Understanding Machine Learning: Exercises

Somnath Sikdar

August 3, 2019

Contents

What this is About

These notes are my attempt to understand and work out material from the textbook *Understanding Machine Learning* by Shai Shalev-Shwartz and Shai Ben-David.

Chapter 3

A Formal Learning Model

Exercise 3.1

Let $m_{\mathcal{H}}(\epsilon, \delta)$ be the sample complexity of a PAC-learnable hypothesis class $\mathcal{H}$ for a binary classification task. For a fixed δ, let $0 < \epsilon_1 \leq \epsilon_2 < 1$ and suppose that $m_{\mathcal{H}}(\epsilon_1, \delta) < m_{\mathcal{H}}(\epsilon_2, \delta)$. Then when running the learning algorithm on $m_{\mathcal{H}}(\epsilon_1, \delta)$ i.i.d examples, we obtain a hypothesis h, which with probability at least $1 - \delta$ has a true error $L_{\mathcal{D},f}(h) \leq \epsilon_1 \leq \epsilon_2$. This implies that for the (ϵ_2, δ) combination of parameters, we can bound the true error of h by ϵ_2 by using a smaller number of i.i.d examples than $m_{\mathcal{H}}(\epsilon_2, \delta)$. This contradicts the minimality of the sample complexity function. Hence we must have $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$.

Next suppose that $0 < \delta_1 \leq \delta_2 < 1$ and that $m_{\mathcal{H}}(\epsilon, \delta_1) < m_{\mathcal{H}}(\epsilon, \delta_2)$, where ϵ is fixed in advance. Then with $m_{\mathcal{H}}(\epsilon, \delta_1)$ i.i.d examples, the learner outputs a hypothesis h which with probability at least $1 - \delta_1 \geq 1 - \delta_2$ has a true error of at most ϵ. This implies that for the (ϵ, δ_2) combination of parameters, we can bound the true error of h by ϵ by using a smaller number of i.i.d examples than $m_{\mathcal{H}}(\epsilon, \delta_2)$. This again contradicts the minimality of the sample complexity function. Hence we must have $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.

Exercise 3.2

Given a sample S, we output a hypothesis h_S with the property that $\forall x \in S_x$,

$$h_S(x) = \begin{cases} 1, & \text{if } (x, 1) \in S \\ 0, & \text{otherwise} \end{cases}$$

For any sample S, this hypothesis has an empirical loss of 0. Note that h_S disagrees with the true labeling function f in at most one point $z \in \mathcal{X}$. It's true loss is therefore $\Pr_{x \sim \mathcal{D}}\{f(x) \neq h_S(x)\} = \Pr_{\mathcal{D}}\{z\} := p_z$.

The true loss of h_S will be 0 if $(z, 1) \in S$. Therefore the probability of getting a "bad" sample is $\Pr_{S \sim \mathcal{D}^m}\{(z, 1) \notin S\}$. Let $z^* \in \mathcal{X}$ be a point at which $(1 - p_z)^m$ is maximized. Since $(1 - p_{z^*})^m \leq e^{-m p_{z^*}}$ and since we want the probability of picking a bad sample to be at most δ, we want $e^{-mp_{z^*}} < \delta$, which gives us the sample size to be:

$$m > \frac{\log(1/\delta)}{p_{z^*}} \tag{3.1}$$

Depending on the value of the error bound ϵ, there are two situations to consider. If $\epsilon \geq p_{z^*}$, then even a sample of size one will guarantee that the true error of h_s is at most ϵ. However if $\epsilon < p_{z^*}$ then we can then use this in (3.1) to obtain:

$$m > \frac{\log(1/\delta)}{\epsilon}.$$

Thus the sample complexity is $m_{\mathcal{H}}(\epsilon, \delta) = \max\left\{1, \frac{\log(1/\delta)}{\epsilon}\right\}$.

Exercise 3.3

Here $\mathcal{X} = \mathbf{R}^2$ and $\mathcal{Y} = \{0, 1\}$. The hypothesis class $\mathcal{H}$ is the set of concentric circles in $\mathbf{R}^2$ centered at the origin. Assuming realizability, this implies that the true labeling function $f = h_r$ for some $r \in \mathbf{R}_+$. Thus f assigns the label 1 to any point (x, y) that is within a distance of r from the origin and 0 otherwise.

Given any sample S, let $q \in \mathbf{R}_+$ be the minimum real number such that all $(x, y) \in S_x$ with a label of 1 are included in a circle centered at the origin with radius q. The output of the ERM procedure is h_q. The empirical error of h_q is zero, but it's true error is:

$$\Pr_{(x,y)\sim\mathcal{D}}\left\{(x, y) \in C_r \setminus C_q\right\}$$

where C_r and C_q are concentric circles centered at the origin with radius r and q respectively. Given an $\epsilon > 0$, let $t \in \mathbf{R}_+$ be such that

$$\epsilon = \Pr_{(x,y)\sim\mathcal{D}}\left\{(x, y) \in C_r \setminus C_t\right\}.$$

That is, we choose t so that the true error matches the probability of picking anything inside the ring described by the circles C_r and C_t. Then the probability that we fail to choose any point in this ring in an i.i.d sample of size m is $(1-\epsilon)^m \leq e^{-\epsilon m}$. This is the probability that we are handed a "bad" sample. Upper bounding this by δ, we obtain that $m > \log(1/\delta)/\epsilon$.

Now a sample of size at least $\log(1/\delta)/\epsilon$ has with probability at least $1 - \delta$ a point from $C_r \setminus C_t$, and hence the true error of the resulting ERM hypothesis is at most ϵ. Hence the sample complexity is upper bounded by $\lceil \log(1/\delta)/\epsilon \rceil$.

Exercise 3.4

In this example, $\mathcal{X} = \{0, 1\}^d$, $\mathcal{Y} = \{0, 1\}$ and the hypothesis class $\mathcal{H}$ is the set of all conjunctions over d Boolean variables. Since there are $\sum_{i=0}^{d} \binom{d}{i} 2^i = 3^d$ Boolean conjunctions

procedure PACBOOLEAN(S) $\triangleright$ S is the sample set with elements $\langle (a_1, \ldots, a_d), b \rangle$, where $(a_1, \ldots, a_d) \in \{0,1\}^d$ and $b \in \{0,1\}$
  $f \leftarrow x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$
  for each $\langle (a_1, \ldots, a_d), b \rangle \in S$ with $b = 1$ **do**
    for j in $[1, \ldots, d]$ **do**
      if $a_j = 1$ **then**
        Delete $\bar{x}_j$ from f, if it exists in the formula
      else
        Delete x_j from f, if it exists in the formula
      end if
    end for
  end for
  return f
end procedure

Figure 3.1: Learning Boolean conjunctions

over d Boolean variables, the hypothesis class is finite. Hence the sample complexity is

$$\begin{aligned} m_{\mathcal{H}}(\epsilon, \delta) &= \left\lceil \frac{\log(\mathcal{H}/\delta)}{\epsilon} \right\rceil \\ &= \left\lceil \frac{d \cdot \log 3 + \log(1/\delta)}{\epsilon} \right\rceil \end{aligned}$$

To prove that the class $\mathcal{H}$ is PAC-learnable, it suffices to exhibit a polynomial-time algorithm that implements the ERM rule. The algorithm outlined in Figure 3.1 starts with the formula $x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$. It runs through the positive examples in the sample S and for each such example, it adjusts the formula so that it satisfies the assignment given in the example. At the end of this procedure, the modified formula satisfies all positive examples of S. The time taken is $O(d \cdot |S|)$.

What may not be immediately apparent is that the formula returned by the algorithm satisfies all negative examples too. This is clear when the sample S has *no* positive examples to begin with as every assignment to $x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$ results in a 0. The point is that if there is even *one* positive example, for each $1 \leq i \leq d$, the algorithm eliminates either x_i or $\bar{x}_i$ depending on the assignment. That is, it eliminates half of the literals on seeing that one example and the modified formula f contains the literals of the true labeling function along with possibly others. Now the literals of the true labeling function produce a 0 on all negative examples and so does f. Hence the sampling error of the function returned by the algorithm is 0.

Exercise 3.5

The first thing to verify is that $\bar{\mathcal{D}}_m$ is a distribution. This is easy since for all $x \in \mathcal{X}$, $\bar{\mathcal{D}}_m(x) \geq 0$ and

$$\begin{aligned}\int_{x\in\mathcal{X}} \bar{\mathcal{D}}_m(x)\mathrm{d}x &= \frac{1}{m}\sum_{i=1}^{m}\int_{x\in\mathcal{X}} \mathcal{D}_i(x)\mathrm{d}x \\ &= \frac{1}{m}\sum_{i=1}^{m} 1 \\ &= 1.\end{aligned}$$

Fix an accuracy parameter $\epsilon > 0$. As in the text, define the set of bad hypotheses to be $\mathcal{H}_B = \{h \in \mathcal{H} \colon L_{\bar{\mathcal{D}}_m,f}(h) > \epsilon\}$ and let $\mathcal{M} = \{S|_x \colon \exists h \in \mathcal{H}_B, L_S(h) = 0\}$ be the set of misleading samples. Since we assume realizability, any hypothesis h output by the ERM procedure has $L_S(h) = 0$. Thus the event $L_{\bar{\mathcal{D}}_m,f}(h) > \epsilon$ and $L_S(h) = 0$ happens only when $S|_x \in \mathcal{M}$. Hence,

$$\begin{aligned}\Pr_{\forall i\colon x_i\sim\mathcal{D}_i}\{S|_x \in \mathcal{M}\} &= \Pr_{\forall i\colon x_i\sim\mathcal{D}_i}\left\{\bigcup_{h\in\mathcal{H}_B}\{S|_x \colon L_S(h) = 0\}\right\} \\ &\leq \sum_{h\in\mathcal{H}_B}\Pr_{\forall i\colon x_i\sim\mathcal{D}_i}\{S|_x \colon L_S(h) = 0\} \\ &= \sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}\Pr_{x_i\sim\mathcal{D}_i}\{f(x_i) = h(x_i)\} \\ &= \sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}\left(1 - L_{\mathcal{D}_i,f}(h)\right) \\ &\leq \sum_{h\in\mathcal{H}_B}\left[\frac{1}{m}\sum_{i=1}^{m}\left(1 - L_{\mathcal{D}_i,f}(h)\right)\right]^m \\ &\leq \sum_{h\in\mathcal{H}_B}\left[1 - L_{\bar{\mathcal{D}}_m,f}(h)\right]^m\end{aligned}$$

The second-last inequality follows from the fact that the arithmetic mean of a set of numbers is at most their geometric mean. The quantity $\sum_{h\in\mathcal{H}_B}[1 - L_{\bar{\mathcal{D}}_m,f}(h)]^m$ is at most $|\mathcal{H}| \cdot (1-\epsilon)^m$ which is at most $|\mathcal{H}| \cdot e^{-\epsilon m}$.

Exercise 3.6

Agnostic PAC-learnability implies PAC-learnability. Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\{0,1\}$ which is agnostic PAC-learnable wrt $\mathcal{X} \times \{0,1\}$ and the 0-1 loss function with sample complexity $m_{\mathcal{H}}$. Let f be a labeling function and let $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$ for which the realizability assumption holds, that is, there exists $h \in \mathcal{H}$ such that $L_{\mathcal{D}_{\mathcal{X}},f}(h) = 0$.

Define a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ as follows: for all $x \in \mathcal{X}$, $\mathcal{D}((x, f(x))) = \mathcal{D}_{\mathcal{X}}(x)$ and $\mathcal{D}((x, 1-f(x))) = 0$. Fix $\epsilon, \delta > 0$. Since $\mathcal{H}$ is agnostic PAC-learnable, there exists a learner A which given a sample S of $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ iid examples generated by $\mathcal{D}$ returns a hypothesis h_S such that

$$\Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \epsilon\right\} > 1-\delta.$$

Note that for any $h' \in \mathcal{H}$, we may write the loss $L_{\mathcal{D}}(h')$ as follows.

$$\begin{aligned} L_{\mathcal{D}}(h') &= \Pr_{(x,y)\in\mathcal{D}}\left\{h'(x) \neq y\right\} \\ &= \Pr_{(x,y)\in\mathcal{D}}\left\{h'(x) \neq f(x)\right\} \\ &= L_{\mathcal{D}_{\mathcal{X}},f}(h'). \end{aligned}$$

The second equality above follows since the only points $(x, y) \in \mathcal{X} \times \{0,1\}$ for which $\mathcal{D}$ places a non-zero probability mass are those for which $y = f(x)$. Since we assume realizability, $\min_{h'\in\mathcal{H}} L_{\mathcal{D}_{\mathcal{X}},f}(h') = 0$. Hence the hypothesis h_S returned by the learner A satisfies:

$$\Pr_{S|_{\mathcal{X}}\sim\mathcal{D}_{\mathcal{X}}^m}\left\{L_{\mathcal{D}_{\mathcal{X}},f}(h_S) \leq \epsilon\right\} > 1-\delta,$$

which is the condition for successful PAC-learnability.

Exercise 3.7

Let us fix some notation. We assume that X and Y are random variables defined over the domains $\mathcal{X}$ and $\{0,1\}$, respectively. Let $\mathcal{D}_{X,Y}$ be a distribution over $\mathcal{X} \times \{0,1\}$; let $\mathcal{D}_{Y|X}$, the conditional distribution of Y given X; l et $\mathcal{D}_X$ be the marginal distribution of X over $\mathcal{X}$; and, finally, let $\eta(x) = \Pr_{\mathcal{D}_{Y|X}}\{Y = 1 \mid X = x\}$.

The Bayes optimal classifier $f_{\mathcal{D}}$ may be written as:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Given any classifier $g\colon \mathcal{X} \to \{0,1\}$, the risk of this classifier is

$$L_{\mathcal{D}}(g) = \Pr_{\mathcal{D}_{X,Y}}\{g(X) \neq Y\} = \int_{x\in\mathcal{X}} \Pr_{\mathcal{D}_{Y|X}}\{g(x) \neq Y \mid X = x\} \cdot \Pr_{\mathcal{D}_X}\{X = x\}\, \mathrm{d}x. \quad (3.2)$$

We may write the first term of this intergrand as follows (where all probabilities are with respect to the conditional distribution $\mathcal{D}_{Y|X}$):

$$\begin{aligned}
\Pr\{g(x) \neq Y \mid X = x\} &= 1 - \Pr\{g(x) = Y \mid X = x\} \\
&= 1 - [\Pr\{g(x) = 1, Y = 1 \mid X = x\} + \Pr\{g(x) = 0, Y = 0 \mid X = x\}] \\
&= 1 - \left[\mathbf{1}_{g(x)=1} \cdot \Pr\{Y = 1 \mid X = x\} + \mathbf{1}_{g(x)=0} \cdot \Pr\{Y = 0 \mid X = x\}\right] \\
&= 1 - \left[\mathbf{1}_{g(x)=1} \cdot \eta(x) + \mathbf{1}_{g(x)=0} \cdot (1 - \eta(x))\right]
\end{aligned}$$

Consider the difference $\Pr\{g(x) \neq Y \mid X = x\} - \Pr\{f_{\mathcal{D}}(x) \neq Y \mid X = x\}$. This may be written as:

$$\begin{aligned}
&\left[\mathbf{1}_{f_{\mathcal{D}}(x)=1} \cdot \eta(x) + \mathbf{1}_{f_{\mathcal{D}}(x)=0} \cdot (1 - \eta(x))\right] - \left[\mathbf{1}_{g(x)=1} \cdot \eta(x) + \mathbf{1}_{g(x)=0} \cdot (1 - \eta(x))\right] \\
&\qquad = \left(\mathbf{1}_{f_{\mathcal{D}}(x)=1} - \mathbf{1}_{g(x)=1}\right) \cdot \eta(x) + \left(\mathbf{1}_{f_{\mathcal{D}}(x)=0} - \mathbf{1}_{g(x)=0}\right) \cdot (1 - \eta(x)).
\end{aligned}$$

This last expression may be written as:

$$\left(\mathbf{1}_{f_{\mathcal{D}}(x)=1} - \mathbf{1}_{g(x)=1}\right) \cdot \eta(x) + \left(\mathbf{1}_{g(x)=1} - \mathbf{1}_{f_{\mathcal{D}}(x)=1}\right) \cdot (1 - \eta(x)).$$

Rearranging terms allows us to write this as:

$$2 \cdot \left(\mathbf{1}_{f_{\mathcal{D}}(x)=1} - \mathbf{1}_{g(x)=1}\right) \cdot \eta(x) + \left(\mathbf{1}_{g(x)=1} - \mathbf{1}_{f_{\mathcal{D}}(x)=1}\right). \tag{3.3}$$

We claim that this last expression is always non-negative. If $f_{\mathcal{D}}(x) = 0$ then $\eta(x) < 1/2$ and the above expression is non-negative. If $f_{\mathcal{D}}(x) = 1$ then $\eta(x) \geq 1/2$ and, in this case too, the expression is non-negative. The result follows by plugging in the difference $\Pr\{g(x) \neq Y \mid X = x\} - \Pr\{f_{\mathcal{D}}(x) \neq Y \mid X = x\}$ in the integral in (3.2).

Chapter 4

Learning via Uniform Convergence

Notes on Chapter 4

Given any hypothesis class $\mathcal{H}$ and a domain $Z = \mathcal{X} \times Y$, let l be a loss function from $\mathcal{H} \times Z \to \mathbf{R}_+$. Finally let $\mathcal{D}$ be a distribution over the domain Z. The risk of a hypothesis $h \in \mathcal{H}$ is

$$L_{\mathcal{D}}(h) = \mathrm{Pr}_{z \sim \mathcal{D}} \{l(h,z)\}$$

A training set S is ϵ-representative w.r.t Z, $\mathcal{H}$, Z and l if for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$. Thus any hypothesis on an ϵ-representative training set has an in-sample error that is close to their true risk.

If S is ϵ-representative, then the $\mathrm{ERM}_{\mathcal{H}}(S)$ learning rule is guaranteed to return a good hypothesis. More specifically,

Lemma 1. *Fix a hypothesis class $\mathcal{H}$, a domain $Z = \mathcal{X} \times Y$, a loss function $l \colon \mathcal{H} \times Z \to \mathbf{R}_+$ and a distribution $\mathcal{D}$ over the domain Z. Let S be an $\epsilon/2$-representative sample. Then any output h_S of $\mathrm{ERM}_{\mathcal{H}}(S)$ satisfies*

$$L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

Therefore in order for the ERM rule to be an agnostic PAC-learner, all we need to do is to ensure that with probability of at least $1 - \delta$ over random choices of the training set, we end up with an $\epsilon/2$-representative training sample. This requirement is baked into the definition of *uniform convergence*.

Definition 1. A hypothesis class $\mathcal{H}$ is uniformly convergent wrt a domain Z and a loss function l, if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} \colon (0,1) \times (0,1) \to \mathbf{N}$ such that for all $\epsilon, \delta \in (0,1)$ and all distributions $\mathcal{D}$ on Z, if a sample of at least $m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$ examples is chosen i.i.d from $\mathcal{D}$, then with probability $1 - \delta$, the sample is ϵ-representative.

By Lemma (1), if $\mathcal{H}$ is uniformly convergent with function $m_{\mathcal{H}}^{\mathrm{UC}}$, then it is agnostically PAC-learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta)$. In this case, the ERM paradigm is a successful agnostic PAC-learner for $\mathcal{H}$.

Exercise 4.1

We first show that (1) $\Rightarrow$ (2). For each $n \in \mathbf{N}$, define $\epsilon_n = 1/2^n$ and $\delta_n = 1/2^n$. Then by (1), for each $n \in \mathbf{N}$, there exists $m(\epsilon_n, \delta_n)$ such that $\forall m \geq m(\epsilon_n, \delta_n)$,

$$\Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) > \epsilon_n\right\} < \delta_n.$$

We can then upper bound $\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right]$ as follows:

$$\begin{aligned}
\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right] &\leq \epsilon_n \cdot \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) \leq \epsilon_n\right\} + (1-\epsilon_n)\cdot \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) > \epsilon_n\right\} \\
&\leq \epsilon_n \cdot (1-\delta_n) + (1-\epsilon_n)\cdot \delta_n \\
&\leq \frac{1}{2^{n-1}} - \frac{1}{2^{2n-1}}.
\end{aligned}$$

The first inequality follows from the fact that the loss function is from $\mathcal{H} \times Z \to [0,1]$, which allows us to upper bound the value of the error when $L_{\mathcal{D}}(h_S) > \epsilon_n$ by $1-\epsilon_n$. As $n \to \infty$, $m \to \infty$ and $\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right] \to 0$, proving that (2) follows.

We next show that (2) $\Rightarrow$ (1). Fix $\epsilon, \delta > 0$. Define $\delta' = \epsilon \cdot \delta$. Since

$$\lim_{m\to\infty} \mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right] = 0,$$

there exists $m_1(\delta')$ such that for all $m \geq m_1(\delta')$ we have $\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right] < \delta'$. We now lower bound $\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right]$ as follows:

$$\begin{aligned}
\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right] &= \int_0^1 x \cdot \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) = x\right\} \mathrm{d}x \\
&\geq \int_\epsilon^1 x \cdot \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) = x\right\} \mathrm{d}x \\
&\geq \epsilon \cdot \int_\epsilon^1 \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) = x\right\} \mathrm{d}x \\
&= \epsilon \cdot \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) \geq \epsilon\right\}.
\end{aligned}$$

Choose $m(\epsilon, \delta) := m_1(\epsilon \cdot \delta)$. Then for all $m \geq m(\epsilon, \delta)$, we have that $\mathrm{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_s)\right] < \epsilon \cdot \delta$, from which it follows that:

$$\epsilon \cdot \Pr_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) \geq \epsilon\right\} < \epsilon \cdot \delta.$$

Condition (1) follows from this.

Chapter 5

The No-Free-Lunch Theorem

Notes on Chapter 5

Consider a binary classification task on a domain $\mathcal{X}$. Assume for the time being that $\mathcal{X}$ is finite. In this case, the set $\mathcal{H}$ of all functions from $\mathcal{X} \to \{0,1\}$ is finite and is hence PAC-learnable with sample complexity $\leq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$. Since $|\mathcal{H}| = 2^{|\mathcal{X}|}$, the sample complexity is $\frac{|\mathcal{X}|+\log(1/\delta)}{\epsilon} = O(|\mathcal{X}|)$.

The first question is what happens wrt PAC-learnability in this situation when we restrict the sample size? The No-Free-Lunch theorem shows that there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a labelling function $f\colon \mathcal{X} \to \{0,1\}$ that learners who are constrained to use at most $|\mathcal{X}|/2$ training examples "cannot learn." There is another way to interpret the No-Free-Lunch theorem: if the domain $\mathcal{X}$ is *infinite*, then the set of all functions from $\mathcal{X}$ to $\{0,1\}$ is not PAC-learnable no matter what the sample size.

Thus the No-Free-Lunch theorem has two interpretations, first, as a lower bound result on the sample complexity of PAC-learning and, second, as the inability to PAC-learn arbitrary hypothesis classes.

Theorem 1. *Consider the task of binary classification over the domain $\mathcal{X}$ wrt the 0-1 loss function. Let A be a learning algorithm that is constrained to use at most $m \leq |\mathcal{X}|/2$ training examples. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a function $f\colon \mathcal{X} \to \{0,1\}$ such that*

1. $L_{\mathcal{D}}(f) = 0$
2. *with probability of at least $1/7$ over the choice of training examples chosen iid from $\mathcal{D}^m$, we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Exercise 5.1

As the hint in the exercise suggests, let θ be a random variable that takes on values in the range $[0,1]$ with expectation $\mathrm{E}[\theta] \geq 1/4$. We want to show that $\Pr\{\theta \geq 1/8\} \geq 1/7$.

We start with Markov's inequality: for any nonnegative random variable X and $a > 0$,

$$\Pr\{X \geq a\} \leq \frac{\mathrm{E}[X]}{a}.$$

This doesn't quite work when we substitute $\theta = X$ and $a = 1/8$. The trick here lies in observing that θ is bounded from above by 1, and hence, if we define $\xi = 1 - \theta$ then ξ is nonnegative and we can use Markov's inequality on ξ. Note that $\mathrm{E}[\xi] = 1 - \mathrm{E}[\theta]$, and hence by Markov's inequality,

$$\begin{aligned}
\Pr\{\xi \geq a\} &\leq \frac{\mathrm{E}[\xi]}{a} \\
1 - \Pr\{\xi \geq a\} &\geq 1 - \frac{\mathrm{E}[\xi]}{a} \\
\Pr\{\xi < a\} &\geq 1 - \frac{1 - \mathrm{E}[\theta]}{a} \\
\Pr\{1 - \theta < a\} &\geq \frac{a-1}{a} + \frac{\mathrm{E}[\theta]}{a}
\end{aligned}$$

At this point, we use the fact that $\mathrm{E}[\theta] \geq 1/4$ to obtain: $\Pr\{\theta > 1 - a\} \geq \frac{a-1}{a} + \frac{1}{4a}$. Now if we substitute $1 - a = 1/8$, or $a = 7/8$, then we obtain:

$$\Pr\{\theta > 1/8\} \geq 1/7.$$

Exercise 5.2

The first algorithm, the one that picks only blood pressure and the BMI as features, is simpler in the sense that the hypothesis class to be learned in simpler. We would expect that this algorithm has a higher bias but a lower variance when compared to the second algorithm which is more feature rich. The second algorithm would probably explain the conditions of a heart attack better as it includes relevant features such as age and the level of physical activity into account. We would expect the second algorithm to have a lower bias but a higher variance because there may be a tendency to overfit on any given sample.

Since the sample complexity is higher for a more complicated hypothesis class, if the sample size is "small," then we might want to choose the first algorithm. If sample size is not a problem, then the second algorithm is probably better.