Solutions to
Understanding Machine Learning
by Shai Shalev-Shwartz and Shai Ben-David

Chapter 2 (A Gentle Start)

Zahra Taheri[1]

Feb 2020

Exercise 2.1

Let $\{$ be the labeling function. We want to show that given a training set S as follows

$$S = ((x_i, \{(x_i)))_{i=1}^m \subseteq (\mathbb{R}^d \times \{0,1\})^m,$$

there exists a polynomial p_S such that $h_S(x) = 1$ if and only if $p_S(x) \geq 0$, where $h_S : \mathcal{X} \to \mathcal{Y}$ is the following predictor:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s. t. } x = x_i \\ 0 & \text{otherwise} \end{cases}$$

where $[m] = \{1, \ldots, m\}$ and $y_i = \{(x_i)$, for all $i \in [m]$.

Obviously, we have:

$$h_S(x) = \begin{cases} 1 & \text{if } \exists i \in [m] \text{ s. t. } x = x_i \text{ and } y_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

It is easy to see that if $y_i = 0$, for all $i \in [m]$, then $h_S(x) = 0$, for all $x \in \mathcal{X}$. In this case, if $p_S(x) := -1$, for all $x \in \mathcal{X}$, then the statement is obviously true.

Suppose that there exists an i in $[m]$ such that $y_i = 1$.

1. Let $m = 1$ and $S = ((x_1, 1))$. Then $h_S(x_1) = 1$ and $h_S(x) = 0$, for all $x \in \mathcal{X} \setminus \{x_1\}$. So we must have $p_S(x_1) \geq 0$ and $p_S(x) < 0$, for all $x \in \mathcal{X} \setminus \{x_1\}$. Let $p_S(x) := -\|x - x_1\|^2$. Such a definition helps us to check the distance between x and x_1. It is easy to see that $p_S(x_1) = 0$ and $p_S(x) < 0$, for all $x \in \mathcal{X} \setminus \{x_1\}$.

[1] https://github.com/zahta/Exercises-Understanding-Machine-Learning

2. Let $m = 2$. Then without loss of generality we may assume that $S = ((x_1, 1), (x_2, 0))$ or $S = ((x_1, 1), (x_2, 1))$. * If $S = ((x_1, 1), (x_2, 0))$, then $h_S(x_1) = 1$ and $h_S(x) = 0$, for all $x \in \mathcal{X} \setminus \{x_1\}$. So, similar to the case (1), by defining $p_S(x) := -\|x - x_1\|^2$, the statement is obviously true. * If $S = ((x_1, 1), (x_2, 1))$, then $h_S(x_1) = h_S(x_2) = 1$ and $h_S(x) = 0$, for all $x \in \mathcal{X} \setminus \{x_1, x_2\}$. So we must have $p_S(x_1) \geq 0$, $p_S(x_2) \geq 0$ and $p_S(x) < 0$, for all $x \in \mathcal{X} \setminus \{x_1, x_2\}$. With a similar discussion as above, let $p_S(x) := -(\|x - x_1\|^2)(\|x - x_1\|^2)$. Then $p_S(x)$ is a polynomial such that $p_S(x_1) = p_S(x_2) = 0$ and $p_S(x) < 0$, for all $x \in \mathcal{X} \setminus \{x_1, x_2\}$, and so the statement is obviously true.

Inductively, we can generalize the obtained polynomials in cases (1) and (2) as follows:

$$p_S(x) := -\prod_{i \in [m] \ y_i = 1} \|x - x_i\|^2$$

Then $p_S(x)$ is a polynomial such that $p_S(x_i) = 0$, for all $i \in [m]$ s. t. $y_i = 1$, and $p_S(x) < 0$, for all $x \in \mathcal{X} \setminus \{x_i | i \in [m] \text{ and } y_i = 1\}$, and so the statement is obviously true.

Exercise 2.2

Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$ and $\mathcal{D}$ be a probability distribution over $\mathcal{X}$. Let $\{$ be the target hypothesis in $\mathcal{H}$. Let $S = ((x_i, \{(x_i)))_{i=1}^m$ be a training set such that $x_1, \ldots, x_m$ are i.i.d with respect to $\mathcal{D}$, denoted by $S|_{\mathcal{X}} \sim \mathcal{D}^m$. Fix some $h \in \mathcal{H}$.

$$\begin{aligned}
\mathbb{E}_{S|_{\mathcal{X}} \sim \mathcal{D}^m}[L_S(h)] &\overset{*}{=} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{x_i \sim \mathcal{D}}[\mathbb{1}_{h(x_i) \neq f(x_i)}] \\
&\overset{**}{=} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{x \sim \mathcal{D}}[\mathbb{1}_{h(x) \neq f(x)}] \\
&= m \left(\frac{1}{m}\right) \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \\
&= L_{(\mathcal{D}, \{)}[h]
\end{aligned}$$

* Expectation is linear

** $x_1, \ldots, x_m$ are i.i.d

Exercise 2.3

Let $S = ((x_i, y_i))_{i=1}^m$ be a training set.

1. Let $R(S)$ be the rectangle returned by A and $h_{A(S)} : \mathcal{X} \to \mathcal{Y}$ be the corresponding hypothesis. Since A returns the rectangle enclosing all positive examples in the training set, $h_{A(S)}(x_i) = 1$, for all

$i \in [m]$ such that $y_i = 1$. On the other hand, by the realizability assumption, there exists $h^* \in \mathcal{H}^2_{\text{rec}}$ such that $L_S(h^*) = 0$, and so $h^*(x_i) = 1$, for all $i \in [m]$ such that $y_i = 1$. Since A returns the smallest rectangle enclosing all positive examples, $L_S(h_{A(S)}) = 0$ and so A is an ERM.

2. Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}$. Also let $R^* = (a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels and { be its corresponding hypothesis. Suppose that $R_1, \ldots, R_4$ are defined as in the hint of this exercise. By the definitions of $R(S)$ and R^* we have $R(S) \subseteq R^*$. By definitions,

$$L_{(\mathcal{D},\{\})}(h_{A(S)}) = \mathcal{D}(\{x \in \mathcal{X} : h_{A(S)}(x) \neq f(x)\}) = \mathcal{D}(\{x \in \mathcal{X} : x \notin S|_{\mathcal{X}} \text{ and } f(x) = 1\}) = \mathcal{D}(R^* \setminus R(S)).$$

Since, the probability mass of the rectangle R_i is exactly $\frac{\varepsilon}{4}$, for all $i \in \{1, 2, 3, 4\}$, if S contains (positive) examples in all of the rectangles $R_1, \ldots, R_4$, then $\mathcal{D}(R^* \setminus R(S)) \leq 4(\frac{\varepsilon}{4}) = \varepsilon$. Therefore, $L_{(\mathcal{D},\{\})}(h_{A(S)}) \leq \varepsilon$. Now, we would like to upper bound $\mathcal{D}^m(\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{\})}(h_S) > \varepsilon\})$. With the discussion above, if S contains (positive) examples in all of the rectangles $R_1, \ldots, R_4$, then $L_{(\mathcal{D},\{\})}(h_{A(S)}) \leq \varepsilon$. Therefore,

$$\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{\})}(h_S) > \varepsilon\} = \bigcup_{i=1}^{4} \{S|_{\mathcal{X}} : S|_{\mathcal{X}} \cap R_i = \varnothing\}.$$

It is easy to see that $\mathcal{D}^m(\{S|_{\mathcal{X}} : S|_{\mathcal{X}} \cap R_i = \varnothing\}) = (1 - \frac{\varepsilon}{4})^m \leq e^{-\frac{\varepsilon}{4}m}$, for all $i \in \{1, 2, 3, 4\}$. With the discussion above and the union bound, $\mathcal{D}^m(\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{\})}(h_S) > \varepsilon\}) \leq \sum_{i=1}^{4} e^{-\frac{\varepsilon}{4}m} = 4e^{-\frac{\varepsilon}{4}m}$. So, the assumption $m \geq \frac{4\log(4/\delta)}{\varepsilon}$ completes the proof.

3. Similar to the definition of axis aligned rectangles in $\mathbb{R}^2$, given real numbers $a_1 \leq b_1, \ldots, a_d \leq b_d$, define the classifier $h_{(a_1,b_1,\ldots,a_d,b_d)}$ as follows:

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

Also, the class of all axis aligned rectangles in $\mathbb{R}^d$ is defined as follows:

$$\mathcal{H}^d_{\text{rec}} = \{h_{(a_1,b_1,\ldots,a_d,b_d)} : a_1 \leq b_1, \ldots, a_d \leq b_d\}.$$

For $i \in [2d]$, we define rectangles R_i similar to $R_1, \ldots, R_4$ in the latter case, with the probability mass $\frac{\varepsilon}{2d}$. By generalizing the algorithm A to the case $\mathbb{R}^d$, the proofs of parts (1) and (2) are straightforward (by considering a training set of size $\geq \frac{2d\log(2d/\delta)}{\varepsilon}$).