mathcal{D}$ be some fixed distribution over $\mathcal{X}$ and define R^* as in the hint. Let f be the hypothesis associated with R^* and S be a training set. We further denote by $R(S)$ the rectangle returned by the learning algorithm A ($A(S)$ the corresponding hypothesis). By definition of the learning algorithm $R(S) \subseteq R^*$ for every S. Thus,

$$L_{(\mathcal{D},f)}(R(S)) = \mathcal{D}(R^* \setminus R(S)).$$

Fix some $\epsilon \in (0,1)$ and define R_1, R_2, R_3 and R_4 as proposed in the hint. For each $i \in [4]$, define the event

$$F_i = \{S|_x : S|_x \cap R_i = \emptyset\}.$$

Thanks to union bound we obtain

$$\mathcal{D}^m(\{S : L_{(\mathcal{D},f)}(R(S)) > \epsilon\}) \leq \mathcal{D}^m(\cup_{i=1}^4 F_i) \leq \sum_{i=1}^4 \mathcal{D}^m(F_i).$$

Thus it is sufficient to ensure that $\mathcal{D}^m(F_i) \leq \delta/4$ for every i. Fix some $i \in [4]$. Then the probability that a sample is in F_i is the probability that all training instances do not fall into R_i, which is exactly $(1-\epsilon/4)^m$. Therefore,

$$\mathcal{D}^m(F_i) = (1-\epsilon/4)^m \leq \exp(-m\epsilon/4),$$

and hence

$$\mathcal{D}^m(\{S : L_{(\mathcal{D},f)}(R(S)) > \epsilon\}) \leq 4\exp(-m\epsilon/4).$$

Plugging in the assumption on m, we conclude our proof.
(c) The hypothesis class of axis aligned rectangles in $\mathbb{R}^d$ is defined as follows. Given real numbers $a_1 \leq b_1, \ldots, a_d \leq b_d$, define the classifier $h_{(a_1,b_1,\ldots,a_d,b_d)}$ by

$$h_{(a_1,b_1,\ldots,a_d,b_d)}(x_1,\ldots,x_d) = \begin{cases} 1 \text{ if } \forall i \in [d], a_i \leq x_i \leq b_i \\ 0 \text{ otherwise} \end{cases}$$

The class of all axis-aligned rectangles in $\mathbb{R}^d$ is defined as

$$\mathcal{H}^{rec} = \{h_{(a_1,b_1,\ldots,a_d,b_d)} : \forall i \in [d], a_i \leq b_i\}.$$

It can be see that the same algorithm proposed above is an ERM for this case as well. The sample complexity is analyzed similarly. The only difference is that instead of 4 strips, we have $2d$ strips (2 for each dimension). Thus, it suffices to draw a training set of size $\lceil \frac{2d\log(2d/\delta)}{\epsilon} \rceil$.

Remark: you can find `R` code on the course web-page (`Rcode1`) implementing the learner A that you can play around with.

Problem 3.1: Sample complexity

The proofs follow (almost) immediately from the definition. We will show that the sample complexity is monotonically decreasing in the accuracy parameter ϵ. The proof that the sample complexity is monotonically decreasing in the confidence parameter δ is analogous. Denote by $\mathcal{D}$ an unknown distribution over $\mathcal{X}$, and let $f \in \mathcal{H}$ be the target hypothesis. Denote by A an algorithm which learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. Fix some $\delta \in (0,1)$. Suppose that $0 < \epsilon_1 \le \epsilon_2 \le 1$. We need to show that $m_1 =^{def} m_{\mathcal{H}}(\epsilon_1, \delta) \ge m_{\mathcal{H}}(\epsilon_2, \delta) =^{def} m_2$. Given an i.i.d. training sequence of size $m \ge m_1$, we have that with probability at least $1-\delta$, A returns a hypothesis h such that

$$L_{\mathcal{D},f}(h) \le \epsilon_1 \le \epsilon_2.$$

By the minimality of m_2, we conclude that $m_2 \le m_1$.

Problem 3.2: Singletons

(b) Let $\epsilon \in (0,1)$, and fix the distribution $\mathcal{D}$ over $\mathcal{X}$. If the true hypothesis is h^-, then our algorithm returns a perfect hypothesis. Assume now that there exists a unique positive instance x_+. It is clear that if x_+ appears in the training sequence S, our algorithm returns a perfect hypothesis. Furthermore, if $\mathcal{D}|\{x_+\}| \le \epsilon$ then in any case, the returned hypothesis has a generalization error of at most ϵ (with probability 1). Thus, it is only left to bound the probability of the case in which $\mathcal{D}|\{x_+\}| > \epsilon$ but x_+ does not appear in S. Denote this event by F. Then

$$\mathbb{P}_{S|x\sim\mathcal{D}^m}[F] \le (1-\epsilon)^m \le e^{-m\epsilon}.$$

Hence $\mathcal{H}_{Singleton}$ is PAC learnable, and its sample complexity is bounded by

$$m_{\mathcal{H}}(\epsilon,\delta) \le \lceil \frac{\log(1/\delta)}{\epsilon} \rceil.$$

Problem 3.3: Concentric circles

Consider the ERM algorithm A which given a training sequence $S = ((x_i, y_i))_{i=1}^m$, returns the hypothesis $\hat{h}$ corresponding to the "tightest" circle which contains all the positive instances. Denote the radius of this hypothesis by $\hat{r}$. Assume realizability and let h^* be a circle with zero generalization error. Denote its radius by r^*.

Let $\epsilon, \delta \in (0,1)$. Let $\overline{r} \le r^*$ be a scalar s.t. $\mathcal{D}_{\mathcal{X}}(\{x : \overline{r} \le \|x\| \le r^*\}) = \epsilon$. Define $E = \{x \in \mathbb{R}^2 : \overline{r} \le \|x\| \le r^*\}$. The probability (over drawing S) that $L_{\mathcal{D}}(h_S) \ge \epsilon$ is bounded above by the probability that no point in S belongs to E. This probability of this event is bounded above by

$$(1-\epsilon)^m \le e^{-\epsilon m}.$$

The desired bound on the sample complexity follows by requiring that $e^{-\epsilon m} \le \delta$.

Problem 3.6: PAC/agnostic PAC

Suppose that $\mathcal{H}$ is agnostic PAC learnable, and let A be a learning algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. We show that $\mathcal{H}$ is PAC learnable using A.

Let $\mathcal{D}, f$ be an (unknown) distribution over $\mathcal{X}$, and the target function respectively. We may assume w.l.o.g. that $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \{0,1\}$, where the conditional probability of y given x is determined deterministically by f. Since we assume realizability, we have

$$\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0.$$

Let $\epsilon, \delta \in (0,1)$. Then, for every positive integer $m \ge m_{\mathcal{H}}(\epsilon,\delta)$, if we equip A with a training set S consisting of m i.i.d. instances which are labeled by f, then with probability at least $1-\delta$ (over the choice of $S|x$), it returns a hypothesis h with

$$\begin{aligned} L_{\mathcal{D}}(h) &\le \inf_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \\ &= 0 + \epsilon \\ &= \epsilon. \end{aligned}$$

Problem 3.7: The Bayes optimal predictor

Let $x \in \mathcal{X}$ and $\eta(x)$ be the conditional probability of a positive label at x. We have

$$\begin{aligned}
\mathbb{P}[h_{\mathcal{D}}(X) \neq Y|X = x] &= \mathbb{1}_{[\eta(x)>1/2]}\mathbb{P}[Y = 0|X = x] + \mathbb{1}_{[\eta(x)\leq 1/2]}\mathbb{P}[Y = 1|X = x] \\
&= \mathbb{1}_{[\eta(x)>1/2]}\eta(x) + \mathbb{1}_{[\eta(x)\leq 1/2]}(1 - \eta(x)) \\
&= \min(\eta(x), 1 - \eta(x)).
\end{aligned}$$

Let g be a classifier from $\mathcal{X}$ to $\{0, 1\}$ (can be non-deterministic). We have

$$\begin{aligned}
\mathbb{P}[h(X) \neq Y|X = x] &= \mathbb{P}[h(X) = 0|X = x]\mathbb{P}[Y = 1|X = x] + \mathbb{P}[h(X) = 1|X = x]\mathbb{P}[Y = 0|X = x] \\
&+ \mathbb{P}[h(X) = 0|X = x]\eta(x) + \mathbb{P}[h(X) = 1|X = x](1 - \eta(x)) \\
&\geq \mathbb{P}[h(X) = 0|X = x] \min(\eta(x), 1 - \eta(x)) + \mathbb{P}[h(X) = 1|X = x] \min(\eta(x), 1 - \eta(x)) \\
&= \min(\eta(x), 1 - \eta(x)).
\end{aligned}$$

The statement follows now from the fact that the above is true for every $x \in \mathcal{X}$.More formally, by the law of total expectation,

$$\begin{aligned}
L_{\mathcal{D}}(h_{\mathcal{D}}) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{[h_{\mathcal{D}}(x)\neq y]}] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}[\mathbb{1}_{[h_{\mathcal{D}}(x)\neq y]}|X = x]] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}[\min(\eta(x), 1 - \eta(x))] \\
&\leq \mathbb{E}_{x\sim\mathcal{D}_x}[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}[\mathbb{1}_{[h(x)\neq y]}|X = x]] \\
&= L_{\mathcal{D}}(h).
\end{aligned}$$

The statement could be also deduced from the following identity for the *excess risk* of any classifier $h : \mathcal{X} \to \{0, 1\}$,

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h) = \mathbb{E}_{x\sim\mathcal{D}}[|2\eta(x) - 1|\mathbb{1}_{h(x)\neq h_{\mathcal{D}}(x)}],$$

As done for $h_{\mathcal{D}}$, we can actually show that for any h,

$$L_{\mathcal{D}}(h) = \mathbb{E}[\mathbb{1}_{h(x)=0}\eta(x) + \mathbb{1}_{h(x)=1}(1 - \eta(x))],$$

which, using that $\mathbb{1}_{h(x)=1} = 1 - \mathbb{1}_{h(x)=0}$, is the same as

$$L_{\mathcal{D}}(h) = \mathbb{E}[\mathbb{1}_{h(x)=0}(2\eta(x) - 1) + 1 - \eta(x)].$$

So then in particular,

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h_{\mathcal{D}}) = \mathbb{E}[(\mathbb{1}_{h(x)=0} - \mathbb{1}_{h_{\mathcal{D}}(x)=0})(2\eta(x) - 1)].$$

Now remark that $(\mathbb{1}_{h(x)=0} - \mathbb{1}_{h_{\mathcal{D}}(x)=0})$ takes values in $\{-1, 0, 1\}$. It turns out that $\mathbb{1}_{h(x)=0} - \mathbb{1}_{h_{\mathcal{D}}(x)=0}$ is non-zero if and only if $h(x) \neq h_{\mathcal{D}}(x)$ and $\mathbb{1}_{h(x)=0} - \mathbb{1}_{h_{\mathcal{D}}(x)=0}$ has the same sign as $2\eta(x) - 1$, so we conclude that

$$\mathbb{E}[(\mathbb{1}_{h(x)=0} - \mathbb{1}_{h_{\mathcal{D}}(x)=0})(2\eta(x) - 1)] = \mathbb{E}[\mathbb{1}_{h(x)\neq h_{\mathcal{D}}(x)}|2\eta(x) - 1|].$$

Problem 4.1: Average losses

(a) Assume that for every $\epsilon, \delta \in (0, 1)$ and every distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, there exists $m(\epsilon, \delta) \in \mathbb{N}$ such that for every $m \geq m(\epsilon, \delta)$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))} > \epsilon] < \delta.$$

Let $\lambda > 0$. We need to show that there exists $m_0 \in \mathbb{N}$ such that for every $m \geq m_0$, $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))}] \leq \lambda$. Let $\epsilon = \min(1/2, \lambda/2)$. Set $m_0 = m_{\mathcal{H}}(\epsilon, \delta)$. For every $m \geq m_0$, since the loss is bounded above by 1, we have

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))}] &\leq \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))} > \lambda/2] \cdot 1 + \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))} \leq \lambda/2] \cdot \lambda/2 \\
&\leq \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))} > \epsilon] + \lambda/2 \\
&\leq \epsilon + \lambda/2 \\
&\leq \lambda/2 + \lambda/2 \\
&= \lambda
\end{aligned}$$

(b) Assume now that
$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))}] = 0.$$
Let $\epsilon, \delta \in (0,1)$. There exists some $m_0 \in \mathbb{N}$ such that for every $m \geq m_0$, $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))}] \leq \epsilon \cdot \delta$. by Markov's inequality,

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))} > \epsilon] &\leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}(A(S))}]}{\epsilon} \\
&\leq \frac{\epsilon\delta}{\epsilon} \\
&= \delta.
\end{aligned}$$