Understanding Machine Learning: Exercises

Somnath Sikdar

June 29, 2020

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 2

Finite Hypothesis Classes

2.1 Setting

Consider a classification problem in which the learning algorithm receives as input a sequence of training examples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{0, 1\}$. The sequence of training examples is drawn iid from some unknown distribution $\mathcal{D}$ and labeled by some target function $f \colon \mathcal{X} \to \mathcal{Y}$.

Given a hypothesis $h \colon \mathcal{X} \to \mathcal{Y}$, the *true error* or the *generalization error* $L_{\mathcal{D},f}(h)$ of the hypothesis is defined to be:

$$\begin{aligned} L_{\mathcal{D},f}(h) &= \mathbf{Pr}_{x \sim \mathcal{D}} \{h(x) \neq f(x)\} \\ &= 1 \cdot \mathbf{Pr}_{x \sim \mathcal{D}} \{h(x) \neq f(x)\} + 0 \cdot \mathbf{Pr}_{x \sim \mathcal{D}} \{h(x) = f(x)\} \\ &= \mathbf{E}_{\mathcal{D}} [|h - f|] \,. \end{aligned}$$

The generalization error is the expected number of points in the domain at which the hypothesis h differs from the true labeling function f, the expectation being calculated wrt the distribution $\mathcal{D}$. The learning algorithm does not directly know the true error. What it can calculate is the *training error* $L_S(h)$ which is defined as:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{h(x_i) \neq y_i},$$

where $1_{h(x_i) \neq y_i} = 1$ if $h(x_i) \neq y_i$ and 0 otherwise. Note that the training error is the expected number of points x at which $h(x)$ differs from the true label y wrt the uniform distribution.

The empirical risk minimization (ERM) paradigm is a learning paradigm where the learner, when given a training sample S, comes up with a hypothesis h_S that minimizes the training error on S. That is, $h_S = \operatorname{argmin}_h L_S(h)$. We may constrain the ERM algorithm to a specific class of hypotheses $\mathcal{H}$. In this case, the ERM algorithm is forced to output an element $h_S \in \mathcal{H}$ where $h_S = \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$.

2.2 Finite Hypothesis Classes

We assume that we have a finite hypothesis class $\mathcal{H}$. For a training sample S labeled by some function $f : \mathcal{X} \to \mathcal{Y}$, let h_S be the hypothesis output by ERM$_{\mathcal{H}}$ when applied to S. In this chapter, we assume that there exists a hypothesis $h^\star$ such that $L_{\mathcal{D},f}(h^\star) = 0$. In particular, this means that for any hypothesis h_S output by the ERM$_{\mathcal{H}}$, we have $L_S(h_S) = 0$ (since $\min_{h \in \mathcal{H}} L_S(h) = 0$).

We wish to upper bound the generalization error $L_{\mathcal{D},f}(h_S)$ of the hypothesis output by ERM$_{\mathcal{H}}$ on the sample S. Note that the hypothesis h_S depends on the sample S and the only way we can connect the sample S to the distribution $\mathcal{D}$ is by making an assumption on how it was generated. The standard assumption is that S is generated i.i.d according to the distribution $\mathcal{D}$. Thus h_S is a random variable and is potentially different for different training samples. Therefore when we talk about bounding the generalization error of the output hypothesis, we must talk about the fraction of samples for which this is possible.

Thus bounding the generalization error of the output hypothesis h involves the specification of two parameters: a *confidence parameter* $\delta \in (0, 1)$ that measures for what fraction of training samples ERM$_{\mathcal{H}}$ outputs a hypothesis h that does *not* generalize well; and an *accuracy parameter* $\varepsilon \in (0, 1)$ that specifies how badly off the generalization accuracy is. In other "words," we want to bound the probability that $L_{\mathcal{D},f}(h_S) > \varepsilon$, since this represents the case where the hypothesis h_S does not generalize well. The probability is calculated w.r.t samples that are chosen i.i.d from the distribution $\mathcal{D}$. That is, we are looking at the following condition:

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_{\mathcal{D},f}(h_S) > \varepsilon\} < \delta. \tag{2.1}$$

To help analyze this probability, define $\mathcal{H}_B$, the set of *bad hypotheses*, to be the subset of hypotheses of $\mathcal{H}$ that have a generalization error exceeding ε.

$$\mathcal{H}_B = \{h \in \mathcal{H} \colon L_{\mathcal{D},f}(h) > \varepsilon\}.$$

Also define the set M of *misleading samples* to be:

$$M = \{S \colon \exists h \in \mathcal{H}_B \text{ s.t. } L_S(h) = 0\}.$$

That is, a sample is misleading if there exists a bad hypothesis that appears to be good on it.

Since we assume realizability, any hypothesis h_S picked up by ERM$_{\mathcal{H}}$ in response to a training sample S has zero training error. Now if $L_{\mathcal{D},f}(h) > \varepsilon$, then we must have $h_S \in \mathcal{H}_B$. Moreover, since $L_S(h_S) = 0$, it must be that $S \in M$. This shows that $\{S \colon L_{\mathcal{D},f}(h_S) > \varepsilon\} \subseteq M$ and that:

$$\begin{aligned}\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_{\mathcal{D},f}(h_S) > \varepsilon\} &\leq \mathbf{Pr}_{S \sim \mathcal{D}^m} \{S \in M\} \\ &= \mathbf{Pr}_{S \sim \mathcal{D}^m} \{\exists h \in \mathcal{H}_B \colon L_S(h) = 0\} \\ &\leq \sum_{h \in \mathcal{H}_B} \mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_S(h) = 0\}, \end{aligned} \tag{2.2}$$

where the last inequality follows from the union bound.

Fix a hypothesis $h \in \mathcal{H}_B$. Then

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_S(h) = 0\} = \prod_{i=1}^{m} \mathbf{Pr}_{x_i \sim \mathcal{D}} \{h(x_i) = f(x_i)\}.$$

Since $L_{\mathcal{D},f}(h) > \varepsilon$, we have that:

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_S(h) = 0\} = (1 - \varepsilon)^m \leq e^{-\varepsilon m}.$$

From (2.2), we get that:

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_{\mathcal{D},f}(h_S) > \varepsilon\} \leq |\mathcal{H}_B| \cdot e^{-\varepsilon m} \leq |\mathcal{H}| \cdot e^{-\varepsilon m}. \tag{2.3}$$

Let us require that $|\mathcal{H}| \cdot e^{-\varepsilon m} < \delta$. This implies that as long as $m > (1/\varepsilon) \cdot \log(|\mathcal{H}|/\delta)$, we have that $\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_{\mathcal{D},f}(h) > \varepsilon\} < \delta$. In words, the hypothesis h_S output by $\mathrm{ERM}_{\mathcal{H}}$ is such that its generalization error on at least $1 - \delta$ fraction of the samples is at most ε.

Chapter 3

A Formal Learning Model

Notes on Chapter 3

The main concept introduced here is that of agnostic PAC learnability. It helps to review the definitions of both PAC learnability with the realizability assumption and that of agnostic PAC learnability.

Definition 3.1 (PAC Learnability)**.** Fix a domain $\mathcal{X}$, a range $\mathcal{Y}$ and let $\mathcal{H}$ be a set of functions from $\mathcal{X} \to \mathcal{Y}$. The class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} \colon (0,1) \times (0,1) \to \mathbf{N}$ and a learning algorithm $\mathcal{A}$ such that the following holds: for all $\varepsilon, \delta \in (0,1)$, all labeling functions $f \colon \mathcal{X} \to \mathcal{Y}$ and all distributions $\mathcal{D}$ over $\mathcal{X}$ such that $\mathcal{H}$ is realizable wrt $\mathcal{D}$ and f, if $\mathcal{A}$ is presented with a sample of at least $m_{\mathcal{H}}(\varepsilon, \delta)$ examples drawn iid from $\mathcal{D}$, $\mathcal{A}$ returns a hypothesis h_S such that

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \left\{ L_{\mathcal{D},f}(h_S) \leq \varepsilon \right\} > 1 - \delta.$$

In the definition above, the sample complexity function $m_{\mathcal{H}}$ must work for all possible labeling functions f and distributions $\mathcal{D}$ as long as $\mathcal{H}$ is realizable (wrt this labeling function and distribution). A learning task is completely specified by $(\mathcal{X}, \mathcal{Y}, f, \mathcal{D})$. Intuitively, what this definition states is that a hypothesis class is PAC learnable if there exists a learner which when given a sufficiently large number of training examples can approximate the true labeling function f with high probability for all learning tasks on the domain $\mathcal{X} \times \mathcal{Y}$ that satisfy the realizability condition.

Definition 3.2 (Agnostic PAC Learnability)**.** A class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $m_{\mathcal{H}} \colon (0,1) \times (0,1) \to \mathbf{N}$ and a learning algorithm $\mathcal{A}$ such that the following holds: for all $\varepsilon, \delta \in (0,1)$ and all distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ such that, if $\mathcal{A}$ is presented with a sample of at least $m_{\mathcal{H}}(\varepsilon, \delta)$ examples drawn iid from $\mathcal{D}$, $\mathcal{A}$ returns a hypothesis h_S such that

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

In the agnostic setting, a learning task is completely specified by the triplet $(\mathcal{X}, \mathcal{Y}, \mathcal{D})$ and a hypothesis class is agnostic PAC learnable if there exists a learner which when given a sufficiently many training examples outputs a “good enough” hypothesis with high probability. The goodness of the hypothesis is measured with respect to the best hypothesis of the class $\mathcal{H}$.

Agnostic PAC learnability presents a *stronger* requirement as one must be able to learn *any* distibution as distinct from PAC learnability where one must be able to learn distibutions for which the hypothesis class is realizable. As such, agnostic PAC learnability implies PAC learnability.

Exercise 3.1

Let $m_{\mathcal{H}}(\varepsilon, \delta)$ be the sample complexity of a PAC-learnable hypothesis class $\mathcal{H}$ for a binary classification task. For a fixed δ, let $0 < \varepsilon_1 \leq \varepsilon_2 < 1$ and suppose that $m_{\mathcal{H}}(\varepsilon_1, \delta) < m_{\mathcal{H}}(\varepsilon_2, \delta)$. Then when running the learning algorithm on $m_{\mathcal{H}}(\varepsilon_1, \delta)$ i.i.d examples, we obtain a hypothesis h, which with probability at least $1 - \delta$ has a true error $L_{\mathcal{D},f}(h) \leq \varepsilon_1 \leq \varepsilon_2$. 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 $m_{\mathcal{H}}(\varepsilon_2, \delta)$. This contradicts the minimality of the sample complexity function. Hence we must have $m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta)$.

Next suppose that $0 < \delta_1 \leq \delta_2 < 1$ and that $m_{\mathcal{H}}(\varepsilon, \delta_1) < m_{\mathcal{H}}(\varepsilon, \delta_2)$, where ε is fixed in advance. Then with $m_{\mathcal{H}}(\varepsilon, \delta_1)$ 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 $m_{\mathcal{H}}(\varepsilon, \delta_2)$. This again contradicts the minimality of the sample complexity function. Hence we must have $m_{\mathcal{H}}(\varepsilon, \delta_1) \geq m_{\mathcal{H}}(\varepsilon, \delta_2)$.

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 $e^{-m p_{z^*}} < \delta$, 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 $\varepsilon \geq p_{z^*}$, then even a sample of size one will guarantee that the true error of h_s is at most ε. However if $\varepsilon < p_{z^*}$ then we can then use this in (3.1) to obtain:

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

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

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:

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

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

$$\varepsilon = \mathbf{Pr}_{(x,y)\sim\mathcal{D}}\left\{(x,y) \in C_r \setminus C_t\right\}.$$

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 $(1-\varepsilon)^m \leq e^{-\varepsilon m}$. This is the probability that we are handed a "bad" sample. Upper bounding this by δ, we obtain that $m > \log(1/\delta)/\varepsilon$.

Now a sample of size at least $\log(1/\delta)/\varepsilon$ 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 $\lceil \log(1/\delta)/\varepsilon \rceil$.

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

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

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

```
procedure PACBoolean($S$)    ▷ $S$ 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

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 $\varepsilon > 0$. As in the text, define the set of bad hypotheses to be $\mathcal{H}_B = \{h \in \mathcal{H} \colon L_{\bar{\mathcal{D}}_m, f}(h) > \varepsilon\}$ 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 $L_{\bar{\mathcal{D}}_m,f}(h) > \varepsilon$ and $L_S(h) = 0$ happens only when $S|_x \in \mathcal{M}$. Hence,

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

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 $|\mathcal{H}| \cdot (1-\varepsilon)^m$ which is at most $|\mathcal{H}| \cdot e^{-\varepsilon m}$.

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 $\varepsilon, \delta > 0$. Since $\mathcal{H}$ is agnostic PAC-learnable, there exists a learner A which given a sample S of $m \ge m_{\mathcal{H}}(\varepsilon,\delta)$ iid examples generated by $\mathcal{D}$ returns a hypothesis h_S such that

$$\mathbf{Pr}_{S \sim \mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) \le \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon\right\} > 1 - \delta.$$

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

$$\begin{aligned}
L_{\mathcal{D}}(h') &= \mathbf{Pr}_{(x,y) \in \mathcal{D}}\{h'(x) \ne y\} \\
&= \mathbf{Pr}_{(x,y) \in \mathcal{D}}\{h'(x) \ne f(x)\} \\
&= L_{\mathcal{D}_{\mathcal{X}},f}(h').
\end{aligned}$$

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:

$$\mathbf{Pr}_{S|_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m}\{L_{\mathcal{D}_{\mathcal{X}},f}(h_S) \le \varepsilon\} > 1 - \delta,$$

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; let $\mathcal{D}_X$ be the marginal distribution of X over $\mathcal{X}$; and, finally, let $\eta(x) = \mathbf{Pr}_{\mathcal{D}_{Y|X}} \{Y = 1 \mid X = x\}$.

Using this notation, 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

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

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

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

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

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

Since $\mathbf{1}_{f_{\mathcal{D}}(x)=0} = 1 - \mathbf{1}_{f_{\mathcal{D}}(x)=1}$ and $\mathbf{1}_{g(x)=0} = 1 - \mathbf{1}_{g(x)=1}$, 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:

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

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 $\mathbf{Pr}\{g(x) \neq Y \mid X = x\} - \mathbf{Pr}\{f_{\mathcal{D}}(x) \neq Y \mid X = x\}$ 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 $\mathcal{H} \times Z \to \mathbf{R}_+$. Let $\mathcal{D}$ be a distribution over the domain Z. The risk of a hypothesis $h \in \mathcal{H}$ is

$$L_{\mathcal{D}}(h) = \mathbf{E}_{z \sim \mathcal{D}}\left[l(h, z)\right]$$

The empirical risk of a hypothesis h w.r.t a sample $S = \{z_i\}_{i=1}^m$ is defined as:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$$

Definition 4.1. A training set S is ε-representative w.r.t the domain Z, the hypothesis class $\mathcal{H}$, the distribution $\mathcal{D}$ and the loss function l if for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$.

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 $\text{ERM}_{\mathcal{H}}(S)$ learning rule is guaranteed to return a good hypothesis. More specifically,

Lemma 4.1. *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 $\varepsilon/2$-representative sample. Then any output h_S of* $\text{ERM}_{\mathcal{H}}(S)$ *satisfies*

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

Proof. The output h_S of $\text{ERM}_{\mathcal{H}}(S)$ is such that

$$h_S = \text{argmin}_{h \in \mathcal{H}} L_S(h).$$

Since S is $\varepsilon/2$ representative,

$$\begin{aligned}
L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\varepsilon}{2} \\
&\leq \min_{h\in\mathcal{H}} L_S(h) + \frac{\varepsilon}{2} \\
&\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.
\end{aligned}$$

In the above derivation, the second inequality follows from the fact that h_S minimizes the empirical risk among all hypotheses in $\mathcal{H}$; the third inequality follows from the fact that S is $\varepsilon/2$-representative and hence $L_S(h) \leq L_{\mathcal{D}}(h) + \varepsilon/2$ for all $h \in \mathcal{H}$. □

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 $\varepsilon/2$-representative training sample. This requirement is baked into the definition of *uniform convergence*.

Definition 4.2. 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 $\varepsilon, \delta \in (0,1)$ and all distributions $\mathcal{D}$ on Z, if a sample with at least $m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon,\delta)$ examples is chosen i.i.d from $\mathcal{D}$, then with probability $1-\delta$, the sample is ε-representative.

By Lemma (4.1), if $\mathcal{H}$ is uniformly convergent with function $m_{\mathcal{H}}^{\mathrm{UC}}$, then it is agnostically PAC-learnable with sample complexity $m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon/2,\delta)$. In this case, the ERM paradigm is a successful agnostic PAC-learner for $\mathcal{H}$.

Corollary 4.1. *If a class $\mathcal{H}$ is uniformly convergent with sample complexity function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC-learnable with sample complexity $m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2,\delta)$.*

The other main result is that all finite hypothesis classes are uniformly convergent and hence agnostic PAC learnable.

Theorem 4.1. *Let $Z = \mathcal{X} \times \mathcal{Y}$ be a domain, $\mathcal{H} = \{h \colon \mathcal{X} \to \mathcal{Y}\}$ be a finite hypothesis class and let $l \colon \mathcal{H} \times Z \to [0,1]$ be a loss function that is bounded so that $a \leq l(h,z) \leq b$ for all $h \in \mathcal{H}$ and $z \in Z$ for some $a, b \in \mathbf{R}_+$. Then $\mathcal{H}$ is uniformly convergent with sample complexity*

$$m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \leq \frac{(b-a)^2 \cdot \log(2|\mathcal{H}|/d)}{2\varepsilon^2}.$$

Furthermore, $\mathcal{H}$ is agnostically PAC-learnable with sample complexity

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq \frac{2\cdot(b-a)^2 \cdot \log(2|\mathcal{H}|/d)}{\varepsilon^2}.$$

Proof. It is sufficient to show that $\mathcal{H}$ is uniformly convergent under these conditions. Fix $\varepsilon, \delta > 0$. Let $\mathcal{D}$ be a distribution on Z. We wish to show that there exists $m \in \mathbf{N}$ such that when a sample S from Z is picked i.i.d according to $\mathcal{D}$

$$\mathbf{Pr}_{S\sim\mathcal{D}^m}\left\{S \colon \forall\, h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\right\} \geq 1-\delta.$$

This probability is equivalent to saying that

$$\mathbf{Pr}_{S\sim\mathcal{D}^m}\left\{S\colon \exists\, h\in\mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\right\}<\delta.$$

Moreover, we can upper bound this probability using the Union Bound as follows:

$$\mathbf{Pr}_{S\sim\mathcal{D}^m}\left\{S\colon \exists\, h\in\mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\right\}\leq\sum_{h\in\mathcal{H}}\mathbf{Pr}_{S\sim\mathcal{D}^m}\left\{S\colon |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\right\}. \quad (4.1)$$

Note that we are using the fact that $\mathcal{H}$ is finite here.

Recall that $L_{\mathcal{D}}(h)=\mathbf{E}_{z\sim\mathcal{D}}\left[l(h,z)\right]$ and that $L_S(h)=\frac{1}{m}\sum_{i=1}^m l(h,z_i)$. Since the points z_i are sampled i.i.d according to $\mathcal{D}$, the random variables $l(h,z_i)$ have expectation $L_{\mathcal{D}}(h)$. The expectation of $L_S(h)$ by the linearity of expectation is also $L_{\mathcal{D}}(h)$. Thus the quantity $|L_S(h)-L_{\mathcal{D}}(h)|$ is the deviation of a random variable $L_S(h)$ from its expected value. By the Law of Large Numbers, as $m\to\infty$, $|L_S(h)-L_{\mathcal{D}}(h)|\to 0$. Thus the intuition why large-enough samples are ε-representative follows directly from this. However the Law of Large Numbers is an asymptotic result and in order to be able to provide a bound on the deviation between an empirically estimated error and its true value for a finite sample size, we need a measure concentration inequality.

Lemma 4.2 (Hoeffding's Inequality)**.** *Let $\theta_1,\dots,\theta_m$ be a sequence of i.i.d random variables. Suppose that for all i, $\mathbf{E}\left[\theta_i\right]=\mu$ and $\mathbf{Pr}\left\{a\leq\theta_i\leq b\right\}=1$. Then for any $\varepsilon>0$*

$$\mathbf{Pr}\left\{\left|\frac{1}{m}\sum_{i=1}^m\theta_i-\mu\right|>\varepsilon\right\}\leq 2\exp\left\{-\frac{2m\varepsilon^2}{(b-a)^2}\right\}.$$

Applying this to our case, we let $\theta_i=l(h,z_i)$ and $\mu=L_{\mathcal{D}}(h)$. We also assumed that the loss function values lie in the interval $[a,b]$. Then by Hoeffding's Inequality, we have that

$$\mathbf{Pr}_{S\sim\mathcal{D}^m}\left\{S\colon |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\right\}\leq 2\exp\left\{-\frac{2m\varepsilon^2}{(b-a)^2}\right\}. \quad (4.2)$$

Plugging this in equation (4.1), we obtain:

$$\mathbf{Pr}_{S\sim\mathcal{D}^m}\left\{S\colon \exists h\in\mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)|>\varepsilon\right\}\leq 2\cdot|\mathcal{H}|\cdot\exp\left\{-\frac{2m\varepsilon^2}{(b-a)^2}\right\}. \quad (4.3)$$

Choose m sufficiently large so that

$$2\cdot|\mathcal{H}|\cdot\exp\left\{-\frac{2m\varepsilon^2}{(b-a)^2}\right\}\leq\delta.$$

Simplifying, this gives us the bound:

$$\frac{(b-a)^2}{2\varepsilon^2}\log\left(\frac{2\cdot|\mathcal{H}|}{\delta}\right)\leq m.$$

This gives us the bound on $m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon,\delta)$:

$$m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon,\delta)\leq\frac{(b-a)^2\cdot\log(2|\mathcal{H}|/d)}{2\varepsilon^2}.$$

The bound on $m_{\mathcal{H}}(\varepsilon,\delta)$ follows from Corollary 4.1. □

Exercise 4.1

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

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \left\{ L_{\mathcal{D}}(h_S) > \varepsilon_n \right\} < \delta_n.$$

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

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

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 $L_{\mathcal{D}}(h_S) > \varepsilon_n$ by $1 - \varepsilon_n$. As $n \to \infty$, $m \to \infty$ and $\mathbf{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(h_s) \right] \to 0$, proving that (2) follows.

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

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

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

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

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

$$\varepsilon \cdot \mathbf{Pr}_{S \sim \mathcal{D}^m} \left\{ L_{\mathcal{D}}(h_S) \geq \varepsilon \right\} < \varepsilon \cdot \delta.$$

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 $\leq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$. Since $|\mathcal{H}| = 2^{|\mathcal{X}|}$, the sample complexity is $\frac{|\mathcal{X}|+\log(1/\delta)}{\varepsilon} = O(|\mathcal{X}|)$.

Let us suppose that A is a learning algorithm for the task of binary classification w.r.t 0-1 loss over the domain $\mathcal{X}$. Furthermore, assume that A has no access to any prior knowledge in the sense that the hypothesis class from which it chooses its hypotheses is the set of all functions from $\mathcal{X} \to \{0,1\}$. 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." However, given specific prior knowledge in the form of a hypothesis class that contains this function f, the ERM algorithm is a successful learner.

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. Firstly, it shows that there is no universal learner in the sense of a learning algorithm that succeeds on learning tasks without prior information. Secondly, it shows that arbitrary hypothesis classes are not PAC-learnable.

Theorem 5.1. *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 exist a function $f\colon \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ 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 $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Corollary 5.1. *Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all boolean functions on $\mathcal{X}$. Then $\mathcal{H}$ is not PAC-learnable.*

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 $\mathbf{E}[\theta] \geq 1/4$. We want to show that $\mathbf{Pr}\{\theta \geq 1/8\} \geq 1/7$.

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

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

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 $\mathbf{E}[\xi] = 1 - \mathbf{E}[\theta]$, and hence by Markov's inequality,

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

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

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

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.

Chapter 6

The VC-Dimension

6.1 Notes on Chapter 6

We know that finite hypothesis classes are agnostic PAC learnable (and hence PAC learnable). What about infinite hypothesis classes? The first example is that of an infinite hypothesis class that is PAC learnable.

Example 6.1 (Threshold Functions)**.** Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}$. For $r \in [0,1]$, define $h_r \colon \mathcal{X} \to \mathcal{Y}$ as:

$$h_r(x) = \begin{cases} 0 & \text{if } x \leq r \\ 1 & \text{if } x > r \end{cases}$$

Let $\mathcal{H}_{\text{thr}}$ be the set of all threshold functions h_r for $r \in [0,1]$. Since $\mathcal{H}_{\text{thr}}$ is not finite, it is not immediately obvious whether it is PAC learnable (in the realizable case).

Fix $\varepsilon, \delta \in (0,1)$. Let $f = h_s$ be the true labeling function where $s \in [0,1]$ and let $\mathcal{D}$ be the underlying distribution over the domain $[0,1]$. Let $s_0 \in [0,s)$ and $s_1 \in [s,1]$ be numbers such that

$$\mathcal{D}\{x \in [s_0, s)\} = \varepsilon = \mathcal{D}\{x \in [s, s_1]\}$$

If $\mathcal{D}\{[0,s)\} < \varepsilon$, then set $s_0 = 0$; similarly, if $\mathcal{D}\{[s,1]\} < \varepsilon$, set $s_1 = 1$. Since $\mathcal{D}$ is a distribution, it must place a probability mass of ε either to the left or to the right of s.

Given a sample S, let $t_0 = \max\{t \colon (t,0) \in S\}$ and $t_1 = \min\{t \colon (t,1) \in S\}$. The ERM algorithm outputs h_p, where $p \in (t_0, t_1)$. In particular, if the sample presented to the ERM algorithm is such that $s_0 \leq t_0$ and $t_1 \leq s_1$, then hypothesis h_p returned by the ERM algorithm will incur a loss of $L_{\mathcal{D}}(h_p) \leq \varepsilon$.

Thus the probability that the hypothesis $\text{ERM}(S)$ output by the ERM algorithm has a loss greater than ε on a sample S of size m is:

$$\begin{aligned}
\textbf{Pr}_{S\sim\mathcal{D}^m}\{L_{\mathcal{D}}(\text{ERM}(S)) > \varepsilon\} &= \textbf{Pr}_{S\sim\mathcal{D}^m}\{S \colon t_0 < s_0 \vee s_1 < t_1\} \\
&\leq \textbf{Pr}_{S\sim\mathcal{D}^m}\{S \colon S|_x \cap [s_0, s) = \varnothing\} + \textbf{Pr}_{S\sim\mathcal{D}^m}\{S \colon S|_x \cap [s, s_1] = \varnothing\} \\
&\leq 2 \cdot (1-\varepsilon)^m \\
&\leq 2 \cdot e^{-\varepsilon m}
\end{aligned}$$

Setting the last expression to be at most δ, we obtain that $m > \frac{1}{\varepsilon} \cdot \log \frac{2}{\delta}$. Hence if we have samples of size at least $\frac{1}{\varepsilon} \cdot \log \frac{2}{\delta}$,

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_{\mathcal{D}}(\mathrm{ERM}(S)) \leq \varepsilon\} \geq 1 - \delta,$$

which is the condition for PAC learnability.

The second example shows that there are infinite hypothesis classes that are not PAC learnable at least by using an ERM strategy.

Example 6.2 (Identity Function for Finite Sets)**.** Let $\mathcal{X} = \mathbf{R}$ and $\mathcal{Y} = \{0, 1\}$. Given a set $A \subseteq \mathcal{X}$, define h_A as follows:

$$h_A = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathcal{H}_{\text{finite}}$ be the set of all such functions h_A for *finite* subsets A of $\mathbf{R}$ along with the function h_1 which maps every point in $\mathbf{R}$ to 1. We claim that $\mathcal{H}_{\text{finite}}$ is not PAC learnable by an ERM algorithm.

Consider the case when the true labeling function $f = h_1$, the all-ones function on $\mathbf{R}$ and $\mathcal{D}$ is the uniform distribution on $[0, 1]$. Since $f \in \mathcal{H}_{\text{finite}}$, we are assuming that the hypothesis class is realizable. Fix any sample size m. A sample S in this case looks like $\{(x_1, 1), \ldots, (x_m, 1)\}$ and an obvious ERM strategy is to output h_A for $A = \{x_1, \ldots, x_m\}$. Clearly $L_S(h_A) = 0$ but $L_{\mathcal{D}}(h_A) = 1$.

The previous examples show that the size of the hypothesis class does not characterize whether it is learnable. This characterization is provided by the so-called VC-dimension.

6.2 The VC Dimension

To motivate the definition of VC-dimension, we note that the proof of the No-Free-Lunch Theorem relied on the fact that there exists a finite subset $C \subset \mathcal{X}$ such that the adversary could choose a target function from the set of *all* possible functions from $C \to \{0, 1\}$. This leads us to the following definition

Definition 6.1 (Restriction of $\mathcal{H}$ to C)**.** Let $\mathcal{H}$ be the set of functions from $\mathcal{X} \to \{0, 1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. Then the restriction of $\mathcal{H}$ to C is the set of functions $\mathcal{H}_C$ from $C \to \{0, 1\}$ that can be derived from $\mathcal{H}$.

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) \colon h \in \mathcal{H}\}.$$

It turns out that whether a hypothesis class is learnable or not $\mathcal{H} = \{h \colon \mathcal{X} \to \{0, 1\}\}$ can be characterized by studying the restriction of the hypothesis class to finite subsets $C \subset \mathcal{X}$. This leads us to the next definition.

Definition 6.2 (Shattering)**.** A hypothesis class $\mathcal{H} = \{h \colon \mathcal{X} \to \{0, 1\}\}$ shatters a set $C \subset \mathcal{X}$ if its restriction to C is the set of all boolean functions on C.

We can now connect the notion of shattering to the No-Free-Lunch Theorem. For the proof of Theorem 5.1 to go through, we require a hypothesis class that given any sample size m shatters some set $C \subset \mathcal{X}$ of size $2m$. In the proof given, this hypothesis class is the set of all boolean functions on $\mathcal{X}$ but this can now be relaxed.

Corollary 6.1. *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$. Assume that for any $m \in \mathbf{N}$ representing a training set size, there exists a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$. Then for any learning algorithm A there exists a function $f \colon \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that*

1. $L_{\mathcal{D}}(f) = 0$
2. $\mathbf{Pr}_{S \in \mathcal{D}^m} \{L_{\mathcal{D}}(A(S)) \geq 1/8\} > 1/7.$

Thus if $\mathcal{H}$ shatters a set of size $2m$, it cannot be PAC-learned using m examples. Since the size of a set shattered by a hypothesis class plays a definitive role in whether it can be PAC-learned or not, the next defintion follows naturally.

Definition 6.3. The VC-dimension of a hypothesis class, VCdim$(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrary size, VCdim$(\mathcal{H}) = \infty$.

As a direct consequence of Corollary 6.1, we have that:

Theorem 6.1. *If a hypothesis class $\mathcal{H}$ has infinite VC-dimension, then it is not PAC-learnable.*

6.3 The Fundamental Theorem of Statistical Learning

The fundamental theorem of statistical learning goes as follows.

Theorem 6.2 (The Fundamental Theorem)**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X} \to \{0,1\}$ and let the loss be the* 0-1 *loss. Then the following are equivalent:*

1. *$\mathcal{H}$ is uniformly convergent*
2. *$\mathcal{H}$ is agnostic PAC-learnable*
3. *$\mathcal{H}$ is PAC-learnable*
4. *$\mathcal{H}$ has finite VC-dimension.*

There is also a "quantitative" version of this theorem wherein the sample complexities are made explicit in terms of the VCdim$(\mathcal{H})$.

Theorem 6.3 (Quantitative Version)**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X} \to \{0,1\}$ and let the loss be the* 0-1 *loss. Assume that VCdim$(\mathcal{H}) = d < \infty$. Then there are absolute constants C_1, C_2 such that:*

1. *$\mathcal{H}$ is uniformly convergent with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}$$
2. *$\mathcal{H}$ is agnostic PAC-learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}$$

3. *$\mathcal{H}$ is PAC-learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq C_2 \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$$

Proof of Theorem 6.2. We saw in Chapter 4 that if a hypothesis class is unifomly convergent then it is agnostic PAC-learnable. An agnostic PAC-learnable class is PAC-learnable. Moreover, PAC-learnability implies that the hypothesis class has finite VC-dimension. The only thing that remains to show is that finite VC-dimension implies uniform convergence. This makes use of Sauer's Lemma. □

6.3.1 Sauer's Lemma and the Growth Function

The growth function measures the size of a hypothesis class when it is restricted to a finite subset of the domain.

Definition 6.4. The growth function of a hypothesis class $\mathcal{H}$ is a function $\tau \colon \mathbf{N} \to \mathbf{N}$ defined as follows:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X} \colon |C| = m} |\mathcal{H}_C|.$$

Thus the growth function is the maximum number of different functions that can be obtained when restricting $\mathcal{H}$ to an m-sized subset of the domain.

For hypothesis classes of finite VC-dimension, Sauer's Lemma bounds the growth function in terms of the VC-dimension and m. If VCdim$(\mathcal{H}) = d$, then for all $m \leq d$ there exists an m-sized subset of the domain that is shattered by $\mathcal{H}$; consequently, $\tau_{\mathcal{H}}(m) = 2^m$. What Sauer's Lemma shows is that for $m > d$, $\tau_{\mathcal{H}}(m) \leq (em/d)^d$. That is, for sets $C \subset \mathcal{X}$ of size larger than VCdim$(\mathcal{H})$, $|\mathcal{H}_C|$ is a polynomial in $|C|$. Thus the VC-dimension marks the point where $|\mathcal{H}_C|$ transitions from an exponential function of $|C|$ to a polynomial function of $|C|$.

Lemma 6.1 (Sauer's Lemma)**.** *Let $\mathcal{H}$ be a hypothesis class with VCdim$(\mathcal{H}) = d < \infty$. Then for all $m \in \mathbf{N}$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, for $m > d$, $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.*

The next result ties the growth function to ε-representativeness of samples.

Theorem 6.4. *Let $\mathcal{H} = \{h \colon h \colon \mathcal{X} \to \{0,1\}\}$ and let $\tau_{\mathcal{H}}$ be its growth function. Then for every distribution $\mathcal{D}$ over $\mathcal{X}$ and every $\delta \in (0,1)$ we have that*

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \left\{ \forall h \in \mathcal{H} \colon |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}} \right\} \geq 1 - \delta.$$

Assuming that VCdim$(\mathcal{H}) = d < \infty$ and $m > d$, we have $\tau_{\mathcal{H}}(m) \leq (em/d)^d$. Substituting this in the righthand side of the above expression it can be shown that for

$$m \geq 4 \frac{2d}{(\delta\varepsilon)^2} \log \frac{2d}{(\delta\varepsilon)^2} + \frac{4d \log(2e/d)}{(\delta\varepsilon)^2}$$

it holds that

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{ \forall h \in \mathcal{H} \colon |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon \} \geq 1 - \delta.$$

This then shows that hypothesis classes of finite VC-dimension are uniformly convergent, completing the proof of Theorem 6.2.

Exercise 6.1

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\{0,1\}$ and let $\mathcal{H}' \subseteq \mathcal{H}$. Assume that VCdim$(\mathcal{H}') >$ VCdim$(\mathcal{H})$. Then there exists a set $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}'$ but not by $\mathcal{H}$. This implies that for all $g\colon C \to \{0,1\}$ there exists $h' \in \mathcal{H}'$ such that $g(x) = h'(x)$ for all $x \in C$. Since $h' \in \mathcal{H}$, this implies that $\mathcal{H}$ shatters C, a contradiction.

Exercise 6.2

In this exercise, $\mathcal{X}$ is finite and $k \leq |\mathcal{X}| =: n$.

6.2.1

We claim that

$$\text{VCdim}(\mathcal{H}_{=k}) = \begin{cases} k & \text{if } k \leq \lfloor n/2 \rfloor \\ n-k & \text{if } k > \lfloor n/2 \rfloor \end{cases}$$

Suppose that $k \leq \lfloor n/2 \rfloor$ and consider a subset $C \subset \mathcal{X}$ of size $k+1$. Then the all-one function on C cannot be extended to a function in $\mathcal{H}_{=k}$ as it maps $k+1$ elements of $\mathcal{X}$ to 1. Hence VCdim$(\mathcal{H}_{=k}) \leq k$. If $|C| = k$ and $g\colon C \to \{0,1\}$ that maps k' elements of C to 1, we can extend g to a function on $\mathcal{X}$ that maps exactly k elements of $\mathcal{X}$ to 1. This shows that VCdim$(\mathcal{H}_{=k}) \geq k$. Hence VCdim$\mathcal{H}_{=k} = k$.

Now consider the case $k > \lfloor n/2 \rfloor$. If C is subset of size $n-k+1$, then the all-zero function on C cannot be extended to a function in $\mathcal{H}_{=k}$. This happens because there are only $n-(n-k+1) < k$ elements in $\mathcal{X} \setminus C$. Hence VCdim$(\mathcal{H}_{=k}) \leq n-k$. If $|C| = n-k$ and $g\colon C \to \{0,1\}$ that assigns 1 to k' elements of C, then we can extend g to a function in $\mathcal{H}_{=k}$ as we have at least $k-k'$ elements in $\mathcal{X} \setminus C$ which we can map to 1. This shows that VCdim$(\mathcal{H}_{=k}) \geq n-k$. Hence VCdim$(\mathcal{H}_{=k}) = n-k$.

6.2.2

First observe that if $k \geq \lfloor n/2 \rfloor$, then $\mathcal{H}_{\leq k}$ includes all possible functions from $\mathcal{X}$ to $\{0,1\}$. This is because any function $g\colon \mathcal{X} \to \{0,1\}$ maps at most half the elements of $\mathcal{X}$ to either 0 or 1 and hence is in $\mathcal{H}_{\leq k}$. Hence in this case every subset of $\mathcal{X}$ is shattered by $\mathcal{H}_{\leq k}$ and VCdim$(\mathcal{H}_{\leq k}) = n$.

If $k < \lfloor n/2 \rfloor$, then we claim that VCdim$(\mathcal{H}_{\leq k}) = 2k+1$. Let $C \subset \mathcal{X}$ of size $2k+1$ and consider a function $g\colon C \to \{0,1\}$. Such a function maps at most k elements to either 0 or 1. Suppose that it maps at most k elements to 1. Extend g to a function on $\mathcal{X}$ by mapping all elements of $\mathcal{X} \setminus C$ to 0. This extension is a function on $\mathcal{X}$ that maps at most k elements to 1 and hence is an element of $\mathcal{H}_{\leq k}$. The reasoning is similar had g mapped at most k elements to 0. This show that VCdim$(\mathcal{H}_{\leq k}) \geq 2k+1$.

Now suppose that $C \subset \mathcal{X}$ is of size $2k+2$. Consider a map that assigns half the elements of C to 0 and the other half to 1. This map cannot be extended to a function in $\mathcal{H}_{\leq k}$. This proves that VCdim$(\mathcal{H}_{\leq k}) \leq 2k+1$. Thus:

$$\text{VCdim}(\mathcal{H}_{\leq k}) = \begin{cases} 2k+1 & \text{if } k < \lfloor n/2 \rfloor \\ n & \text{if } k \geq \lfloor n/2 \rfloor \end{cases}$$

Exercise 6.3

Since $|\mathcal{H}_{n\text{-parity}}| = 2^n$, using the upper bound on the VC-dimension,

$$\text{VCdim}(\mathcal{H}_{n\text{-parity}}) \leq \log_2 |\mathcal{H}_{n\text{-parity}}| = n.$$

We claim that $\text{VCdim}(\mathcal{H}_{n\text{-parity}}) = n$. Let $C = \{c_1, \ldots, c_n\} \subset \mathcal{X}$ be the set of standard basis vectors such that c_i is the basis vector with a 1 in the ith position and 0's elsewhere. Let $(b_1, \ldots, b_n)$ be a function from C to $\{0, 1\}$. Construct an index set $I \subseteq \{1, \ldots, n\}$ as follows: start with $I \leftarrow \varnothing$; for $1 \leq i \leq n$, if $b_i = 1$ then $I \leftarrow I \cup \{i\}$.

We claim that $h_I(c_j) = b_j$ for all $1 \leq j \leq n$. For if $b_j = 0$, then $j \notin I$ and $\sum_{i \in I} c_{ji} = 0 \pmod 2$; if $b_j = 1$, then $j \in I$ and $\sum_{i \in I} c_{ji} = 1 \pmod 2$, proving the claim. This shows that every function from C to $\{0, 1\}$ can be extended to a function in $\mathcal{H}_{n\text{-parity}}$. Hence $\text{VCdim}(\mathcal{H}_{n\text{-parity}}) \geq n$ and together with the upper bound for the VC-dimension, this implies that $\text{VCdim}(\mathcal{H}_{n\text{-parity}}) = n$.

Exercise 6.5

Let $\mathcal{H}^d_{\text{rect}}$ be the set of axis-aligned rectangles in $\mathbf{R}^d$. A function in $\mathcal{H}^d_{\text{rect}}$ is defined via $2d$ parameters $(a^1_1, a^1_2, a^2_1, a^2_2, \ldots, a^d_1, a^d_2)$, where for $1 \leq i \leq d$, $a^i_1 \leq a^i_2$ are the boundaries of the rectangle in dimension i.

We claim that $\mathcal{H}^d_{\text{rect}} = 2d$. Consider a set of $2d$ points that correspond to the centres of the faces of an axis-aligned rectangle in $\mathbf{R}^d$. For example, if the faces of the rectangle are defined by the equations: $x_i = a^i_1$ and $x_i = a^i_2$ for $1 \leq i \leq d$, then the centres of the face defined by $x_1 = a^1_1$ and $x_1 = a^i_2$ are $(a^1_1, \frac{a^2_1+a^2_2}{2}, \ldots, \frac{a^d_1+a^d_2}{2})$ and $(a^1_2, \frac{a^2_1+a^2_2}{2}, \ldots, \frac{a^d_1+a^d_2}{2})$. Similarly, there are two points for each of the remaining dimensions, with a total of $2d$ points. Call this set of points C. Such a set C will be shattered by $\mathcal{H}^d_{\text{rect}}$.

On the other hand, no subset C' of $\mathbf{R}^d$ of size $2d + 1$ can be shattered by $\mathcal{H}^d_{\text{rect}}$. The reasoning is similar to that given in the book. For each dimension i select a point $c^i_{\text{min}} \in C'$ whose ith co-ordinate is a minimum among all points in C'; also select $c^i_{\text{max}} \in C'$ whose ith co-ordinate is a maximum. This procedure yields $2d$ points and the rectangle that contains all these $2d$ points must necessarily contain the $2d + 1$st point. Hence a function that maps these $2d$ points to 1, and the $2d + 1$st point to 0 cannot be extended in $\mathcal{H}^d_{\text{rect}}$, proving that the set cannot be shattered.

Exercise 6.6

1. $|\mathcal{H}^d_{\text{con}}| \leq 3^d + 1$. One way of counting the number of boolean conjunctions is to first select a set of indices from among $\{1, \ldots, d\}$ and then from among these select either a positive or the negative version of the variables. When we select no indices, we obtain the all-positive hypothesis. The number of such conjunctions is:

$$\sum_{i=0}^{d} 2^i = 3^d.$$

This does not include the all-negative conjunction. Hence the *total* number of conjunctions is $3^d + 1$, which is a tight upper bound for $|\mathcal{H}^d_{\text{con}}|$.

2. VCdim($\mathcal{H}^d_{\text{con}}$) $\leq 3 \cdot \log_2 d$. This immediately follows from the upper bound in the text.

3. Let $\mathbf{e}_1, \ldots, \mathbf{e}_d$ be the standard basis vectors of $\{0,1\}^d$. Let $(b_1, \ldots, b_d)$ be a mapping from this set of basis vectors to $\{0,1\}$. Note that there are only 2^d such functions and we will show that each such function can be represented by a boolean conjunction on d variables.

   Start with the conjunction: $f := x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$; for $1 \leq i \leq d$, if $b_i = 1$ then drop $\bar{x}_i$ and all x_j for $j \neq i$ from f. Note that after this step we have that: $f(\mathbf{e}_j) = b_j$ for all $j \leq i$. Thus at step d, we end up with a formula f that matches the function $(b_1, \ldots, b_d)$. Thus the set of basis vectors is shattered by $\mathcal{H}^d_{\text{con}}$ and hence VCdim($\mathcal{H}^d_{\text{con}}$) $\geq d$.

4. VCdim($\mathcal{H}^d_{\text{con}}$) $\leq d$. Suppose that $C = \{c_1, \ldots, c_{d+1}\}$ is shattered by $\mathcal{H}^d_{\text{con}}$. This implies that every function from C to $\{0,1\}$ can be extended to a function in $\mathcal{H}^d_{\text{con}}$. Consider the $d+1$ functions $g_1, \ldots, g_{d+1}$, where g_i maps c_i to 0 and all c_j to 1 for $j \neq i$. Let $h_1, \ldots, h_{d+1}$ be the extensions of these functions in $\mathcal{H}^d_{\text{con}}$. Then for each $i \in \{1, \ldots, d+1\}$, there exists a literal l_i in h_i such that l_i is false for c_i but true for all c_j, $j \neq i$. Furthermore, each h_i has at most d literals since none of these functions is the all-zero function. By the Pigeon Hole Principle, there exists i, such that $1 \leq i \leq d$ and l_i and l_{d+1} use the same variable, say x_k.

   Since h_i maps c_i to 0 and c_{d+1} to 1 and h_{d+1} maps these the other way around, it cannot be that both l_i and l_{d+1} use x_k in the same form. That is, either $l_i = x_k$ and $l_{d+1} = \bar{x}_k$ or vice versa. Consider the effect of the literals on the bit strings in $C \setminus \{c_i, c_{d+1}\}$. Both map each bit string to 1, an impossibility since l_i and l_{d+1} will have the opposite effect on each of these bit strings too. This contradition shows that the assumption that there exist functions $h_1, \ldots, h_{d+1}$ that extend $g_1, \ldots, g_{d+1}$ is incorrect. Thus C is not shattered by $\mathcal{H}^d_{\text{con}}$ and VCdim($\mathcal{H}^d_{\text{con}}$) $\leq d$.

5. VCdim($\mathcal{H}^d_{\text{mcon}}$) $= d$. Since $\mathcal{H}^d_{\text{mcon}} \subseteq \mathcal{H}^d_{\text{con}}$, we know that VCdim($\mathcal{H}^d_{\text{mcon}}$) $\leq d$. Now consider $C = \{c_1, \ldots, c_d\} \subset \{0,1\}^d$, where c_i has ones in all locations except the ith. Let $(b_1, \ldots, b_d)$ be any function from C to $\{0,1\}$. Start with the conjunction: $f := x_1 \wedge \cdots \wedge x_d$; for $1 \leq i \leq d$, if $b_i = 1$ then drop x_i from f. Note that after this step we have that: $f(c_j) = b_j$ for all $j \leq i$. Thus at step d, we end up with a formula f that matches the function $(b_1, \ldots, b_d)$. Thus the set C is shattered by $\mathcal{H}^d_{\text{mcon}}$ and hence VCdim($\mathcal{H}^d_{\text{mcon}}$) $= d$.

Exercise 6.7

6.7.1

Let $\mathcal{H}$ be the set of all threshold functions h_a for $a \in [0,1]$. Then VCdim($\mathcal{H}$) $= 1$ and $|\mathcal{H}| = \infty$.

6.7.2

Define $\mathcal{H}$ to consist of the single threshold function $h_{1/2}$. In this case, $\log_2 |\mathcal{H}| = 0$ and if $a \in [0,1]$, the function $g(a) = 1 - h_{1/2}(a)$ cannot be extended in $\mathcal{H}$. Hence the VCdim($\mathcal{H}$) $= 0$, matching the upper bound.

Exercise 6.8

Fix a $d \in \mathbf{R}$. We will construct a set $C = \{x_1, \ldots, x_d\} \subset [0, 1]$ and show that for any boolean function $(b_1, \ldots, b_d)$ on C, there exists $\theta \in \mathbf{R}$ such that: $\lceil \sin(\theta x_i) \rceil = b_i$ for all $1 \le i \le d$. In particular, we will construct the binary representations of the elements of C. Consider a matrix with d rows and $2^d + 1$ columns, where the ith row represents the number x_i. Fill in the first 2^d columns of the matrix (from top to bottom) with the binary representations of the numbers $0, 1, \ldots, 2^d - 1$. Finally fill in the last column with ones. The number $x_i = 0.a_1a_2, \ldots, a_{2^d}a_{2^d+1}$, where $a_1, a_2, \ldots, a_{2^d}, a_{2^d+1}$ are the elements of the ith row of the matrix. This completes the construction of the d numbers from $[0, 1]$.

From the way this matrix has been constructed, it is clear that the bit patterns in the first 2^d columns are all the elements of $\{0, 1\}^d$. Let $(b_1, \ldots, b_d)$ be any boolean function defined on the set C. Then the bitwise complement $(\bar{b}_1, \ldots, \bar{b}_d)$ of the pattern $(b_1, \ldots, b_d)$ is in some column j of the matrix created in the last paragraph, where $1 \le j \le 2^d$. Define $\theta = 2^j\pi$, and using the hint provided in the text, we obtain that for all $1 \le i \le d$:

$$\lceil \sin(2^j \pi x_i) \rceil = 1 - B_j(x_i),$$

where $B_j(x_i)$ is the jth bit in the binary representation of x_i. Now this bit is simply $\bar{b}_i$, since the jth column of the matrix is $(\bar{b}_1, \ldots, \bar{b}_d)$. Hence the right-hand side of the equation is $1 - \bar{b}_i = b_i$. This shows that every boolean function defined on C can be extended in $\mathcal{H}$. Since one can do this for sets of any size d, VCdim$(\mathcal{H}) = \infty$.

Exercise 6.9

The VC-dimension of the class of signed intervals. Let c_1, c_2, c_3 be any three reals with $c_1 < c_2 < c_3$. We wish to show that for any function from $\{c_1, c_2, c_3\}$ to $\{-1, 1\}$ can be extended to a function in $\mathcal{H}$. The only contentious candidates are $(1, -1, 1)$ and $(-1, 1, -1)$. The first function $(1, -1, 1)$ can be extended to $h_{b_1, b_2, -1}$, where b_1 and b_2 are two reals such that $c_1 < b_1 < c_2$ and $c_2 < b_2 < c_3$. Similarly, the second function $(-1, 1, -1)$ can be extended to $h_{b_1, b_2, +1}$. This shows that VCdim$(\mathcal{H}) \ge 3$.

Now consider any set of four reals c_1, c_2, c_3, c_4 with $c_1 < c_2 < c_3 < c_4$. Then the function $(1, -1, 1, -1)$ cannot be extended to any function in $\mathcal{H}$. For such a function to exist, it would have to map an interval around c_2 to -1 and the rest of the reals to 1; but that also map c_4 to 1. This shows that such a function cannot exist and hence VCdim$(\mathcal{H}) \le 3$. Together with the upper bound, we have that VCdim$(\mathcal{H}) = 3$.

Exercise 6.11

6.11.1

Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over a fixed domain $\mathcal{X}$ and let $d := \max_i \text{VCdim}(\mathcal{H}_i)$. Suppose that $d \geq 3$. We need to show that

$$\text{VCdim}\left(\bigcup_{i=1}^{r} \mathcal{H}_i\right) \leq 4d\log(2d) + 2\log r$$

Let $C = \{c_1, \ldots, c_k\}$ be a set that is shattered by the union set $\bigcup_{i=1}^{r} \mathcal{H}_i$. Then all 2^k binary functions on C have extensions in $\bigcup_{i=1}^{r} \mathcal{H}_i$. By Sauer's Lemma, for any hypothesis class $\mathcal{H}_i$, $1 \leq i \leq r$, the number of possible extensions on a set of size k is $\sum_{i=0}^{d} \binom{k}{i} \leq (ek/d)^d$. This is strictly less than k^d since $d \geq 3$. Hence the total number of extensions possible in the union on a set of size k is strictly less than $r \cdot k^d$. Hence we must have $2^k < r \cdot k^d$, which implies that $k < d \cdot \log_2 k + \log_2 r$.

Lemma A.2 states that if $a \geq 1$ and $b > 0$ then $x \geq 4a\log(2a) + 2b$ implies that $x \geq a\log(x) + b$. The contrapositive states that if $x < a\log(x) + b$ then $x < 4a\log(2a) + 2b$. If we apply this to our case, we obtain that $k < 4d\log(2d) + 2\log(r)$.

Exercise 6.12

6.12.1

Let g and $\mathcal{F}$ be as stated. We first show that VCdim(POS($\mathcal{F} + g$)) $\geq$ VCdim(POS($\mathcal{F}$)). Let $C \subseteq \mathbf{R}^n$ be a set that is shattered by POS($\mathcal{F}$). Then for every $C' \subseteq C$, there exist $h_1, h_2 \in \mathcal{F}$ such that for all $x \in C$:

$$h_1(x) > 0 \text{ iff } x \in C'$$
$$h_2(x) > 0 \text{ iff } x \in C \setminus C'.$$

The idea here is to use a linear combination of h_1 and h_2 and g to obtain a function that takes on strictly positive values on C' and non-positive values on $C \setminus C'$. Define α_1 and α_2 as follows:

$$\alpha_1 := 1 + \frac{\max_{x \in C'} |g(x)|}{\min_{x \in C'} h_1(x)}, \qquad \alpha_2 := 1 + \frac{\max_{x \in C \setminus C'} |g(x)|}{\min_{x \in C \setminus C'} h_2(x)}$$

Then $\alpha_1 h_1(x) - \alpha_2 h_2(x) + g(x) > 0$ iff $x \in C'$, showing that POS($\mathcal{F} + g$) also shatters C. Hence VCdim(POS($\mathcal{F} + g$)) $\geq$ VCdim(POS($\mathcal{F}$)).

We next show that VCdim(POS($\mathcal{F} + g$)) $\leq$ VCdim(POS($\mathcal{F}$)). Now let $C \subseteq \mathbf{R}^n$ be shattered by POS($\mathcal{F} + g$). This means that for every $C' \subseteq C$, there exist $h_1, h_2 \in \mathcal{F}$ such that for all $x \in C$:

$$h_1(x) + g(x) > 0 \text{ iff } x \in C'$$
$$h_2(x) + g(x) > 0 \text{ iff } x \in C \setminus C'.$$

This immediately shows that for all $x \in C$:

$$(h_1(x) + g(x)) + (-h_2(x) - g(x)) > 0 \text{ iff } x \in C'.$$

Hence POS($\mathcal{F}$) shatters C and VCdim(POS($\mathcal{F} + g$)) $\leq$ VCdim(POS($\mathcal{F}$)).

6.12.2

We wish to show that VCdim(POS($\mathcal{F}$)) $=$ dim($\mathcal{F}$), where dim($\mathcal{F}$) is the dimension of $\mathcal{F}$ as a vector space. Let $\mathcal{F}$ be a finite dimensional vector space of dimension d with basis $f_1, \dots, f_d$. For any $h \in \mathcal{F}$ there exist real numbers $\alpha_1, \dots, \alpha_d$ such that

$$h = \alpha_1 f_1 + \cdots + \alpha_d f_d.$$

Note that

$$\text{POS}(\mathcal{F}) = \text{POS}(\{\alpha_1 f_1 + \cdots + \alpha_d f_d \mid (\alpha_1, \dots, \alpha_d) \in \mathbf{R}^d\}).$$

Hence each function in POS($\mathcal{F}$) can be associated with a d-tuple $(\alpha_1, \dots, \alpha_d)$ which, in turn, represents a homogeneous linear halfspace in $\mathbf{R}^d$. Since the VC-dimension of the set of homogeneous linear halfspaces in $\mathbf{R}^d$ is d, we have that VCdim(POS($\mathcal{F}$)) $= d$.

6.12.3

6.12.3.1

The set HS$_n$ $=$ POS($\mathcal{F}$), where $\mathcal{F}$ consists of functions $h_{\mathbf{w}}$, with $\mathbf{w} \in \mathbf{R}^n$, defined as follows: for $x \in \mathbf{R}^n$, $h_{\mathbf{w}}(x) = \langle \mathbf{w}, x \rangle$. Note that $\mathcal{F}$ is a n-dimensional vector space with basis $f_1, \dots, f_n$, where $f_i(x) = x_i$, that is, f_i maps a point in $\mathbf{R}^n$ to its ith coordinate. This shows that HS$_n$ is a Dudley class.

6.12.3.2

Similarly, the set HHS$_n$ $=$ POS($\mathcal{F}$), where $\mathcal{F}$ is a set of functions $h_{\mathbf{w},b}$, with $\mathbf{w} \in \mathbf{R}^n$, $b \in \mathbf{R}$ and defined as follows: for $x \in \mathbf{R}^n$, $h_{\mathbf{w},b}(x) = \langle \mathbf{w}, x \rangle + b$. It is sufficient to note that $\mathcal{F}$ is a vector space of dimension $n+1$ with the basis vectors $f_1, \dots, f_n, f_{n+1}$, where for $1 \le i \le n$, f_i maps points in $\mathbf{R}^n$ to their ith coordinate and f_{n+1} maps points to the constant 1.

6.12.3.3

The class B_d of open balls in $\mathbf{R}^d$ is the set of functions $h_{\mathbf{a},r}$, where $\mathbf{a} \in \mathbf{R}^d$ is the center of the ball and $r \in \mathbf{R}$ is its radius, such that for $\mathbf{x} \in \mathbf{R}^d$,

$$\begin{aligned} h_{\mathbf{a},r}(\mathbf{x}) &= r - \sum_{i=1}^{d} (x_i - a_i)^2 \\ &= 2a_1 x_1 + \cdots + 2a_d x_d + \left(r - \sum_{i=1}^{d} a_i^2 \right) \cdot 1 - \sum_{i=1}^{d} x_i^2. \end{aligned}$$

Define $g(\mathbf{x}) = -\sum_{i=1}^{d} x_i^2$ and for $1 \le i \le d$, define $f_i(\mathbf{x}) = x_i$; finally, define $f_{d+1}(\mathbf{x}) = 1$. Then $B_d \subseteq \mathcal{F} + g$, where $\mathcal{F}$ is a space of functions defined by:

$$\mathcal{F} := \{a_1 f_1 + \cdots + a_{d+1} f_{d+1} \mid (a_1, \dots, a_{d+1}) \in \mathbf{R}^d\}.$$

Since VCdim(POS($\mathcal{F} + g$)) $= d + 1$, we have that VCdim(POS(B_d)) $\leq d + 1$. Now a set U of $d + 1$ points in **R**^{d} can be shattered by d-dimensional half spaces. This means that for any subset $S \subseteq U$, there exists a hyperplane that separates S and $U \setminus S$. Thus there exist d-dimensional balls B_1, B_2 (of sufficiently large radius) on either side of this hyperplane and tangent to it such that $S \subseteq B_1$ and $U \setminus S \subseteq B_2$. Thus there is a set of size $d + 1$ shattered by POS(B_d) and hence VCdim(POS(B_d)) $\geq d+1$. Together with the upper bound, this gives: VCdim(POS(B_d)) $= d+1$.

6.12.3.4

1. A polynomial $p = a_0 + a_1 x + \cdots + a_d x^d$ of degree d may be thought of as the vector space of $(d + 1)$-dimensional tuples over **R**. Thus VCdim(P_1^d) $= d + 1$.

2. Since the class of polynomial classifiers of degree d has VC-dimension d, the VC-dimension of the class of all polynomial classifiers is unbounded.

3. In order to find out the VC-dimension of the class of polynomial classifiers of degree d on n variables, it is sufficient to establish the number of terms that such a polynomial can have. The number of terms equals the dimension of the function space of all such polynomials, which by 6.12.2 equals the VC-dimension. The degree of any such term is some number k, $1 \leq k \leq d$. For a fixed k, the number of terms of degree k is the number of non-negative solutions to the equation:
$$y_1 + \cdots + y_n = k,$$
where $y_i \geq 0$ for all $1 \leq i \leq n$. This, in turn, can be cast into the problem of finding out the number of bit patterns on $n - 1$ ones and k zeros. This is just $\binom{n-1+k}{k}$ and hence, the number of terms in such a polynomial is:
$$\sum_{k=0}^{d} \binom{n - 1 + k}{k},$$
which is the VC-dimension of P_n^d.

Chapter 7

Non-Uniform Learnability

In the (agnostic) PAC learning setting, the sample size depended only on the accuracy parameter ε and the confidence parameter δ. It was "uniform" w.r.t the hypothesis class and the underlying distribution. In non-uniform learnability, the sample size is allowed to depend on the hypothesis class. In particular, when a learning algorithm is competing against a specific hypothesis in the hypothesis class, then it is allowed to have a training sample size that depends on that hypothesis (and also on ε, δ).

Definition 7.1. Let $\mathcal{H}$ be a set of binary functions over a domain $\mathcal{X}$. The class $\mathcal{H}$ is *non-uniformly learnable* if there exist a learning algorithm $\mathcal{A}$ and a function $m_{\mathcal{H}}^{\mathrm{NUL}}\colon (0,1) \times (0,1) \times \mathcal{H} \to \mathbf{N}$ such that for all $\varepsilon, \delta \in (0,1)$, for all $h \in \mathcal{H}$ and all distributions $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the following holds for samples S with $m \geq m_{\mathcal{H}}^{\mathrm{NUL}}(\varepsilon, \delta, h)$ examples:

$$\mathbf{Pr}_{S \sim \mathcal{D}^m} \{L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(h) + \varepsilon\} \geq 1 - \delta.$$

It should be clear that non-uniform learnability is a relaxation of agnostic PAC-learnability in that if a hypothesis class is agnostic PAC-learnable then it is also non-uniformly learnable. In fact, the notion of non-uniform learnability is a strict generalization of the notion of agnostic PAC-learnability. This can be shown by considering the set of all thresholded polynomials. The VC-dimension of this class is infinite and hence it is not agnostic PAC-learnable, but it is non-uniformly learnable as the set of thresholded polynomials of any fixed degree d has VC-dimension $d+1$. The non-uniform learnability of all thresholded polynomials follows from the following characterization:

Theorem 7.1. *A hypothesis class $\mathcal{H}$ of binary classifiers is non-uniformly learnable iff it is the countable union of agnostically PAC-learnable hypothesis classes.*

On the other hand, the notion of non-uniform learnability is restricted enough that it does not allow all hypothesis classes to be learnable. This follows from Lemma 7.1 and the characterization of non-uniform learnable classes in Theorem 7.1.

Lemma 7.1. *For every infinite set $\mathcal{X}$, the set of all binary valued functions on $\mathcal{X}$ is not expressible as $\bigcup_{n=1}^{\infty} H_n$, where each H_n has finite VC-dimension.*

The main topic is in proving Theorem 7.1. We do this in two steps.

Lemma 7.2. *If $\mathcal{H}$ is non-uniformly learnable then there exist classes H_n, $n \in \mathbf{N}$, each of finite VC-dimension, such that $\mathcal{H} = \bigcup_{n=1}^{\infty} H_n$.*

Proof. Let $\mathcal{H}$ be as stated and for each $n \in \mathbf{N}$, define

$$H_n := \{h \in \mathcal{H} \colon m_{\mathcal{H}}^{\text{NUL}}(1/8, 1/7, h) \leq n\}.$$

Then clearly $H = \bigcup_{n=1}^{\infty} H_n$. By definition, each class H_n can be learned to an accuracy of $1/8$ with a confidence of $1/7$ using at most n examples. We claim that $\text{VCdim}(H_n) \leq 2n$ for all $n \in \mathbf{N}$. Suppose not. Then there exist $n \in \mathbf{N}$ and a set $A \subset \mathcal{X}$ of size $2n+1$ that is shattered by H_n. This means that every binary function on A admits an extension in H_n to a function on $\mathcal{X}$. By the No-Free-Lunch Theorem, in order to be able to learn H_n to an accuracy of $1/8$ with a confidence of $1/7$, we need at least $n+1$ examples. This contradiction shows that there cannot be any set of size $2n+1$ that is shattered by H_n. Hence $\text{VCdim}(H_n) \leq 2n$. □

To prove the other direction, we need to introduce the notion of the *Structural Risk Minimization (SRM)* paradigm. Let $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$ be a hypothesis class that is the union of a countably many uniformly convergent classes $\mathcal{H}_n$ with sample complexity function $m_n(\varepsilon, \delta)$. Define $\varepsilon_n \colon \mathbf{N} \times (0,1) \to (0,1)$ by

$$\varepsilon_n(m, \delta) := \min\{\varepsilon \in (0,1) \colon m_n(\varepsilon, \delta) \leq m\}.$$

Thus $\varepsilon_n(m, \delta)$ is just the best possible accuracy obtainable when learning functions from the class $\mathcal{H}_n$ with a sample size of at most m and with confidence δ.

Since a given $h \in \mathcal{H}$ may belong to an infinite number of $\mathcal{H}_n$, define $n(h) := \min\{n \in \mathbf{N} \colon h \in \mathcal{H}_n\}$. The other component of SRM is a weight function $w \colon \mathbf{N} \to [0,1]$ such that $\sum_{n=1}^{\infty} w(n) \leq 1$. The weight function $w(n)$ represents the importance that the learner attributes to each hypothesis class $\mathcal{H}_n$.

Theorem 7.2. *Let $\mathcal{H}$ be a hypothesis class that can be written as $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, where each class $\mathcal{H}_n$ is uniformly convergent with sample complexity $m_n(\varepsilon, \delta)$. Let $w \colon \mathbf{N} \to [0,1]$ be a weight function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. Then for every $\delta \in (0,1)$, for every distribution $\mathcal{D}$, with probability of at least $1-\delta$ over the choice of samples $S \sim \mathcal{D}^m$, the following holds for all $n \in \mathbf{N}$ and all $h \in \mathcal{H}_n$:*

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon_n(m, w(n) \cdot \delta).$$

The main claim of this theorem can be restated as either one of these statements:

- for all $h \in \mathcal{H}$: $L_{\mathcal{D}}(h) \leq L_S(h) + \varepsilon_n(m, w(n) \cdot \delta)$
- for all $h \in \mathcal{H}$: $L_S(h) \leq L_{\mathcal{D}}(h) + \varepsilon_n(m, w(n) \cdot \delta)$

Given a training sample S and a confidence parameter δ, the SRM paradigm is to select a hypothesis $h \in \mathcal{H}$ that minimizes

$$L_S(h) + \varepsilon_{n(h)}(m, w(n(h)) \cdot \delta).$$

Theorem 7.3. *Let $\mathcal{H}$ be a hypothesis class that can be written as $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, where each class $\mathcal{H}_n$ is uniformly convergent with sample complexity $m_n(\varepsilon, \delta)$. Let $w\colon \mathbf{N} \to [0,1]$ be a weight function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. Then $\mathcal{H}$ is non-uniformly learnable using the SRM rule with rate:*

$$m_{\mathcal{H}}^{NUL}(\varepsilon, \delta, h) \leq m_{n(h)}(\varepsilon/2, w(n(h)) \cdot \delta).$$

Proof. Let $\mathcal{A}$ be the SRM algorithm. Given ε, δ and $h \in \mathcal{H}$, choose $m \geq m_{n(h)}(\varepsilon/2, w(n(h)) \cdot \delta)$. By Theorem 7.2, for every distribution $\mathcal{D}$, with probability at least $1-\delta$ over the choice of samples $S \sim \mathcal{D}^m$ for all $h' \in \mathcal{H}$:

$$L_{\mathcal{D}}(h') \leq L_S(h') + \varepsilon_{n(h')}(m, w(n(h')) \cdot \delta).$$

Since $\mathcal{A}$ uses the SRM rule:

$$\begin{aligned} L_{\mathcal{D}}(A(S)) &\leq \min_{h' \in \mathcal{H}} \{L_S(h') + \varepsilon_{n(h')}(m, w(n(h')) \cdot \delta)\} \\ &\leq L_S(h) + \varepsilon_{n(h)}(m, w(n(h)) \cdot \delta). \end{aligned}$$

By Theorem 7.2, with probability at least $1-\delta$ over the choice of samples $S \sim \mathcal{D}^m$:

$$\begin{aligned} L_S(h) + \varepsilon_{n(h)}(m, w(n(h)) \cdot \delta) &\leq L_{\mathcal{D}}(h) + 2 \cdot \varepsilon_{n(h)}(m, w(n(h)) \cdot \delta) \\ &\leq L_{\mathcal{D}}(h) + 2 \cdot \varepsilon/2 \\ &\leq L_{\mathcal{D}}(h) + \varepsilon. \end{aligned}$$

□

An immediate implication is that every countably infinite hypothesis class for binary classification is non-uniformly learnable using a suitably defined weight function. For if, $\mathcal{H} = \bigcup_{n=1}^{\infty} \{h_n\}$ then, by Corollary 4.6, each singleton class $\{h_n\}$ has the uniform convergence property with sample complexity $m(\varepsilon, \delta) = \frac{\log(2/\delta)}{\varepsilon^2}$.

Ex 7.1

Let l be the maximum description length of a hypothesis in $\mathcal{H}$. Since the descriptions are over the alphabet $\{0,1\}$, the maximum number of hypotheses that can possibly be represented is:

$$2^l + 2^{l-1} + \cdots + 2^1 + 2^0 = 2^{l+1} - 1.$$

Hence $|\mathcal{H}| \leq 2^{l+1} - 1 < 2^{l+1}$. Since VCdim$(\mathcal{H}) \leq \log_2 |\mathcal{H}|$, we have that: VCdim$(\mathcal{H}) < l + 1$.

If, in addition, the descriptions are prefix-free, then by Kraft's inequality:

$$\frac{|\mathcal{H}|}{2^l} \leq \sum_{h \in \mathcal{H}} \frac{1}{2^{|h|}} \leq 1,$$

which yields $|\mathcal{H}| \leq 2^l$, and hence VCdim$(\mathcal{H}) \leq l$.

Ex 7.2

It is sufficient to prove that there exists no weighting function $w\colon \mathbf{N} \to [0,1]$ that is not indentically zero such that

1. $\sum_{n=1}^{\infty} w(n) \leq 1$
2. w is non-decreasing: for all $i < j$, $w(i) \leq w(j)$.

Since, by hypothesis, w is not identically zero, there exists an index n_0 such that $w_{n_0} > 0$. Since w is non-decreasing, this implies that for all $n \geq n_0$, $w(n) \geq w(n_0)$. Hence $\sum_{i=0}^{k} w(n_0+i) \geq k \cdot w(n_0)$, which implies that the sum $\sum_{i=1}^{\infty} w(i)$ diverges to infinity, contradicting requirement (1).

Ex 7.3

Ex 7.3.1

Let $\mathcal{H} = \bigcup_{n \in \mathbf{N}} \mathcal{H}_n$, where $\mathcal{H}_n$ is finite for all $n \in \mathbf{N}$. For $h \in \mathcal{H}$, define

$$w(h) = \frac{1}{|H_{n(h)}| \cdot 2^{n(h)}}.$$

Then

$$\begin{aligned}\sum_{h \in \mathcal{H}} w(h) &= \sum_{h \in \mathcal{H}} \frac{1}{|H_{n(h)}| \cdot 2^{n(h)}} \\ &\leq \sum_{j=1}^{\infty} \frac{|H_j|}{|H_j| \cdot 2^j} \\ &\leq 1.\end{aligned}$$

Ex 7.3.2

Let $\mathcal{H} = \bigcup_{n \in \mathbf{N}} \mathcal{H}_n$, where $\mathcal{H}_n$ is countably infinite. In this case, every $h \in \mathcal{H}$ can be uniquely identified by a pair of natural numbers (i,j), where $i = n(h)$ and $j = \mathrm{id}(h)$ is the index of h in $\mathcal{H}_{n(h)}$. Now define $w(h)$ as $\frac{1}{2^{n(h)+\mathrm{id}(h)}}$. With this definition, a countably infinite number of $h \in \mathcal{H}$ may have the same first component $n(h)$, but since each hypothesis has a distinct index j in $\mathcal{H}_{n(h)}$, we ensure that the weights associated with the class $\mathcal{H}_{n(h)}$ are decreasing. Indeed,

$$\begin{aligned}\sum_{h\in\mathcal{H}} w(h) &= \sum_{h\in\mathcal{H}} \frac{1}{2^{n(h)+\text{id}(h)}}\\ &\le \sum_{n=1}^{\infty}\sum_{j=1}^{\infty}\frac{1}{2^{n+j}}\\ &\le \sum_{n=1}^{\infty}\frac{1}{2^n}\sum_{j=1}^{\infty}\frac{1}{2^j}\\ &\le \sum_{n=1}^{\infty}\frac{1}{2^n}\cdot 1\\ &\le 1.\end{aligned}$$

Ex 7.5

A No-Free-Lunch result for non-uniform learnability.

Theorem 7.4. *Let $\mathcal{H}$ be a class that shatters an infinite set. Then for every sequence of classes $\{\mathcal{H}_n\}_{n\in\mathbf{N}}$ such that $\mathcal{H} = \bigcup_{n\in\mathbf{N}}\mathcal{H}_n$, there exists some n for which VCdim$(\mathcal{H}_n) = \infty$.*

Proof. Let $\mathcal{H}$ be a hypothesis class that shatters a countably infinite set $K = \{k_i\}_{i=\mathbf{N}}$ and suppose that $\mathcal{H} = \bigcup_{n\in\mathbf{N}}\mathcal{H}_n$, where every $\mathcal{H}_n$ has finite VC-dimension. For $n \in \mathbf{N}$, define $K_n \subseteq K$ such that $|K_n| = \text{VCdim}(\mathcal{H}_n) + 1$ and $K_n \cap K_m = \varnothing$ for $n \neq m$. One possible definition that satisfies these conditions is the following:

$$K_n = \{k_{r+1}, \dots, k_{r+\text{VCdim}(\mathcal{H}_n)+1}\}$$

where $r = \sum_{j=1}^{n-1}\text{VCdim}(\mathcal{H}_j)$. Clearly, $K_1, \dots, K_n, \dots$ are pairwise disjoint and $K = \bigcup_{n\in\mathbf{N}} K_n$.

Since $|K_n| > \text{VCdim}(\mathcal{H}_n)$, there exists a function $f_n \colon K_n \to \{0,1\}$ that does not agree with any function in $\mathcal{H}_n$ on K_n. Define $f \colon K \to \{0,1\}$ as follows: for $k \in K_n$, $f(k) = f_n(k)$. Since $\mathcal{H}$ shatters K, we must have $f \in \mathcal{H}$. By the definition of f, for all n, f differs from every function in $\mathcal{H}_n$ on the set K_n. Thus $f \notin \bigcup_{n\in\mathbf{N}}\mathcal{H}_n$, a contradiction to the assumption that $\mathcal{H} = \bigcup_{n\in\mathbf{N}}\mathcal{H}_n$, where each $\mathcal{H}_n$ has finite VC-dimension. This shows that if $\mathcal{H}$ shatters a countably infinite set, then it does not admit such a represtation. Since any non-uniformly learnable class does in fact admit such a representation, it follows that $\mathcal{H}$ is not non-uniformly learnable. □

Chapter 8

The Computational Complexity of Learning

Exercise 8.1

In this case, an input sample consists of a set $S = \{(x, b) \colon x \in \mathbf{R}, b \in \{0, 1\}\}$. If we were working in the realizable case, then we could simply define $l = \min_x\{(x, 1) \in S\}$ and $L = \max_x\{(x, 1) \in S\}$ and output $[l, L]$ as the interval. This would take time $O(m)$, where $m = |S|$. In the agnostic case, we can consider all possible pairs (x, y) for $(x, 1), (y, 1) \in S$ with $x < y$. For each pair (x, y), we evaluate the empirical loss $L_S([x, y])$ and output that interval that minimizes the empirical loss. This takes time $O(m^2)$.

Chapter 9

Boosting

9.1 Notes

The Fundamental Theorem of Statistical Learning guarantees that if a hypothesis class has finite VC-dimension d, the ERM algorithm will be able to learn it with a sample of size $\Omega(\frac{d+\log(1/\delta)}{\varepsilon})$ in the realizable setting, and $\Omega(\frac{d+\log(1/\delta)}{\varepsilon^2})$ in the agnostic setting. From the statistical perspective, there is *no* difference between the realizable and agnostic setting. Learning is solely determined by the VC-dimension of the hypothesis class.

However the computational complexity of implementing the ERM algorithm varies widely between these two settings. Implementing the ERM algorithm for learning Boolean conjunctions or the class of axis-aligned rectangles can be efficiently done (as in polynomial time in the input size) in the realizable case; however, these problems are NP-hard in the agnostic case. Since the notion of PAC-learning (in the realizable setting) deals with being able to approximate the true hypothesis with arbitrary accuracy, it makes sense to ask when learning is computationally feasible if we drop this requirement and consider classifiers that are just slightly better than making a random guess. This leads us to the notion of γ-weak-learnability.

Definition 9.1 (γ-weak-learnability)**.** A learning algorithm A is a γ-weak-learner for a hypothesis class $\mathcal{H}$ if there exists a function $m_{\mathcal{H}}\colon (0,1) \to \mathbf{N}$ such that for every $\delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$ and every labeling function $f\colon \mathcal{X} \to \{\pm 1\}$, if A is presented with $m \geq m_{\mathcal{H}}(\delta)$ examples chosen i.i.d. according to $\mathcal{D}$, it outputs a hypothesis that with probability at least $1-\delta$ has a true error of at most $1/2-\gamma$.

Exercise 10.1

For $\varepsilon, \delta \in (0,1)$, define

$$m_{\mathcal{H}}(\varepsilon,\delta) := k \cdot m_{\mathcal{H}}(\varepsilon/2) + \left\lceil \frac{2\log(4k/\delta)}{\varepsilon^2} \right\rceil,$$

where $k = \lceil \log(\delta/2)/\log(\delta_0) \rceil$. Now given a sample of size at least $m_{\mathcal{H}}(\varepsilon,\delta)$, divide it into $k+1$ chunks $S_1,\ldots,S_{k+1}$ such that the first k chunks has size $m_{\mathcal{H}}(\varepsilon/2)$.

Using algorithm A, train the first k chunks to obtain hypotheses $\hat{h}_1, \ldots, \hat{h}_k$. Then with probability at least $1 - \delta_0^k \geq 1 - \delta/2$, we have that

$$\min_{1 \leq i \leq k} \{L_{\mathcal{D}}(\hat{h}_i)\} \leq \min_{h' \in \mathcal{H}} \{L_{\mathcal{D}}(h')\} + \varepsilon/2.$$

Call this event 1.

Now that we have a finite hypothesis class $\{\hat{h}_1, \ldots, \hat{h}_k\}$, and since the $(k+1)$st chunk has size at least $\lceil 2\log(4k/\delta)/\varepsilon^2 \rceil$, by Corollary 4.6, we can use ERM to obtain a hypothesis $\hat{h}$ such that with probability at least $1 - \delta/2$, we have that

$$L_{\mathcal{D}}(\hat{h}) \leq \min_{1 \leq i \leq k} \{L_{\mathcal{D}}(\hat{h}_i)\} + \varepsilon/2 \leq \min_{h' \in \mathcal{H}} \{L_{\mathcal{D}}(h')\} + \varepsilon.$$

Call this event 2.

Using the union bound, the probability that either event 1 does *not* happen or event 2 does *not* happen is at most δ. Hence the probability that *both* events do happen is at least $1 - \delta$. Hence with probability $1 - \delta$, one can find a hypothesis $\hat{h}$ such that $L_{\mathcal{D}}(\hat{h}) \leq \min_{h' \in \mathcal{H}} \{L_{\mathcal{D}}(h')\} + \varepsilon$.

Exercise 10.2

Let $\theta_0, \theta_1, \ldots, \theta_T \in \mathbf{R}$ such that $\theta_0 = -\infty$ and $\theta_T = +\infty$. Define $g \colon \mathbf{R} \to \{\pm 1\}$ such that $g(x) = (-1)^t$ when $x \in (\theta_{t-1}, \theta_t]$ for $1 \leq t \leq T-1$ and $g(x) = (-1)^T$ for $x > \theta_{T-1}$. Define

$$h(x) = \text{sign}\left(\sum_{t=0}^{T} w_t \,\text{sign}(x - \theta_t)\right)$$

where $w_0 = 0.5$ and $w_t = (-1)^{t+1}$ for $1 \leq t \leq T$. We will show that $h = g$ by inducting on T.

Let $T = 1$. Then for all $x \in \mathbf{R}$, $g(x) = -1$ and

$$h(x) = \text{sign}(0.5 \cdot \text{sign}(x - \theta_0) + \text{sign}(x - \theta_1)) = \text{sign}(0.5 - 1) = -1.$$

Let us assume that the result holds when $T = k \geq 1$. Consider the case when $T = k + 1$. We distinguish two cases here: when k is even and when k is odd.

$$\begin{aligned} h(x) = \text{sign}\,(0.5 \cdot \text{sign}(x - \theta_0) + \text{sign}(x - \theta_1) + \cdots \\ + (-1)^{k+1} \cdot \text{sign}(x - \theta_k) + (-1)^{k+2} \cdot \text{sign}(x - \theta_{k+1})\big) \end{aligned}$$

First assume that k is even. By induction hypothesis, we know that for all $x < \theta_k$, $h(x) = g(x)$. For $x = \theta_k$, $g(x) = +1$ and

$$\begin{aligned} h(x) = \text{sign}(0.5 + (1 - 1) + \ldots + (1 + (-1)^{k+1} \cdot \text{sign}(x - \theta_k)) + \\ (-1)^{k+2}\,\text{sign}(x - \theta_{k+1})). \end{aligned}$$

This equals $\text{sign}(0.5 + 2 - 1) = +1$. For $x > \theta_k$, $g(x) = -1$ and

$$\begin{aligned} h(x) &= \text{sign}(0.5 + (1 - 1) + \ldots + (1 - 1) + (-1)^{k+2}\,\text{sign}(x - \theta_{k+1})) \\ &= \text{sign}(0.5 - 1) \\ &= -1. \end{aligned}$$

One can show that $h = g$ for k odd.