Understanding Machine Learning
Solution Manual

Written by Alon Gonen*
Edited by Dana Rubinstein

November 17, 2014

2 Gentle Start

1. Given $S = ((\mathbf{x}_i, y_i))_{i=1}^m$, define the multivariate polynomial

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

Then, for every i s.t. $y_i = 1$ we have $p_S(\mathbf{x}_i) = 0$, while for every other $\mathbf{x}$ we have $p_S(\mathbf{x}) < 0$.

2. By the linearity of expectation,

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

*The solutions to Chapters 13,14 were written by Shai Shalev-Shwartz

3. (a) First, observe that by definition, A labels positively all the positive instances in the training set. Second, as we assume realizability, and since the tightest rectangle enclosing all positive examples is returned, all the negative instances are labeled correctly by A as well. We conclude that A is an ERM.

   (b) Fix some distribution $\mathcal{D}$ over $\mathcal{X}$, and define $R^\star$ as in the hint. Let f be the hypothesis associated with $R^\star$ a training set S, denote by $R(S)$ the rectangle returned by the proposed algorithm and by $A(S)$ the corresponding hypothesis. The definition of the algorithm A implies that $R(S) \subseteq R^*$ for every S. Thus,

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

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

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

   Applying the union bound, we obtain

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

   Thus, it suffices 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 of the instances don't fall in 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,

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

   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 $a_1 \leq b_1, a_2 \leq b_2, \ldots, a_d \leq b_d$, define the classifier $h_{(a_1,b_1,\ldots,a_d,b_d)}$ by

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

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

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

It can be seen 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 strips for each dimension). Thus, it suffices to draw a training set of size $\left\lceil \frac{2d\log(2d/\delta)}{\epsilon} \right\rceil$.

(d) For each dimension, the algorithm has to find the minimal and the maximal values among the positive instances in the training sequence. Therefore, its runtime is $O(md)$. Since we have shown that the required value of m is at most $\left\lceil \frac{2d\log(2d/\delta)}{\epsilon} \right\rceil$, it follows that the runtime of the algorithm is indeed polynomial in $d, 1/\epsilon$, and $\log(1/\delta)$.

3 A Formal Learning Model

1. 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 $0 < \epsilon_1 \leq \epsilon_2 \leq 1$. We need to show that $m_1 \stackrel{\text{def}}{=} m_{\mathcal{H}}(\epsilon_1,\delta) \geq m_{\mathcal{H}}(\epsilon_2,\delta) \stackrel{\text{def}}{=} m_2$. Given an i.i.d. training sequence of size $m \geq m_1$, we have that with probability at least $1-\delta$, A returns a hypothesis h such that

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

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

2. (a) We propose the following algorithm. If a positive instance x_+ appears in S, return the (true) hypothesis h_{x_+}. If S doesn't contain any positive instance, the algorithm returns the all-negative hypothesis. It is clear that this algorithm is an ERM.

   (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's clear that if x_+ appears in the training sequence S, our algorithm returns a perfect hypothesis. Furthermore, if $\mathcal{D}[\{x_+\}] \leq \epsilon$ 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_+ doesn't appear in S. Denote this event by F. Then

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

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

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

3. Consider the ERM algorithm A which given a training sequence $S = ((\mathbf{x}_i, y_i))_{i=1}^m$, 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 $h^\star$ be a circle with zero generalization error. Denote its radius by $r^\star$.

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

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

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

4. We first observe that $\mathcal{H}$ is finite. Let us calculate its size accurately. Each hypothesis, besides the all-negative hypothesis, is determined by deciding for each variable x_i, whether x_i, $\bar{x}_i$ or none of which appear in the corresponding conjunction. Thus, $|\mathcal{H}| = 3^d + 1$. We conclude that $\mathcal{H}$ is PAC learnable and its sample complexity can be bounded by

$$m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{d\log 3 + \log(1/\delta)}{\epsilon} \right\rceil .$$

   Let's describe our learning algorithm. We define $h_0 = x_1 \cap \bar{x}_1 \cap \ldots \cap x_d \cap \bar{x}_d$. Observe that h_0 is the always-minus hypothesis. Let $((\mathbf{a}^1, y^1), \ldots, (\mathbf{a}^m, y^m))$ be an i.i.d. training sequence of size m. Since

we cannot produce any information from negative examples, our algorithm neglects them. For each positive example a, we remove from h_i all the literals that are missing in a. That is, if $a_i = 1$, we remove $\bar{x}_i$ from h and if $a_i = 0$, we remove x_i from h_i. Finally, our algorithm returns h_m.

By construction and realizability, h_i labels positively all the positive examples among $\mathbf{a}^1, \ldots, \mathbf{a}^i$. From the same reasons, the set of literals in h_i contains the set of literals in the target hypothesis. Thus, h_i classifies correctly the negative elements among $\mathbf{a}^1, \ldots, \mathbf{a}^i$. This implies that h_m is an ERM.

Since the algorithm takes linear time (in terms of the dimension d) to process each example, the running time is bounded by $O(m \cdot d)$.

5. Fix some $h \in \mathcal{H}$ with $L_{(\overline{\mathcal{D}}_m, f)}(h) > \epsilon$. By definition,

$$\frac{\mathbb{P}_{X \sim \mathcal{D}_1}[h(X) = f(X)] + \ldots + \mathbb{P}_{X \sim \mathcal{D}_m}[h(X) = f(X))]}{m} < 1 - \epsilon \ .$$

We now bound the probability that h is consistent with S (i.e., that $L_S(h) = 0$) as follows:

$$\begin{aligned}
\mathbb{P}_{S \sim \prod_{i=1}^m \mathcal{D}_i}[L_S(h) = 0] &= \prod_{i=1}^m \mathbb{P}_{X \sim \mathcal{D}_i}[h(X) = f(X)] \\
&= \left(\left(\prod_{i=1}^m \mathbb{P}_{X \sim \mathcal{D}_i}[h(X) = f(X)] \right)^{\frac{1}{m}} \right)^m \\
&\leq \left(\frac{\sum_{i=1}^m \mathbb{P}_{X \sim \mathcal{D}_i}[h(X) = f(X)]}{m} \right)^m \\
&< (1 - \epsilon)^m \\
&\leq e^{-\epsilon m} \ .
\end{aligned}$$

The first inequality is the geometric-arithmetic mean inequality. Applying the union bound, we conclude that the probability that there exists some $h \in \mathcal{H}$ with $L_{(\overline{\mathcal{D}}_m, f)}(h) > \epsilon$, which is consistent with S is at most $|\mathcal{H}| \exp(-\epsilon m)$.

6. 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 $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, 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

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

7. Let $x \in \mathcal{X}$. Let α_x be the conditional probability of a positive label given x. We have

$$\begin{aligned} \mathbb{P}[f_{\mathcal{D}}(X) \neq y | X = x] &= \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot \mathbb{P}[Y = 0 | X = x] + \mathbb{1}_{[\alpha_x < 1/2]} \cdot \mathbb{P}[Y = 1 | X = x] \\ &= \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot (1 - \alpha_x) + \mathbb{1}_{[\alpha_x < 1/2]} \cdot \alpha_x \\ &= \min\{\alpha_x, 1 - \alpha_x\}. \end{aligned}$$

Let g be a classifier[1] from $\mathcal{X}$ to $\{0,1\}$. We have

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

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

$$\begin{aligned} L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}_{[f_{\mathcal{D}}(x) \neq y]}] \\ &= \mathbb{E}_{x \sim \mathcal{D}_X} \left[\mathbb{E}_{y \sim \mathcal{D}_{Y|x}} [\mathbb{1}_{[f_{\mathcal{D}}(x) \neq y]} | X = x] \right] \\ &= \mathbb{E}_{x \sim \mathcal{D}_X} [\alpha_x] \\ &\leq \mathbb{E}_{x \sim \mathcal{D}_X} \left[\mathbb{E}_{y \sim \mathcal{D}_{Y|x}} [\mathbb{1}_{[g(x) \neq y]} | X = x] \right] \\ &= L_{\mathcal{D}}(g) \ . \end{aligned}$$

[1] As we shall see, g might be non-deterministic.