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

## Chapter 3 (A Formal Learning Model)

**Zahra Taheri**[1]

Mar 2020

**Exercise** 3.2

1. As it is mentioned in the exercise, the realizability assumption here implies that the true hypothesis $\{$ labels negatively all examples in the domain $\mathcal{X}$, perhaps except one. Let $A$ be the algorithm that returns an hypothesis $h_S$ with the following property:

$$h_S = \begin{cases} h_x & \text{if } \exists x \in S \text{ s. t. } \{(x) = 1 \\ h^- & \text{otherwise} \end{cases}$$

It is easy to see that $L_S(h_S) = 0$, and so $A$ is an $ERM$.

2. Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}$ and $\varepsilon \in (0, 1)$. If $\{ = h^-$, then $A$ returns the true hypothesis. Suppose that there exists an $x$ in $\mathcal{X}$ such that $\{(x) = 1$. Such an element is unique, by the realizability assumption. Let $S|_{\mathcal{X}} = (x_1, \ldots, x_m)$ be the instances of the training set. We would like to upper bound $\mathcal{D}^m(\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{)}(h_S) > \varepsilon\})$. If $x \in S|_{\mathcal{X}}$, then $A$ returns the true hypothesis and so, $L_{(\mathcal{D},\{)}(h_S) = 0$. Therefore we are interested in cases that $x \notin S|_{\mathcal{X}}$. Also, if $\mathcal{D}(x) \leq \varepsilon$, then $L_{(\mathcal{D},\{)}(h) \leq \varepsilon$, for all $h$ in the hypothesis class. So, we should suppose that $\mathcal{D}(x) > \varepsilon$. Then, $\mathcal{D}(x') \leq 1 - \varepsilon$, for all $x' \in \mathcal{X} \setminus x$. Hence we have

$$\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{)}(h_S) > \varepsilon\} = \{S|_{\mathcal{X}} : x \notin S|_{\mathcal{X}} \text{ and } \mathcal{D}(x) > \varepsilon\} = \{S|_{\mathcal{X}} : \forall x' \in S|_{\mathcal{X}} \quad \mathcal{D}(x') \leq 1 - \varepsilon\}.$$

Therefore we have:

$$\mathcal{D}^m(\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{)}(h_S) > \varepsilon\}) = \mathcal{D}^m(\{S|_{\mathcal{X}} : \forall x' \in S|_{\mathcal{X}} \;\; \mathcal{D}(x') \leq 1 - \varepsilon\}) \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}.$$

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

Let $\delta \in (0,1)$ such that $e^{-\varepsilon m} \leq \delta$. So, $m \geq \frac{log(1/\delta)}{\varepsilon}$. Therefore, $\mathcal{H}_{singleton}$ is PAC learnable with $m_{\mathcal{H}_{singleton}} \leq \lceil \frac{log(1/\delta)}{\varepsilon} \rceil$.

**Exercise** 3.3

Similar to the Exercise 3 of Chapter 2, let $A$ be the algorithm that returns the smallest circle enclosing all positive examples in the training set $S$. Let $C(S)$ be the circle returned by $A$ with the radius $r(S)$ and $h_{A(S)} : \mathcal{X} \to \mathcal{Y}$ be the corresponding hypothesis. Similar to the Exercise 2.3, it is easy to see that $L_S(h_{A(S)}) = 0$ and so $A$ is an ERM.

Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}$, $\varepsilon \in (0,1)$, and { be the target hypothesis in $\mathcal{H}$. By the realizability assumption, there exists a circle $C^*$ with the radius $r^*$ and the corresponding hypothesis $h^*$ related to the zero generalization error. By the definitions of $C(S)$ and $C^*$ we have $C(S) \subseteq C^*$. Also we have:

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

Let $r_1 \leq r^*$ be a number such that the probability mass (with respect to $\mathcal{D}$) of the strip $C_1 = \{x \in \mathcal{R}^2 : r_1 \leq \|x\| \leq r^*\}$ is $\varepsilon$. Since $L_{(\mathcal{D},\{)}(h_{A(S)}) = \mathcal{D}(C^* \setminus C(S))$, by the discussion above we have $L_{(\mathcal{D},\{)}(h_{A(S)}) \leq \varepsilon$. Now, we would like to upper bound $\mathcal{D}^m(\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{)}(h_S) > \varepsilon\})$. With the discussion above,

$$\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{)}(h_S) > \varepsilon\} = \{S|_{\mathcal{X}} : S|_{\mathcal{X}} \cap C_1 = \varnothing\}.$$

Therefore we have :

$$\mathcal{D}^m(\{S|_{\mathcal{X}} : L_{(\mathcal{D},\{)}(h_S) > \varepsilon\}) \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

Let $\delta \in (0,1)$ such that $e^{-\varepsilon m} \leq \delta$. So, $m \geq \frac{log(1/\delta)}{\varepsilon}$. Therefore, $\mathcal{H}$ is PAC learnable with $m_{\mathcal{H}} \leq \lceil \frac{log(1/\delta)}{\varepsilon} \rceil$.

**Exercise** 3.4

1. Let $\mathcal{H}$ be the hypothesis class of all conjunctions over $d$ variables. If we show that $\mathcal{H}$ is finite, then by corollary 3.2, $\mathcal{H}$ is PAC learnable. Let $h \in \mathcal{H}$ and $h$ is not the all-negative hypothesis. Let $x = (x_1, \ldots, x_d) \in \mathcal{X}$. Then $h(x) = \bigwedge_{i=1}^{d} a_i$, where $a_i \in \{x_i, \bar{x_i}, none\}$, for all $i \in d$, in which by *none* we means that the literals $x_i$ and $\bar{x_i}$ are not appear in $h(x)$. Therefore, $|\mathcal{H}| = 3^d + 1$ and so by corollary 3.2, $\mathcal{H}$ is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \lceil \frac{log(|\mathcal{H}|/\delta)}{\varepsilon} \rceil.$$

2. Suppose that $S$ is a training set of size $m$ such that $x'_1, \ldots, x'_l$ are all positively labeled instances in $S$. By induction on $i \leq l$, we define conjunctions $h_i$. Let $h_0$ be the all-negative hypothesis with definition $h_0(x) := \bigwedge_{j=1}^{d} x_j \overline{x_j}$. Let $i+1 \leq l$ and $x'_{i+1} = (x_1^{i+1}, \ldots, x_d^{i+1})$. We obtain $h_{i+1}$ from $h_i$ as follows: * For all $j \in [d]$, if $x_j^{i+1} = 1$ and $\overline{x_j^{i+1}}$ is a literal of $h_i$ then delete $\overline{x_j^{i+1}}$. * For all $j \in [d]$, if $x_j^{i+1} = 0$ and $x_j^{i+1}$ is a literal of $h_i$ then delete $x_j^{i+1}$.

The algorithm returns $h_l$. It is easy to see that $h_l$ labels $x'_1, \ldots, x'_l$ as positive. Since $h_l$ is the most restrictive conjunction that labels positively all the positively labeled members of $S$ and by the realizability assumption, $L_S(h_l) = 0$ and so the algorithm implements the ERM rule.

(Note: The solution of this exercise is explained in Section 8.2.3 of the book)

**Exercise** 3.5

Let $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\overline{\mathcal{D}}_{\Updownarrow},\{\})}(h) > \varepsilon\}$ and $M = \{S|_{\mathcal{X}} : \exists h \in \mathcal{H}_B \;\; s.t. \;\; L_{(S,f)}(h) = 0\}$. Then we have $\mathbb{P}\left[\exists h \in \mathcal{H} \; s.t. \; L_{(\overline{\mathcal{D}}_{\Updownarrow},\{\})}(h) > \varepsilon \; and \; L_{(S,f)}(h) = 0\right] = \mathbb{P}[M] = \mathbb{P}\left[\bigcup_{h \in \mathcal{H}_B} \{S|_{\mathcal{X}} : L_{(S,f)}(h) = 0\}\right]$. So by the union bound we have:

$$\mathbb{P}\left[\exists h \in \mathcal{H} \; s.t. \; L_{(\overline{\mathcal{D}}_{\Updownarrow},\{\})}(h) > \varepsilon \; and \; L_{(S,f)}(h) = 0\right] \leq |\mathcal{H}| \times \mathbb{P}\left[\{S|_{\mathcal{X}} : L_{(S,f)}(h) = 0\}\right] \tag{1}$$

On the other hand, if there exists $h \in \mathcal{H}$ such that $L_{(\overline{\mathcal{D}}_{\Updownarrow},\{\})}(h) > \varepsilon$ then by definition of the generalization error we have:

$$\frac{\mathbb{P}_{x \sim \mathcal{D}_i}[h(x) \neq f(x)] + \cdots + \mathbb{P}_{x \sim \mathcal{D}_m}[h(x) \neq f(x)]}{m} > \varepsilon$$

Therefore

$$\frac{\mathbb{P}_{x \sim \mathcal{D}_i}[h(x) = f(x)] + \cdots + \mathbb{P}_{x \sim \mathcal{D}_m}[h(x) = f(x)]}{m} \leq 1 - \varepsilon \tag{2}$$

Also we have:

$$\mathbb{P}\left[\{S|_{\mathcal{X}} : L_{(S,f)}(h) = 0\}\right] = \prod_{i=1}^{m} \mathbb{P}_{x \sim \mathcal{D}_i}[h(x) = f(x)] = \left(\left(\prod_{i=1}^{m} \mathbb{P}_{x \sim \mathcal{D}_i}[h(x) = f(x)]\right)^{1/m}\right)^m$$

So by geometric-arithmetic mean inequality and (2) we have:

$$\mathbb{P}\left[\{S|_{\mathcal{X}} : L_{(S,f)}(h) = 0\}\right] \leq \left(\frac{\mathbb{P}_{x \sim \mathcal{D}_i}[h(x) = f(x)] + \cdots + \mathbb{P}_{x \sim \mathcal{D}_m}[h(x) = f(x)]}{m}\right)^m \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

Therefore by (1) we have:

$$\mathbb{P}\left[\exists h \in \mathcal{H} \; s.t. \; L_{(\overline{\mathcal{D}}_{\Updownarrow},\{\})}(h) > \varepsilon \; and \; L_{(S,f)}(h) = 0\right] \leq |\mathcal{H}| e^{-\varepsilon m}.$$

**Exercise** 3.6

Suppose that $\mathcal{H}$ is agnostic PAC learnable. So there exist an algorithm $A$ and a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running $A$ on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, $A$ returns a hypothesis $h$ such that, with probability of at least $1 - \delta$ (over the choice of the $m$ training examples), $L_{\mathcal{D}}(h) \leq min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$.

Now we want to show that if the realizability assumption holds, $\mathcal{H}$ is PAC learnable using $A$. Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}$ and { be the target hypothesis in $\mathcal{H}$. Consider the distribution $\mathcal{D}'$ over $\mathcal{X} \times \{0,1\}$ obtained by drawing $x \in \mathcal{X}$ according to $\mathcal{D}$ and taking the pair $(x, \{(x))$. By the realizability assumption, $min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') = 0$. Let $\varepsilon, \delta \in (0,1)$. Therefore when running $A$ on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples which are labeled by {, $A$ returns a hypothesis $h$ such that, with probability of at least $1 - \delta$ (over the choice of the $m$ training examples) we have:

$$L_{\mathcal{D}}(h) \leq min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon = \varepsilon.$$

**Exercise** 3.7

The Bayes Optimal Predictor: Given any probability distribution $\mathcal{D}$ over $X \times \{0,1\}$, the Bayes Optimal Predictor is the following label predicting function from $X$ to $\{0,1\}$:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1 | x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

We want to verify that for every probability distribution $\mathcal{D}$, the Bayes optimal predictor $f_{\mathcal{D}}$ is optimal. It means that for every classifier $g : X \to \{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$. Note that for every classifier $g$,

$$L_{\mathcal{D}}(g) = \mathbb{P}[\{g(X) \neq Y\}] = \mathbb{E}_X [\mathbb{P}[\{g(X) \neq Y | X = x\}]].$$

We prove that $\mathbb{P}[\{g(X) \neq Y\}] \geq \mathbb{P}[\{f_{\mathcal{D}}(X) \neq Y\}]$.

It is easy to see that for every classifier $g : X \to \{0,1\}$ we have:

$$\mathbb{P}[\{g(X) \neq Y | X = x\}] = 1 - \mathbb{P}[\{g(X) = Y | X = x\}] = 1 - \mathbb{P}[\{g(X) = 1, Y = 1 | X = x\}] - \mathbb{P}[\{g(X) = 0, Y = 0 | X = x\}$$

Also we have 1. $\mathbb{P}[\{g(X) = 1, Y = 1 | X = x\}] = \mathbb{P}[\{g(X) = 1 | X = x\}]\,\mathbb{P}[\{Y = 1 | X = x\}]$

2. $\mathbb{P}[\{g(X) = 0, Y = 0 | X = x\}] = \mathbb{P}[\{g(X) = 0 | X = x\}]\,\mathbb{P}[\{Y = 0 | X = x\}]$

Note that if $g(x) = 1$ then $\mathbb{P}[\{g(X) = 1|X = x\}] = 1$ and if $g(x) = 0$ then $\mathbb{P}[\{g(X) = 1|X = x\}] = 0$. Also if $g(x) = 0$ then $\mathbb{P}[\{g(X) = 0|X = x\}] = 1$ and if $g(x) = 1$ then $\mathbb{P}[\{g(X) = 0|X = x\}] = 0$.

Therefore

$$1 - \mathbb{P}[\{g(X) = Y|X = x\}] = 1 - (\mathbb{1}_{g(x)=1}\mathbb{P}[\{Y = 1|X = x\}] + \mathbb{1}_{g(x)=0}\mathbb{P}[\{Y = 0|X = x\}])$$

So we have:

$$\mathbb{P}[\{f_{\mathcal{D}}(X) = Y|X = x\}] - \mathbb{P}[\{g(X) = Y|X = x\}] = \mathbb{P}[\{Y = 1|X = x\}](\mathbb{1}_{f_{\mathcal{D}}(x)=1} - \mathbb{1}_{g(x)=1}) + \mathbb{P}[\{Y = 0|X = x\}](\mathbb{1}_{f_{\mathcal{D}}($$

Since $\mathbb{P}[\{Y = 0|X = x\}] = 1 - \mathbb{P}[\{Y = 1|X = x\}]$ and for every classifier $g : X \to \{0, 1\}$, $\mathbb{1}_{g(x)=0} = 1 - \mathbb{1}_{g(x)=1}$, we have:

$$\mathbb{P}[\{f_{\mathcal{D}}(X) = Y|X = x\}] - \mathbb{P}[\{g(X) = Y|X = x\}] = \mathbb{P}[\{Y = 1|X = x\}](\mathbb{1}_{f_{\mathcal{D}}(x)=1} - \mathbb{1}_{g(x)=1}) + (1 - \mathbb{P}[\{Y = 1|X = x\}])($$

Therefore

$$\mathbb{P}[\{f_{\mathcal{D}}(X) = Y|X = x\}] - \mathbb{P}[\{g(X) = Y|X = x\}] = (2\mathbb{P}[\{Y = 1|X = x\}] - 1)\left(\mathbb{1}_{f_{\mathcal{D}}(x)=1} - \mathbb{1}_{g(x)=1}\right)$$

So by the definition of the Bayes predictor, for all $x \in X$ we have

$$\mathbb{P}[\{f_{\mathcal{D}}(X) = Y|X = x\}] - \mathbb{P}[\{g(X) = Y|X = x\}] \geq 0$$

Hence, for all $x \in X$ we have

$$\mathbb{P}[\{g(X) \neq Y|X = x\}] \geq \mathbb{P}[\{f_{\mathcal{D}}(X) \neq Y|X = x\}]$$

Since $\mathbb{E}_{X,Y}[f(X, Y)] = \mathbb{E}_X\mathbb{E}_{Y|X=x}[f(X, Y)|X = x]$, by the latter inequality we have:

$$L_{\mathcal{D}}(g) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{g(x)\neq y}\right] = \mathbb{E}_{x\sim\mathcal{D}_X}\left[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}\left[\mathbb{1}_{g(x)\neq y}|X = x\right]\right] = \mathbb{E}_{x\sim\mathcal{D}_X}\left[\mathbb{P}[\{g(X) \neq Y|X = x\}]\right] \geq \mathbb{E}_{x\sim\mathcal{D}_X}\left[\mathbb{P}[\{f_{\mathcal{D}}\right.$$