Learning axis-aligned rectangles

(In particular, PAC learnability of axis-aligned rectangles)

We consider the hypothesis class $\mathcal{H}$ of axis-aligned rectangles in $\mathcal{X} = \mathbb{R}^2$. Our label space is $\mathcal{Y} = \{0, 1\}$. A hypothesis $h \in \mathcal{H}$ is

$$h : \mathcal{X} \longrightarrow \mathcal{Y}, \quad \mathbf{x} \mapsto h_{b,t,r,s}(\mathbf{x}) = \begin{cases} 1 & \text{if } b \leq x \leq t \text{ and } r \leq y \leq s \\ 0 & \text{else} \end{cases}$$

with $\mathbf{x} = (x, y) \in \mathbb{R}^2$. For this class, $|\mathcal{H}| = \infty$.

We are given a sample

$$S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$$

where $\mathbf{x}_i \sim \mathcal{D}$ and labeled according to a hypothesis $f \in \mathcal{H}$, i.e., $y_i = f(\mathbf{x}_i)$. Assuming realizability, $\exists h \in \mathcal{H}$, s.t. $L_{\mathcal{D},f}(h) = 0$. We identify the corresponding rectangle with R.

Learning algorithm. We choose a very simple algorithm A which receives S as input, i.e., $A(S)$, and selects the *tightest rectangle* around the positively labeled points in S. We denote the hypothesis returned by this algorithm as h_S and its corresponding rectangle as R_S.

Claim 1. *A is an ERM algorithm.*

Proof. First, A labels all positive samples in S correctly. Second, since we assume realizability and A returns the tightest rectangle around the positive samples, all negative samples are labeled correctly as well $\Longrightarrow L_S(A(S) = L_S(h_S) = 0$. □

Claim 2. *$\mathcal{H}$ is PAC learnable via the proposed ERM algorithm A.*

Proof. Assume the rectangle corresponding to f is given by

$$R = [b^*, t^*] \times [r^*, s^*] \ ,$$

where $[b^*, t^*]$ denotes an interval on the x-axis and $[r^*, s^*]$ an interval on the y-axis.

We now fix some $\epsilon \in (0, 1)$ and assume that $\mathcal{D}([b^*, t^*] \times [r^*, s^*]) > \epsilon$. Otherwise, any selected hypothesis would have error $< \epsilon$ and the result is trivial.

We also observe that $R_S \subseteq R$, by construction, and the only region where errors can occur is $R \setminus R_S$ (here: $\setminus$ denotes the set difference).

We start by constructing four axis-parallel "strips", $r_1, \dots, r_4$, as follows (here, only r_1 is formally defined):

$$r_1 = [b^*, t_1] \times [r^*, s^*], \text{ with }, t_1 = \inf\{t : \mathcal{D}([b^*, t] \times [r^*, s^*]) \geq \epsilon/4\}$$

The infimum exists, as the measure, $\mathcal{D}$, is continuous from below and above. Letting

$$r_1' = [b^*, t_1[\ \times \ [r^*, s^*] \ ,$$

it follows that

$$\mathcal{D}(r_1') < \epsilon/4 \ ,$$

i.e., the probability drops below $\epsilon/4$ as soon as t_1 is excluded.

Lets assume for a moment that all r_i overlap R_S, i.e.,

$$\forall i : R_S \cap r_i \neq \emptyset \iff S \cap r_i \neq \emptyset$$

In that case,

$$R \setminus R_S \subset \bigcup_{i=1}^{4} r_i',$$

which implies

$$\begin{aligned} \mathcal{D}(R \setminus R_S) &< \mathcal{D}\left(\bigcup_{i=1}^{4} r_i'\right) \\ &\leq \sum_{i=1}^{4} \mathcal{D}(r_i') \\ &< \epsilon \end{aligned} \tag{1}$$

and $L_{\mathcal{D},f}(h_S) < \epsilon$ would hold.

> ***Question.*** *How large does $m = |S|$ have to be such that $\forall i : R_S \cap r_i \neq \emptyset$ with probability $1 - \delta$, $\delta \in (0, 1/2)$?*

We turn this question around and aim to bound the probability of the event that $R_S \cap r_i = 0$ for *some i*. In fact,

$$\begin{aligned}
\mathcal{D}^m\Big(\{S|_x \,:\, L_{\mathcal{D},f}(h_S) > \epsilon\}\Big) &\leq \mathcal{D}^m\Big(\bigcup_i \{S|_x \,:\, R_S \cap r_i = 0\}\Big) \\
&\leq \sum_i \mathcal{D}^m\Big(\{S|_x \,:\, R_S \cap r_i = 0\}\Big) \\
&= \sum_i \mathcal{D}^m\Big(\{S|_x \,:\, S \cap r_i = 0\}\Big) \\
&= 4(1 - \mathcal{D}(r_i))^m
\end{aligned} \tag{2}$$

By construction, $\mathcal{D}(r_i) \geq \epsilon/4$, hence, $1 - \mathcal{D}(r_i) \leq \epsilon/4$ and we get

$$\mathcal{D}^m\Big(\{S|_x \,:\, L_{\mathcal{D},f}(h_S) > \epsilon\}\Big) \leq 4(1 - \epsilon/4)^m \leq 4e^{-\epsilon/4} \tag{3}$$

Requiring that the last term is $< \delta$, yields

$$\begin{aligned}
& 4e^{-m\epsilon/4} < \delta \\
\iff & e^{-m\epsilon/4} < \delta/4 \\
\iff & -m\epsilon/4 < \log(\delta/4) \\
\iff & m\epsilon/4 > \log(\delta/4) \\
\iff & m > \frac{4}{\epsilon} \log\left(\frac{\delta}{4}\right)
\end{aligned}$$

In summary, we have

$$\mathcal{D}^m\Big(\{S|_x \,:\, L_{\mathcal{D},f}(A(S)) > \epsilon\}\Big) < \delta$$

for $\epsilon \in (0, 1)$, $\delta \in (0, 1/2)$ as long as

$$m > \frac{4}{\epsilon} \log\left(\frac{\delta}{4}\right) \quad .$$

This establishes *PAC learnability* of $\mathcal{H}$ by the chosen ERM algorithm A, as we have shown that given enough samples, labeled by f, the generalization error of a hypothesis selected by A being $> \epsilon$ holds with probability $< \delta$ (over S). $\square$