Understanding Machine Learning

Shai Shalev-Shwartz and Shai Ben-David

Triple-S dedicates the book to triple-M

Preface

The term *machine learning* refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets. We are surrounded by a machine learning based technology: search engines learn how to bring us the best results (while placing profitable ads), anti-spam software learns to filter our email messages, and credit card transactions are secured by a software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident prevention systems that are built using machine learning algorithms. Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy.

One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, fine-detailed specification of how such tasks should be executed. Taking example from intelligent beings, many of our skills are acquired or refined through *learning* from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with endowing programs with the ability to "learn" and adapt.

The first goal of this book is to provide a rigorous, yet easy to follow, introduction to the main concepts underlying machine learning: What is learning? How can a machine learn? How do we quantify the resources needed to learn a given concept? Is learning always possible? Can we know if the learning process succeeded or failed?

The second goal of this book is to present several key machine learning algorithms. We chose to present algorithms that on one hand are successfully used in practice and on the other hand give a wide spectrum of different learning techniques. Additionally, we pay specific attention to algorithms appropriate for large scale learning (a.k.a. "Big Data"), since in recent years, our world has become increasingly "digitized" and the amount of data available for learning is dramatically increasing. As a result, in many applications data is plentiful and computation time is the main bottleneck. We therefore explicitly quantify both the amount of data and the amount of computation time needed to learn a given concept.

The book is divided into four parts. The first part aims at giving an initial rigorous answer to the fundamental questions of learning. We describe a generalization of Valiant's Probably Approximately Correct (PAC) learning model, which is a first solid answer to the question "what is learning?". We describe the Empirical Risk Minimization (ERM), Structural Risk Minimization (SRM), and Minimum Description Length (MDL) learning rules, which shows "how can a machine learn". We quantify the amount of data needed for learning using the ERM, SRM, and MDL rules and show how learning might fail by deriving

a "no-free-lunch" theorem. We also discuss how much computation time is required for learning. In the second part of the book we describe various learning algorithms. For some of the algorithms, we first present a more general learning principle, and then show how the algorithm follows the principle. While the first two parts of the book focus on the PAC model, the third part extends the scope by presenting a wider variety of learning models. Finally, the last part of the book is devoted to advanced theory.

We made an attempt to keep the book as self-contained as possible. However, the reader is assumed to be comfortable with basic notions of probability, linear algebra, analysis, and algorithms. The first three parts of the book are intended for first year graduate students in computer science, engineering, mathematics, or statistics. It can also be accessible to undergraduate students with the adequate background. The more advanced chapters can be used by researchers intending to gather a deeper theoretical understanding.

Acknowledgements

The book is based on Introduction to Machine Learning courses taught by Shai Shalev-Shwartz at the Hebrew University and by Shai Ben-David at the University of Waterloo. The first draft of the book grew out of the lecture notes for the course that was taught at the Hebrew University by Sbai Shalev-Shwartz during 2010–2013. We greatly appreciate the help of Ohad Shamir, who served as a TA for the course in 2010, and of Alon Gonen, who served as a TA for the course in 2011–2013. Ohad and Alon prepared few lecture notes and many of the exercises. Alon, to whom we are indebted for his help throughout the entire making of the book, has also prepared a solution manual.

We are deeply grateful for the most valuable work of Dana Rubinstein. Dana has scientifically proofread and edited the manuscript, transforming it from lecture-based chapters into fluent and coherent text.

Special thanks to Amit Daniely, who helped us with a careful read of the advanced part of the book and also wrote the advanced chapter on multiclass learnability. We are also grateful for the members of a book reading club in Jerusalem that have carefully read and constructively criticized every line of the manuscript. The members of the reading club are: Maya Alroy, Yossi Arjevani, Aharon Birnbaum, Alon Cohen, Alon Gonen, Roi Livni, Ofer Meshi, Dan Rosenbaum, Dana Rubinstein, Shahar Somin, Alon Vinnikov, and Yoav Wald. We would also like to thank Gal Elidan, Amir Globerson, Nika Haghtalab, Shie Mannor, Amnon Shashua, Nati Srebro, and Ruth Urner for helpful discussions.

Shai Shalev-Shwartz, Jerusalem, Israel
Shai Ben-David, Waterloo, Canada

Contents

1 Introduction

The subject of this book is automated learning, or, as we will more often call it, Machine Learning (ML). That is, we wish to program computers so that they can "learn" from input available to them. Roughly speaking, learning is the process of converting experience into expertise or knowledge. The input to a learning algorithm is training data, representing experience, and the output is some expertise, which usually takes the form of another computer program that can perform some task. Seeking a formal-mathematical understanding of this concept, we'll have to be more explicit about what we mean by each of the involved terms; What is the training data our programs will access? How can the process of learning be automated? How can we evaluate the success of such a process (namely the quality of the output of a learning program)?

1.1 What is learning?

Let us begin by considering a couple of examples from naturally occurring animal learning. Some of the most fundamental issues in ML arise already in that context, that we are all familiar with.

Bait Shyness—rats learning to avoid poisonous baits: When rats encounter food items with novel look or smell, they will first eat very small amounts, and subsequent feeding will depend on the flavor of the food and its physiological effect. If the food produces an ill effect, the novel food would often be associated with the illness, and subsequently, the rats will not eat it. Clearly, there is a learning mechanism in play here – the animal used past experience with some food to acquire expertise in detecting the safety of this food. If past experience with the food was negatively labeled, the animal predicts that it will also have a negative effect when encountered in the future.

Inspired by the above example of successful learning, let us demonstrate a typical machine learning task. Suppose we would like to program a machine that learns how to filter spam emails. A naive solution would be seemingly similar to the way rats learn how to avoid poisonous baits. The machine would simply *memorize* all previous emails, that had been labeled as spam emails by the human user. When a new email arrives, the machine would search for it in the set of

previous spam emails. If it matches one of them it will be trashed. Otherwise, it will be moved to the user's inbox folder.

While the above "learning by memorization" approach is sometimes useful, it lacks an important aspect of learning systems—the ability to label unseen email messages. A successful learner should be able to progress from individual examples to broader *generalization*. This is also referred to as *inductive reasoning* or *inductive inference*. In the bait shyness example presented above, after the rats encounter an example of a certain type of food, they apply their attitude towards it on new, unseen examples of food of similar smell and taste. To achieve generalization in the spam filtering task, the learner can scan the previously seen emails, and extract a set of words whose appearance in an email message is indicative of spam. Then, when a new email arrives, the machine can check if one of the suspicious words appear in it, and predict its label accordingly. Such a system would potentially be able to correctly predict the label of unseen emails.

However, inductive reasoning might lead us to false conclusions. To illustrate this, let us consider again an example from animal learning.

Pigeon superstition: In an experiment performed by the psychologist B.F. Skinner, he placed a bunch of hungry pigeons in a cage. An automatic mechanism has been attached to the cage, delivering food to the pigeons at regular intervals with no reference whatsoever to the birds' behavior. The hungry pigeons go around the cage, and when food is first delivered, it finds each pigeon engaged in some activity (pecking, turning the head, etc.). The arrival of food reinforces each bird's specific action, and consequently, each bird tends to spend some more time doing that very same action. That, in turn, increases the chance that the next random food delivery will find each bird engaged in that activity again. What results is a chain of events that reinforces the pigeons' association of the delivery of the food with whatever chance actions they had been performing when it was first delivered. They subsequently continue to perform these same actions diligently.[1]

What distinguishes learning mechanisms that result in superstition from useful learning? This question is crucial to the development of automated learners. While human learners can rely on common sense to filter out random meaningless learning conclusions, once we export the task of learning to a machine, we must provide well defined crisp principles that will protect the program from reaching senseless or useless conclusions. The development of such principles is a central goal of the theory of machine learning.

What, then, made the rats' learning more successful than that of the pigeons? As a first step towards answering this question, let us have a closer look at the bait shyness phenomenon in rats.

Bait Shyness revisited—rats fail to acquire conditioning between food and electric shock or between sound and nausea: The bait shyness mechanism in rats turns out to be more complex than what one may expect. In experiments car-

[1] See: http://psychclassics.yorku.ca/Skinner/Pigeon

ried out by Garcia ((Garcia & Koelling 1996)), it was demonstrated that if the unpleasant stimulus that follows food consumption is replaced by, say, electrical shock (rather than nausea), then no conditioning occurs. Even after repeated trials in which the consumption of some food is followed by the administration of unpleasant electrical shock, the rats do not tend to avoid that food. Similar failure of conditioning occurs when the characteristic of the food that implies nausea (such as taste or smell) is replaced by a vocal signal. The rats seem to have some "built in" prior knowledge telling them that, while temporal correlation between food and nausea can be causal, it is unlikely that there would be a causal relationship between food consumption and electrical shocks or between sounds and nausea.

We conclude that one distinguishing feature between the bait shyness learning and the pigeon superstition is the incorporation of *prior knowledge* that biases the learning mechanism. This is also referred to as *inductive bias*. The pigeons in the experiment are willing to adopt *any* explanation to the occurrence of food. However, the rats "know" that food cannot cause an electric shock and that the co-occurrence of noise with some food is not likely to effect the nutritional value of that food. The rats' learning process is biased towards detecting some kind of patterns while ignoring other temporal correlations between events.

It turns out that the incorporation of prior knowledge, biasing the learning process, is inevitable for the success of learning algorithms (this is formally stated and proved as the "No Free Lunch theorem" in Chapter 5). The development of tools for expressing domain expertise, translating it into a learning bias, and quantifying the effect of such a bias on the success of learning, is a central theme of the theory of machine learning. Roughly speaking, the stronger the prior knowledge (or prior assumptions) that one starts the learning process with, the easier it is to learn from further examples. However, the stronger these prior assumptions are, the less flexible the learning is - it is bound, a priory, by the commitment to these assumptions. We shall discuss these issues explicitly in Chapter 5.

1.2 When do we need machine learning?

When do we need machine learning rather than directly program our computers to carry out the task at hand? Two aspects of a given problem may call for the use of programs that learn and improve based on their "experience": the problem's complexity and the need for adaptivity.

Tasks that are too complex to program.

- *Tasks performed by animals/humans:* there are numerous tasks that we, human beings, perform routinely, yet our introspection concerning how we do them is not sufficiently elaborate to extract

a well defined program. Examples of such tasks include driving, speech recognition, and image understanding. In all of these tasks, state of the art machine learning programs, programs that "learn from their experience", achieve quite satisfactory results, once exposed to sufficiently many training examples.

- *Tasks beyond human capabilities:* another wide family of tasks that benefit from machine learning techniques are related to the analysis of very large and complex data sets: Astronomical data, turning medical archives into medical knowledge, weather prediction, analysis of genomic data, web search engines, and electronic commerce. With more and more available digitally recorded data, it becomes obvious that there are treasures of meaningful information buried in data archives that are way too large and too complex for humans to make sense of. Learning to detect meaningful patterns in large and complex data sets is a promising domain in which the combination of programs that learn with the almost unlimited memory capacity and ever increasing processing speed of computers open up new horizons.

Adaptivity. One limiting feature of programmed tools is their rigidity - once the program has been written down and installed, it stays unchanged. However, many tasks change over time or from one user to another. Machine learning tools - programs whose behavior adapts to their input data - offer a solution to such issues; they are, by nature, adaptive to changes in the environment they interact with. Typical successful applications of machine learning to such problems include programs that decode hand written text, where a fixed program can adapt to variations between the handwriting of different users, spam detection programs, adapting automatically to changes in the nature of spam emails, and speech recognition programs.

1.3 Types of learning

Learning is, of course, a very wide domain. Consequently, the field of machine learning has branched into several subfields dealing with different types of learning tasks. We give a rough taxonomy of learning paradigms, aiming to provide some perspective of where the content of this book sits within the wide field of machine learning.

We describe four parameters along which learning paradigms can be classified.

Supervised vs. Unsupervised Since learning involves an interaction between the learner and the environment, one can divide learning tasks according to the nature of that interaction. The first distinction to note is the difference between supervised and unsupervised learning. As an illustrative

example, consider the task of learning to detect spam email versus the task of anomaly detection. For the spam detection task, we consider a setting in which the learner receives training emails for which the label `spam/not-spam` is provided. Based on such training the learner should figure out a rule for labeling a newly arriving email message. In contrast, for the task of anomaly detection, all the learner gets as training is a large body of email messages (with no labels) and the learner's task is to detect "unusual" messages.

More abstractly, viewing learning as a process of "using experience to gain expertise", supervised learning describes a scenario in which the "experience", a training example, contains significant information (say, the `spam/not-spam` labels) that is missing in the unseen "test examples" to which the learned expertise is to be applied. In this setting, the acquired expertise is aimed to predict that missing information for the test data. In such cases, we can think of the environment as a teacher that "supervises" the learner by providing the extra information (labels). In unsupervised learning, however, there is no distinction between training and test data. The learner processes input data with the goal of coming up with some summary, or compressed version of that data. Clustering a data set into subsets of similar objets is a typical example of such a task.

There is also an intermediate learning setting in which, while the training examples contain more information than the test examples, the learner is required to predict even more information for the test examples. For example, one may try to learn a value function, that describes for each setting of a chess board the degree by which White's position is better than the Black's. Yet, the only information available to the learner at training time is positions that occurred throughout actual chess games, labeled by who eventually won that game. Such learning frameworks are mainly investigated under the title of *reinforcement learning*.

Active vs. Passive learners Learning paradigms can vary by the role played by the learner. We distinguish between 'active' and 'passive' learners. An active learner interacts with the environment at training time, say by posing queries or performing experiments, while a passive learner only observes the information provided by the environment (or the teacher) without influencing or directing it. Note that the learner of a spam filter is usually passive - waiting for users to mark the emails arriving to them. In an active setting, one could imagine asking users to label specific emails chosen by the learner, or even composed by the learner to enhance its understanding of what spam is.

Helpfulness of the teacher When one thinks about human learning, of a baby at home, or a student at school, the process often involves a helpful teacher, who is trying to feed the learner with the information most use-

ful for achieving the learning goal. In contrast, when a scientist learns about nature, the environment, playing the role of the teacher, can be best thought of as passive - apples drop, stars shine and the rain falls without regards to the needs of the learner. We model such learning scenarios by postulating that the training data (or the learner's experience) is generated by some random process. This is the basic building block in the branch of 'statistical learning'. Finally, learning also occurs when the learner's input is generated by an adversarial "teacher". This may be the case in the spam filtering example (if the spammer makes an effort to mislead the spam filtering designer) or in learning to detect fraud. One also uses an adversarial teacher model as a worst-case-scenario, when no milder setup can be safely assumed. If you can learn against an adversarial teacher, you are guaranteed to succeed interacting any odd teacher.

Online vs. Batch learning protocol The last parameter we mention is the distinction between situations in which the learner has to respond online, throughout the learning process, and settings in which the learner has to engage the acquired expertise only after having a chance to process large amounts of data. For example, a stock broker has to make daily decisions, based on the experience collected so far. He may become an expert over time, but might have made costly mistakes in the process. In contrast, in many data mining settings, the learner - the data miner - has large amounts of training data to play with before having to output conclusions.

In this book we shall discuss only a subset of the possible learning paradigms. Our main focus is on supervised statistical batch learning with a passive learner (like for example, trying to learn how to generate patients' prognosis, based on large archives of records of patients that were independently collected and are already labeled by the fate of the recorded patients). We shall also briefly discuss online learning and batch unsupervised learning (in particular, clustering).

1.4 Relations to other fields

As an interdisciplinary field, machine learning shares common threads with the mathematical fields of statistics, information theory, game theory, and optimization. It is naturally a sub-field of computer science, as our goal is to program machines so that they will learn. In a sense, machine learning can be viewed as a branch of AI (Artificial Intelligence), since after all, the ability to turn experience into expertise or to detect meaningful patterns in complex sensory data is a corner stone of human (and animal) intelligence. However, one should note that, in contrast with traditional AI, machine learning is not trying to build automated imitation of intelligent behavior, but rather to use the strengths and

special abilities of computers to complement human intelligence, often performing tasks that fall way beyond human capabilities. For example, the ability to scan and process huge databases allows machine learning programs to detect patterns that are outside the scope of human perception.

The component of experience, or training, in machine learning often refers to data that is randomly generated. The task of the learner is to process such randomly generated examples towards drawing conclusions that hold for the environment from which these examples are picked. This description of machine learning highlights its close relationship with statistics. Indeed there is a lot in common between the two disciplines, in terms of both the goals and techniques used. There are, however, a few significant differences of emphasis; If a doctor comes up with the hypothesis that there is a correlation between smoking and heart disease, its the statistician's role to view samples of patients and check the validity of that hypothesis (this is the common statistical task of hypothesis testing). In contrast, machine learning aims to use the data gathered from samples of patients to come up with a description of the causes of heart disease. The hope is that automated techniques may be able to figure out meaningful patterns (or hypotheses) that may have been missed by the human observer.

In contrast with traditional statistics, in machine learning in general, and in this book in particular, algorithmic considerations play a major role. Machine learning is about the execution of learning by computers, hence algorithmic issues are pivotal. We develop algorithms to perform the learning tasks and are concerned with their computational efficiency. Another difference is that while statistics is often interested in asymptotic behavior (like the convergence of sample-based statistical estimates as the sample sizes grow to infinity), the theory of machine learning focuses on finite sample bounds. Namely, given the size of available samples, machine learning theory aims to figure out the degree of accuracy that a learner can expect based on such samples.

There are further differences between these two disciplines, of which we shall mention only one more here. While in statistics it is common to work under the assumption of certain pre-subscribed data models (such as assuming the normality of data-generating distributions, or the linearity of functional dependencies), in machine learning the emphasis is on working under a "distribution-free" setting, where the learner assumes as little as possible about the nature of the data distribution and allows the learning algorithm to figure out which models best approximate the data generating process. A precise discussion of this issue requires some technical preliminaries, and we will come back to it along the book, and in particular in Chapter 5.

1.5 How to read this book

The first part of the book provides the basic theoretical principles that underlie machine learning. In a sense, this is the foundation upon which the rest of the

book is built. This part could serve as a basis for a mini-course on the theoretical foundations of ML.

The second part of the book introduces the most commonly used algorithmic approaches to supervised machine learning. A subset of these chapters may also be used for introducing machine learning in a general AI course to CS, Math, or Engineering students.

The third part of the book extends the scope of discussion from statistical classification to other learning models. It covers online learning, unsupervised learning, dimensionality reduction, generative models, and feature learning.

The fourth part of the book, Advanced Theory, is geared towards readers who have interest in research and provides the more technical mathematical techniques that serve to analyze and drive forward the field of theoretical machine learning.

The appendix provides some technical tools used in the book. In particular, we list basic results from measure concentration and linear algebra.

Few sections are marked by a 'star', which means they are addressed to more advanced students. Each chapter is concluded with a list of exercises. A solution manual is provided in the course website.

1.5.1 Possible course plans based on this book

A 14 week introduction course for graduate students:

1. Chapters 2-4.
2. Chapter 9 (without the VC calculation).
3. Chapters 5-6 (without proofs).
4. Chapter 10.
5. Chapters 7,11 (without proofs).
6. Chapters 12,13 (with some of the easier proofs).
7. Chapter 14 (with some of the easier proofs).
8. Chapter 15.
9. Chapter 16.
10. Chapter 18.
11. Chapter 22.
12. Chapter 23 (without proofs for compressed sensing).
13. Chapter 24.
14. Chapter 25.

A 14 week advanced course for graduate students:

1. Chapters 26,27.
2. (continued)
3. Chapters 6,28.
4. Chapter 7.
5. Chapter 31.

6. Chapter 30.
7. Chapters 12, 13.
8. Chapter 14.
9. Chapter 8.
10. Chapter 17.
11. Chapter 29.
12. Chapter 19.
13. Chapter 20.
14. Chapter 21.

1.6 Notation

Most of the notation we use throughout the book is either standard or defined on the spot. In this section we describe our main conventions and provide a table summarizing our notation (Table 1.1). The reader is encouraged to skip this section and come back to it if during the reading of the book some notation is unclear.

We denote scalars and abstract objects with lower case letters (e.g. x and λ). Often, we would like to emphasize that some object is a vector and then we use boldface letters (e.g. $\mathbf{x}$ and $\boldsymbol{\lambda}$). The ith element of a vector $\mathbf{x}$ is denoted by x_i. We use upper case letters to denote matrices, sets, and sequences. The meaning should be clear from the context. As we will see momentarily, the input of a learning algorithm is a sequence of training examples. We denote by z an abstract example and by $S = z_1, \ldots, z_m$ a sequence of m examples. Historically, S is often referred to as a training *set*, however, we will always assume that S is a *sequence* rather than a set. A sequence of m vectors is denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_m$. The ith element of $\mathbf{x}_t$ is denoted by $x_{t,i}$.

Throughout the book, we make use of basic notions from probability. We denote by $\mathcal{D}$ a distribution over some set,[2] for example Z. We use the notation $z \sim \mathcal{D}$ to denote that z is sampled according to $\mathcal{D}$. Given a random variable $f : Z \to \mathbb{R}$, its expected value is denoted by $\mathbb{E}_{z\sim\mathcal{D}}[f(z)]$. We sometimes use the shorthand $\mathbb{E}[f]$ when the dependence on z is clear from the context. For $f : Z \to \{\texttt{true}, \texttt{false}\}$ we also use $\mathbb{P}_{z\sim\mathcal{D}}[f(z)]$ to denote $\mathcal{D}(\{z : f(z) = \texttt{true}\})$. In the next chapter we will also introduce the notation $\mathcal{D}^m$ to denote the probability over Z^m induced by sampling $(z_1, \ldots, z_m)$ where each point z_i is sampled from $\mathcal{D}$ independently of the other points.

In general, we made an effort to avoid asymptotic notation. However, we occasionally use it to clarify the main results. In particular, given $f : \mathbb{R} \to \mathbb{R}_+$ and $g : \mathbb{R} \to \mathbb{R}_+$ we write $f = O(g)$ if there exist $x_0, \alpha \in \mathbb{R}_+$ such that for all $x > x_0$ we have $f(x) \leq \alpha g(x)$. We write $f = o(g)$ if for every $\alpha > 0$ there exists

[2] To be mathematically precise, $\mathcal{D}$ should be defined over some σ-algebra of subsets of Z. The user who is not familiar with measure theory can skip the few footnotes and remarks regarding more formal measurability definitions and assumptions.

x_0 such that for all $x > x_0$ we have $f(x) \leq \alpha g(x)$. We write $f = \Omega(g)$ if there exist $x_0, \alpha \in \mathbb{R}_+$ such that for all $x > x_0$ we have $f(x) \geq \alpha g(x)$. The notation $f = \omega(g)$ is defined analogously. The notation $f = \Theta(g)$ means that $f = O(g)$ and $g = O(f)$. Finally, the notation $f = \tilde{O}(g)$ means that there exists $k \in \mathbb{N}$ such that $f(x) = O(g(x) \log^k(g(x)))$.

The inner product between vectors $\mathbf{x}$ and $\mathbf{w}$ is denoted by $\langle \mathbf{x}, \mathbf{w} \rangle$. Whenever we do not specify the vector space we assume that it is the d-dimensional Euclidean space and then $\langle \mathbf{x}, \mathbf{w} \rangle = \sum_{i=1}^{d} x_i w_i$. The Euclidean (or ℓ_2) norm of a vector $\mathbf{w}$ is $\|\mathbf{w}\|_2 = \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}$. We omit the subscript from the ℓ_2 norm when it is clear from the context. We also use other ℓ_p norms, $\|\mathbf{w}\|_p = \left(\sum_i |w_i|^p\right)^{1/p}$, and in particular $\|\mathbf{w}\|_1 = \sum_i |w_i|$ and $\|\mathbf{w}\|_\infty = \max_i |w_i|$.

We use the notation $\min_{x \in C} f(x)$ to denote the minimum value of the set $\{f(x) : x \in C\}$. To be mathematically more precise, we should use $\inf_{x \in C} f(x)$ whenever the minimum is not achievable. However, in the context of this book the distinction between infimum and minimum is often of little interest. Hence, to simplify the presentation, we sometimes use the min notation even when inf is more adequate. An analogous remark applies to max vs. sup.

symbol	meaning
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^d$	the set of d dimensional vectors over $\mathbb{R}$
$\mathbb{R}_+$	the set of non-negative real numbers
$\mathbb{N}$	the set of natural numbers
$O, o, \Theta, \omega, \Omega, \tilde{O}$	asymptotic notation (see text)
$\mathbb{1}_{[\text{Boolean expression}]}$	indicator function (equals 1 if expression is true and 0 o.w.)
$[a]_+$	$= \max\{0, a\}$
$[n]$	the set $\{1, \ldots, n\}$ (for $n \in \mathbb{N}$)
$\mathbf{x}, \mathbf{v}, \mathbf{w}$	(column) vectors
x_i, v_i, w_i	the i'th element of a vector
$\langle \mathbf{x}, \mathbf{v} \rangle$	$= \sum_{i=1}^d x_i v_i$ (inner product)
$\|\mathbf{x}\|_2$ or $\|\mathbf{x}\|$	$= \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ (the ℓ_2 norm of $\mathbf{x}$)
$\|\mathbf{x}\|_1$	$= \sum_{i=1}^d \lvert x_i \rvert$ (the ℓ_1 norm of $\mathbf{x}$)
$\|\mathbf{x}\|_\infty$	$= \max_i \lvert x_i \rvert$ (the ℓ_∞ norm of $\mathbf{x}$)
$\|\mathbf{x}\|_0$	the number of non-zero elements of $\mathbf{x}$
$A \in \mathbb{R}^{d,k}$	a $d \times k$ matrix over $\mathbb{R}$
$A^\top$	the transpose of A
$A_{i,j}$	the (i, j) element of A
$\mathbf{x}\mathbf{x}^\top$	the $d \times d$ matrix A s.t. $A_{i,j} = x_i x_j$ (where $\mathbf{x} \in \mathbb{R}^d$)
$\mathbf{x}_1, \ldots, \mathbf{x}_m$	a sequence of m vectors
$x_{i,j}$	the j'th element of the i'th vector in the sequence
$\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$	the values of a vector $\mathbf{w}$ during an iterative algorithm
$w_i^{(t)}$	the i'th element of the vector $\mathbf{w}^{(t)}$
$\mathcal{X}$	instances domain (a set)
$\mathcal{Y}$	labels domain (a set)
Z	examples domain (a set)
$\mathcal{H}$	hypothesis class (a set)
$\ell : \mathcal{H} \times Z \to \mathbb{R}_+$	loss function
$\mathcal{D}$	a distribution over some set (usually over Z or over $\mathcal{X}$)
$\mathcal{D}(A)$	the probability of a set $A \subseteq Z$ according to $\mathcal{D}$
$z \sim \mathcal{D}$	sampling z according to $\mathcal{D}$
$S = z_1, \ldots, z_m$	a sequence of m examples
$S \sim \mathcal{D}^m$	sampling $S = z_1, \ldots, z_m$ i.i.d. according to $\mathcal{D}$
$\mathbb{P}, \mathbb{E}$	probability and expectation of a random variable
$\mathbb{P}_{z \sim \mathcal{D}}[f(z)]$	$= \mathcal{D}(\{z : f(z) = \texttt{true}\})$ for $f : Z \to \{\texttt{true}, \texttt{false}\}$
$\mathbb{E}_{z \sim \mathcal{D}}[f(z)]$	expectation of the random variable $f : Z \to \mathbb{R}$
$N(\mu, C)$	Gaussian distribution with expectation μ and covariance C
$f'(x)$	the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at x
$f''(x)$	the second derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at x
$\frac{\partial f(\mathbf{w})}{\partial w_i}$	the partial derivative of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{w}$ w.r.t. w_i
$\nabla f(\mathbf{w})$	the gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{w}$
$\partial f(\mathbf{w})$	the differential set of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{w}$
$\min_{x \in C} f(x)$	$= \min\{f(x) : x \in C\}$ (minimal value of f over C)
$\max_{x \in C} f(x)$	$= \max\{f(x) : x \in C\}$ (maximal value of f over C)
$\operatorname{argmin}_{x \in C} f(x)$	the set $\{x \in C : f(x) = \min_{z \in C} f(z)\}$
$\operatorname{argmax}_{x \in C} f(x)$	the set $\{x \in C : f(x) = \max_{z \in C} f(z)\}$
log	the natural logarithm

Table 1.1 Summary of notation

Part I

Foundations

2 A Gentle Start

Let us begin our mathematical analysis by showing how successful learning can be achieved in a relatively simplistic setting. Imagine you have just arrived in some small Pacific island. You soon find out that papayas are a significant ingredient in the local diet. However, you have never before tasted papayas. You have to learn how to predict whether a papaya you see in the market is tasty or not. First, you need to decide which features of a papaya should your prediction be based on. Based on your previous experience with other fruits, you decide to use two features; the papaya's color, ranging from dark green, through orange and red to dark brown, and the papaya's softness, ranging from rock hard to mushy. Your input for figuring out your prediction rule is a sample of papayas that you have examined for color and softness and then tasted and found out if they were tasty or not. Let us analyze this task as a demonstration of the considerations involved in learning problems.

Our first step is to describe a formal model aimed to capture such learning tasks.

2.1 A Formal Model - the Statistical Learning Framework

The learner's input: In the basic statistical learning setting, the learner has access to the following:

Domain set: An arbitrary set, $\mathcal{X}$. This is the set of objects that we may wish to label. For example, in the papaya learning problem mentioned before, the domain set will be the set of all papayas. Usually, these domain points will be represented by a vector of *features* (like the papaya's color and softness). We also refer to domain points as *instances* and to $\mathcal{X}$ as instance space.

Label set: For our current discussion, we will restrict the label set to be a two-element set, usually, $\{0, 1\}$ or $\{-1, +1\}$. Let $\mathcal{Y}$ denote our set of possible labels. For our papayas example, let $\mathcal{Y}$ be $\{0, 1\}$, where 1 represents being tasty and 0 stands for being not-tasty.

Training data: $S = ((x_1, y_1) \ldots (x_m, y_m))$ is a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$. That is, a sequence of labeled domain points. This is

the input that the learner has access to (like a set of papayas that have been tasted and their color, softness and tastiness). Such labeled examples are often called *training examples*. We sometimes also refer to S as a *training set*.[1]

The learner's output: The learner is requested to output a *prediction rule*, $h : \mathcal{X} \to \mathcal{Y}$. This function is also called a *predictor*, a *hypothesis*, or a *classifier*. The predictor can be used to predict the label of new domain points. In our papayas example, it is a rule that our learner will employ to predict whether future papayas he examines in the farmers market are going to be tasty or not. We use the notation $A(S)$ to denote the hypothesis that a learning algorithm, A, returns upon receiving the training sequence S.

A simple data-generation model We now explain how the training data is generated. First, we assume that the instances (the papayas we encounter) are generated by some probability distribution (in this case, representing the environment). Let us denote that probability distribution over $\mathcal{X}$ by $\mathcal{D}$. It is important to note that we do not assume that the learner knows anything about this distribution. For the type of learning tasks we discuss, this could be any arbitrary probability distribution. As to the labels, in the current discussion we assume that there is some "correct" labeling function, $f : \mathcal{X} \to \mathcal{Y}$, and that $y_i = f(x_i)$ for all i. This assumption will be relaxed in the next chapter. The labeling function is unknown to the learner. In fact, this is just what the learner is trying to figure out. In summary, each pair in the training data S is generated by first sampling a point x_i according to $\mathcal{D}$ and then labeling it by f.

Measures of success: We define the *error of a classifier* to be the probability that it does not predict the correct label on a random data point generated by the aforementioned underlying distribution. That is, the error of h is the probability to draw a random instance x, according to the distribution $\mathcal{D}$, such that $h(x)$ does not equal to $f(x)$.

Formally, given a domain subset[2], $A \subset \mathcal{X}$, the probability distribution, $\mathcal{D}$, assigns a number, $\mathcal{D}(A)$, which determines how likely it is to observe a point $x \in A$. In many cases, we refer to A as an event and express it using a function $\pi : \mathcal{X} \to \{0, 1\}$, namely, $A = \{x \in \mathcal{X} : \pi(x) = 1\}$. In that case, we also use the notation $\mathbb{P}_{x \sim \mathcal{D}}[\pi(x)]$ to express $\mathcal{D}(A)$.

We define the error of a prediction rule, $h : \mathcal{X} \to \mathcal{Y}$ to be:

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}) \,. \tag{2.1}$$

[1] Despite the "set" notation, S is a sequence. In particular, the same example may appear twice in S and some algorithms can take into account the order of examples in S.

[2] Strictly speaking, we should be more careful and require that A is a member of some σ-algebra of subsets of $\mathcal{X}$, over which $\mathcal{D}$ is defined. We will formally define our measurability assumptions in the next chapter.

That is, the error of such h is the probability to randomly choose an example x for which $h(x) \neq f(x)$. The subscript $(\mathcal{D}, f)$ indicates that the error is measured with respect to the probability distribution $\mathcal{D}$ and the correct labeling function f. We omit this subscript when it is clear from the context. $L_{(\mathcal{D},f)}(h)$ has several synonymous names such as the *generalization error*, the *risk*, or the *true error* of h, and we will use these names interchangeably throughout the book. We use the letter L for the error, since we view this error as the *loss* of the learner. We will later also discuss other possible formulations of such loss.

A note about the information available to the learner The learner is blind to the underlying distribution $\mathcal{D}$ over the world and to the labeling function f. In our papayas example, we have just arrived to a new island and we have no clue as to how papayas are distributed and how to predict their tastiness. The only way the learner can interact with the environment is through observing the training set.

In the next section we describe a simple learning paradigm for the above setup and analyze its performance.

2.2 Empirical Risk Minimization

As mentioned above, a learning algorithm receives as input a training set S, sampled from an unknown distribution $\mathcal{D}$ and labeled by some target function f, and should output a predictor $h_S : \mathcal{X} \to \mathcal{Y}$ (the subscript S emphasizes the fact that the output predictor depends on S). The goal of the algorithm is to find h_S that minimizes the error with respect to the unknown $\mathcal{D}$ and f.

Since the learner does not know what $\mathcal{D}$ and f are, the true error is not directly available to the learner. A useful notion of error that can be calculated by the learner is the *training error* - the error the classifier incurs over the training sample:

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}, \tag{2.2}$$

where $[m] = \{1, \ldots, m\}$.

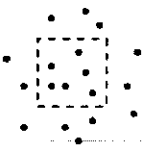
The terms *empirical error*, or *empirical risk*, are often used interchangeably for this error.

Since the training sample is the snapshot of the world that is available to the learner, it makes sense to search for a solution that works well on that data. This learning paradigm – coming up with a predictor h that minimizes $L_S(h)$ – is called *Empirical Risk Minimization* or ERM for short.

2.2.1 Something may go wrong - overfitting

Although the ERM rule seems very natural, without being careful, this approach may fail miserably.

To demonstrate such a failure, let us go back to the problem of learning to predict the taste of a papaya based on its softness and color. Consider a sample as depicted below:



Assume that the probability distribution $\mathcal{D}$ is such that instances are distributed uniformly within the gray square and the labeling function, f, determines the label to be 1 if the instance is within the inner blue square, and 0 otherwise. The area of the gray square in the picture is 2 and the area of the blue square is 1. Consider the following predictor:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{otherwise} \end{cases} . \tag{2.3}$$

While this predictor might seem rather artificial, in Exercise 1 we show a natural representation of it using polynomials. Clearly, no matter what the sample is, $L_S(h_S) = 0$, and therefore this predictor may be chosen by an ERM algorithm (it is one of the empirical-minimum-cost hypotheses, no classifier can have smaller error). On the other hand, the true error of any classifier that predicts the label 1 only on a finite number of instances is, in this case, $1/2$. Thus, $L_{\mathcal{D}}(h_S) = 1/2$. We have found a predictor whose performance on the training set is excellent, yet its performance on the true "world" is very poor. This phenomenon is called *overfitting*. Intuitively, overfitting occurs when our hypothesis fits the training data "too well" (perhaps like the everyday experience that a person that provides a perfect detailed explanation for each of his single actions may raise suspicion).

2.3 Empirical Risk Minimization with Inductive Bias

We have just demonstrated that the ERM rule might lead to overfitting. Rather than giving up on the ERM paradigm, we will look for ways to rectify it. We will search for conditions under which there is a guarantee that ERM does not overfit. Namely, conditions under which when the ERM predictor has good performance with respect to the training data, it is also highly likely to perform well over the underlying data distribution.

A common solution is to apply the ERM learning rule over a restricted search space. Formally, the learner should choose in advance (before seeing the data) a

set of predictors. This set is called a *hypothesis class* and is denoted by $\mathcal{H}$. Each $h \in \mathcal{H}$ is a function mapping from $\mathcal{X}$ to $\mathcal{Y}$. For a given class $\mathcal{H}$, and a training sample, S, the $\text{ERM}_\mathcal{H}$ learner uses the ERM rule to choose a predictor $h \in \mathcal{H}$, with as low as possible error over S. Formally,

$$\text{ERM}_\mathcal{H}(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h) \ ,$$

where argmin stands for the set of hypotheses in $\mathcal{H}$ that achieve the minimum value of $L_S(h)$ over $\mathcal{H}$. By restricting the learner to choosing a predictor from $\mathcal{H}$, we *bias* it toward a particular set of predictors. Such restrictions are often called an *inductive bias*. Since the choice of such a restriction is determined before the learner sees the training data, it should ideally be based on some prior knowledge about the problem to be learnt. For example, for the Papaya taste prediction problem we may choose the class $\mathcal{H}$ to be the set of predictors which are determined by axis aligned rectangles (in the space determined by the color and softness coordinates). We will later show that $\text{ERM}_\mathcal{H}$ over this class is guaranteed not to overfit. On the other hand, the example of overfitting that we have seen above, demonstrates that choosing $\mathcal{H}$ to be a class of predictors that includes all functions that assign the value 1 to a finite set of domain points, does not suffice to guarantee that $\text{ERM}_\mathcal{H}$ will not overfit.

A fundamental question in learning theory is, over which hypothesis classes $\text{ERM}_\mathcal{H}$ learning will not result in overfitting. We will study this question later in the book.

Intuitively, choosing a more restricted hypothesis class better protects us against overfitting but at the same time might cause us a stronger inductive bias. We will get back to this fundamental tradeoff later.

2.3.1 Finite hypothesis classes

The simplest type of restriction on a class is imposing an upper bound on its size (that is, the number of predictors h in $\mathcal{H}$). In this section, we show that if $\mathcal{H}$ is a finite class then $\text{ERM}_\mathcal{H}$ will not overfit, provided it is based on a sufficiently large training sample (this size requirement will depend on the size of $\mathcal{H}$).

Limiting the learner to prediction rules within some finite hypothesis class may be considered as a reasonably mild restriction. For example, $\mathcal{H}$ can be the set of all predictors that can be implemented by a C++ program written in at most 10^9 bits of code. In our papayas example, we mentioned previously the class of axis aligned rectangles. While this is an infinite class, if we discretize the representation of real numbers, say by using a 64 bits floating-point representation, the hypothesis class becomes a finite class.

Let us now analyze the performance of the $\text{ERM}_\mathcal{H}$ learning rule assuming that $\mathcal{H}$ is a finite class. For a training sample, S, labeled according to some $f : \mathcal{X} \to \mathcal{Y}$, let h_S denote a result of applying ERM_H to S. Namely,

$$h_S \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h) \ , \tag{2.4}$$

In this chapter, we make the following simplifying assumption (which would be relaxed in the next chapter).

DEFINITION 2.1 (The Realizability assumption) There exists $h^\star \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^\star) = 0$. Note that this assumption implies that with probability 1 over random samples, S, where the instances of S are sampled according to $\mathcal{D}$ and are labeled by f, we have $L_S(h^\star) = 0$.

The realizability assumption implies that for every ERM hypothesis we have that[3] $L_S(h_S) = 0$. However, we are interested in the *true* risk of h_S, $L_{(\mathcal{D},f)}(h_S)$, rather than its empirical risk.

Clearly, any guarantee on the error with respect to the underlying distribution, $\mathcal{D}$, for an algorithm that has access only to a sample S, should depend on the relationship between $\mathcal{D}$ and S. The common assumption in statistical machine learning is that the training sample S is generated by sampling points from the distribution $\mathcal{D}$ independently of each other. Formally,

The i.i.d. assumption: The examples in the training set are independently and identically distributed (i.i.d.) according to the distribution $\mathcal{D}$. That is, every x_i in S is freshly sampled according to $\mathcal{D}$ and then labeled according to the labeling function, f. We denote this assumption by $S \sim \mathcal{D}^m$ where m is the size of S, and $\mathcal{D}^m$ denotes the probability over m-tuples induced by applying $\mathcal{D}$ to pick each element of the tuple independently of the other members of the tuple.

Intuitively, the training set S is a window through which the learner gets partial information about the distribution $\mathcal{D}$ over the world and the labeling function, f. The larger the sample gets, the more likely it is to reflect more accurately the distribution and labeling used to generate it.

Since $L_{(\mathcal{D},f)}(h_S)$ depends on the training set, S, and that training set is picked by a random process, there is randomness in the choice of the predictor h_S and, consequently, in the risk $L_{(\mathcal{D},f)}(h_S)$. Formally, we say that it is a random variable. It is not realistic to expect that with full certainty S will suffice to direct the learner towards a good classifier (from the point of view of $\mathcal{D}$), as there is always some probability that the sampled training data happens to be very non-representative of the underlying $\mathcal{D}$. If we go back to the papaya-tasting example, there is always some (small) chance that all the papayas we have happened to taste were not tasty, in spite of the fact that, say 70% of the papayas in our island are tasty. In such a case, $\text{ERM}_{\mathcal{H}}(S)$ may be the constant function that labels every papaya as 'not tasty' (and has 70% error on the true distribution of papapyas in the island). We will therefore address the *probability* to sample a training set for which $L_{(\mathcal{D},f)}(h_S)$ is not too large. Usually, we denote the probability of getting a non-representative sample by δ, and call $(1-\delta)$ the *confidence parameter* of our prediction.

[3] Mathematically speaking, this holds with probability 1. To simplify the presentation, we sometimes omit the "with probability 1" specifier.

On top of that, since we cannot guarantee perfect label prediction, we introduce another parameter for the quality of prediction, the *accuracy parameter*, commonly denoted by ϵ. We interpret the event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ as a failure of the learner, while if $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$ we view the output of the algorithm as an approximately correct predictor. Therefore (fixing some labeling function $f : \mathcal{X} \to \mathcal{Y}$), we are interested in upper bounding the probability to sample m-tuple of instances that will lead to failure of the learner. Formally, let $S|_x = (x_1, \ldots, x_m)$ be the instances of the training set. We would like to upper bound

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) .$$

Let $\mathcal{H}_B$ be the set of "bad" hypotheses, that is

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

In addition, let

$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

be the set of misleading samples. Namely, for every $S|_x \in M$, there is a "bad" hypothesis, $h \in \mathcal{H}_B$, which looks like a "good" hypothesis on $S|_x$. Now, recall that we would like to bound the probability of the event $L_{(\mathcal{D},f)}(h_S) > \epsilon$. But, since the realizability assumption implies that $L_S(h_S) = 0$, it follows that the event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ can only happen if for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$. In other words, this event will only happen if our sample is in the set of misleading samples, M. Formally, we have shown that

$$\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \leq M .$$

Note that we can rewrite M as

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\} . \tag{2.5}$$

Hence,

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m(\cup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}) . \tag{2.6}$$

Next, we upper bound the right-hand side of the above using the *union bound* - a basic property of probabilities.

LEMMA 2.2 (Union bound) *For any two sets A, B and a distribution $\mathcal{D}$ we have*

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B) .$$

Applying the union bound to the right-hand side of Equation (2.6) yields

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) . \tag{2.7}$$

Next, let us bound each summand of the right-hand side of the above. Fix some "bad" hypothesis $h \in \mathcal{H}_B$. The event $L_S(h) = 0$ is equivalent to the event

$\forall i, h(x_i) = f(x_i)$. Since the examples in the training set are sampled i.i.d. we get that

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) = \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) = \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) \ . \tag{2.8}$$

For each individual sampling of an element of the training set we have,

$$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \epsilon \ ,$$

where the last inequality follows from the fact that $h \in \mathcal{H}_B$. Combining the above with Equation (2.8) and using the inequality $1 - \epsilon \leq e^{-\epsilon}$ we obtain that for every $h \in \mathcal{H}_B$,

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1 - \epsilon)^m \leq e^{-\epsilon m} \ . \tag{2.9}$$

Combining the above with Equation (2.7) we conclude that

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \ \leq \ |\mathcal{H}_B| e^{-\epsilon m} \ \leq \ |\mathcal{H}| e^{-\epsilon m} \ .$$

A graphical illustration which explains how we used the union bound is given in Figure 2.1.

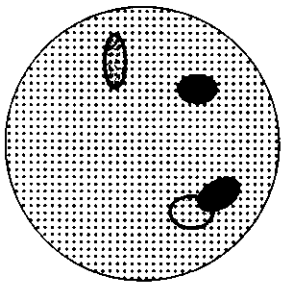


Figure 2.1 Each point in the large circle represents a possible m-tuple of instances. Each colored oval represents the set of 'misleading' m-tuple of instances for some 'bad' predictor $h \in \mathcal{H}_B$. The ERM can potentially overfit whenever it gets a misleading training set S. That is, for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$. Equation (2.9) guarantees that for each individual bad hypothesis, $h \in \mathcal{H}_B$, at most $(1 - \epsilon)^m$-fraction of the training sets would be misleading. In particular, the larger m is, the smaller each of these colored ovals becomes. The union bound formalizes the fact that the area representing the training sets which are misleading with respect to some $h \in \mathcal{H}_B$ (that is, the training sets in M) is at most the sum of the areas of the colored ovals. Therefore, it is bounded by $|\mathcal{H}_B|$ times the maximum size of a colored oval. Any sample S outside the colored ovals cannot cause the ERM rule to overfit.

COROLLARY 2.3 *Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer that satisfies*

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} .$$

Then, for any labeling function, f, and for any distribution, $\mathcal{D}$, for which the realizability assumption holds (that is, for some $h \in \mathcal{H}$, $L_{(\mathcal{D},f)}(h) = 0$), with probability of at least $1-\delta$ over the choice of an i.i.d. sample S of size m we have that for every ERM hypothesis, h_S, it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon .$$

The above corollary tells us that for a sufficiently large m, the $\text{ERM}_{\mathcal{H}}$ rule over a finite hypothesis class will be *probably* (with confidence $1-\delta$) *approximately* (up to an error of ϵ) correct. In the next chapter we formally define the model of Probably Approximately Correct (PAC) learning.

2.4 Exercises

1. **Overfitting of polynomial matching:** We have shown that the predictor defined in Equation (2.3) leads to overfitting. While this predictor seems to be very unnatural, the goal of this exercise is to show that it can be described as a thresholded polynomial. That is, show that given a training set $S = \{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^m \subseteq (\mathbb{R}^d \times \{0,1\})^m$, there exists a polynomial p_S such that $h_S(\mathbf{x}) = 1$ if and only if $p_S(\mathbf{x}) \geq 0$, where h_S is as defined in Equation (2.3). It follows that learning the class of all thresholded polynomials using the ERM rule may lead to overfitting.
2. Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. Fix some $h \in \mathcal{H}$. Show that the expected value of $L_S(h)$ over the choice of $S|_x$ equals to $L_{(\mathcal{D},f)}(h)$, namely,

$$\mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}[L_S(h)] = L_{(\mathcal{D},f)}(h) .$$

3. **Axis-aligned rectangles:** An axis-aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1, a_2 \leq b_2$, define the classifier $h_{(a_1,b_1,a_2,b_2)}$ by

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases} . \tag{2.10}$$

The class of all axis-aligned rectangles in the plane is defined as

$$\mathcal{H}^2_{\text{rec}} = \{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1, \text{ and } a_2 \leq b_2\}.$$

Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

1. Let A be the algorithm which returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM.
2. Show that if A receives a training set of size $\geq \frac{4\log(4/\delta)}{\epsilon}$ then, with probability of at least $1-\delta$ it returns a hypothesis with error of at most ϵ.
   Hint: Fix some distribution $\mathcal{D}$ over $\mathcal{X}$, let $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels, and let f be the corresponding hypothesis. Let $a_1 \geq a_1^*$ be a number such that the probability mass (with respect to $\mathcal{D}$) of the rectangle $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$ is exactly $\epsilon/4$. Similarly, let b_1, a_2, b_2 be numbers such that the probability masses of the rectangles $R_2 = R(b_1, b_1^*, a_2^*, b_2^*), R_3 = R(a_1^*, b_1^*, a_2^*, a_2), R_4 = R(a_1^*, b_1^*, b_2, b_2^*)$ are all exactly $\epsilon/4$. Let $R(S)$ be the rectangle returned by A. See illustration in Figure 2.2.

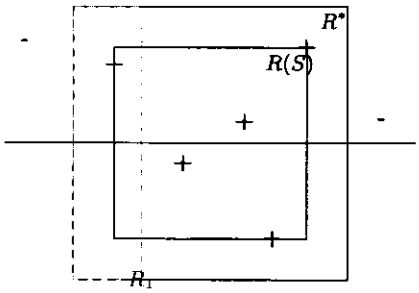


Figure 2.2 Axis aligned rectangles.

   - Show that $R(S) \subseteq R^*$.
   - Show that if S contains (positive) examples in all of the rectangles R_1, R_2, R_3, R_4, then the hypothesis returned by A has error of at most ϵ.
   - For each $i \in \{1, \ldots, 4\}$, upper bound the probability that S does not contain an example from R_i.
   - Use the union bound to conclude the argument.
3. Repeat the previous question for the class of axis aligned rectangles in $\mathbb{R}^d$.
4. Show that the runtime of applying the algorithm A mentioned above is polynomial in $d, 1/\epsilon$, and in $\log(1/\delta)$.

3 A Formal Learning Model

In this chapter we define our main formal learning model — the PAC learning model and its extensions. We will consider other notions of learnability in Chapter 7.

3.1 PAC Learning

In the previous chapter we have shown that for a finite hypothesis class, if the ERM rule with respect to that class is applied on a sufficiently large training sample (whose size is independent of the underlying distribution or labeling function) then the output hypothesis will be probably approximately correct. More generally, we now define *Probably Approximately Correct* (PAC) learning.

DEFINITION 3.1 (PAC learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

The definition of Probably Approximately Correct learnability contains two approximation parameters. The accuracy parameter ϵ determines how far can the output classifier be from the optimal one (this corresponds to the "approximately correct"), and a confidence parameter δ indicating how likely is the classifier to meet that accuracy requirement (corresponds to the "probably" part of "PAC"). Under the data access model that we are investigating, these approximations are inevitable. Since the training set is randomly generated, there may always be a small chance that it will happen to be non-informative (for example, there is always some chance that the training set will contain only one domain point, sampled over and over again). Furthermore, even when we are lucky enough to get a training sample that does faithfully represent $\mathcal{D}$, due to being just a finite sample, there may always be some fine details of $\mathcal{D}$ that it fails to reflect. Our

accuracy parameter, ϵ allows "forgiving" the learner's classifier for making minor errors.

Sample complexity

The function $m_{\mathcal{H}} : (0,1)^2 \to \mathbf{N}$ determines the *sample complexity* of learning $\mathcal{H}$. That is, how many examples are required to guarantee a probably approximately correct solution. The sample complexity is a function of the accuracy (ϵ) and confidence (δ) parameters. It also depends on properties of the hypothesis class $\mathcal{H}$ – for example, for a finite class we showed that the sample complexity depends on log the size of $\mathcal{H}$.

Note that if $\mathcal{H}$ is PAC learnable, there are many functions $m_{\mathcal{H}}$ that satisfy the requirements given in the definition of PAC learnability. Therefore, to be precise, we will define the sample complexity of learning $\mathcal{H}$ to be the "minimal function", in the sense that for any ϵ, δ, $m_{\mathcal{H}}(\epsilon, \delta)$ is the minimal integer that satisfies the requirements of PAC learning with accuracy ϵ and confidence δ.

Let us now recall the conclusion of the analysis of finite hypothesis classes from the previous chapter. It can be rephrased as stating:

COROLLARY 3.2 *Every finite hypothesis class is PAC learnable with sample complexity*

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

There are infinite classes that are learnable as well (see for example Exercise 3). Later on we will show that what determines the PAC learnability of a class is not its finiteness but rather a combinatorial measure called the *VC dimension*.

3.2 A More General Learning Model

The model we have just described can be readily generalized, so that it could be made relevant to a wider scope of learning tasks. We consider generalizations in two aspects:

Removing the realizability assumption

We have required that the learning algorithm succeeds on a pair of data distribution $\mathcal{D}$ and labeling function f provided that the realizability assumption is met. For practical learning tasks, this assumption may be too strong (can we really guarantee that there is a rectangle in the color-hardness space that *fully determines* which papayas are tasty?). In the next sub-section, we will describe the *agnostic PAC* model in which this realizability assumption is waived.

Learning problems beyond binary classification
The learning task that we have been discussing so far, has to do with predicting a binary label to a given example (like being tasty or not). However, many learning tasks take a different form. For example, one may wish to predict a real valued number (say, the temperature at 9pm tomorrow) or a label picked from a finite set of labels (like the topic of the main story in tomorrow's paper). It turns out that our analysis of learning can be readily extended to such and many other scenarios by allowing a variety of loss functions. We shall discuss that in Section 3.2.2 below.

3.2.1 Releasing the realizability assumption – Agnostic PAC learning

A more realistic model for the data-generating distribution:
Recall that the realizability assumption requires that there exists $h^* \in \mathcal{H}$ such that $\mathbb{P}_{x\sim\mathcal{D}}[h^*(x) = f(x)] = 1$. In many practical problems this assumption does not hold. Furthermore, it is maybe more realistic not to assume that the labels are fully determined by the features we measure on input elements (in the case of the papayas, it is plausible that two papayas of the same color and softness will have a different taste). In the following, we relax the realizability assumption by replacing the "target labeling function" with a more flexible notion, a data-labels generating distribution.

Formally, from now on, let $\mathcal{D}$ be a probability distribution over $\mathcal{X} \times \mathcal{Y}$, where, as before, $\mathcal{X}$ is our domain set and $\mathcal{Y}$ is a set of labels (usually we will consider $\mathcal{Y} = \{0, 1\}$). That is, $\mathcal{D}$ is a *joint distribution* over domain points and labels. One can view such a distribution as being composed of two parts; a distribution $\mathcal{D}_x$ over unlabeled domain points (sometimes called the *marginal distribution*) and a *conditional* probability over labels for each domain point, $\mathcal{D}((x, y)|x)$. In the papaya example, $\mathcal{D}_x$ determines the probability of encountering a papaya whose color and hardness fall in some color-hardness values domain, and the conditional probability is the probability that a papaya with color and hardness represented by x is tasty. Indeed, such modeling allows for two papayas that share the same color and hardness to belong to different taste categories.

The empirical and the true error revised:
For a probability distribution, $\mathcal{D}$, over $\mathcal{X} \times \mathcal{Y}$, one can measure how likely is h to make an error when labeled points are randomly drawn according to $\mathcal{D}$. We redefine the true error (or risk) of a prediction rule h to be

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x, y) : h(x) \neq y\}). \tag{3.1}$$

We would like to find a predictor, h, for which that error will be minimized. However, the learner does not know the data generating $\mathcal{D}$. What the learner does have access to is the training data, S. The definition of the empirical risk

remains the same as before, namely,

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}.$$

Given S, a learner can compute $L_S(h)$ for any function $h : X \to \{0,1\}$. Note that $L_S(h) = L_{D(\text{uniform over } S)}(h)$.

The goal:
We wish to find some hypothesis, $h : \mathcal{X} \to \mathcal{Y}$, that (probably approximately) minimizes the true risk, $L_D(h)$.

The Bayes optimal predictor.
Given any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the best label predicting function from $\mathcal{X}$ to $\{0,1\}$ will be

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

It is easy to verify (see Exercise 7) that for every probability distribution $\mathcal{D}$, the Bayes optimal predictor $f_{\mathcal{D}}$ is optimal, in the sense that no other classifier, $g : \mathcal{X} \to \{0,1\}$ has a lower error. That is, for every classifier g, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Unfortunately, since we do not know $\mathcal{D}$, we cannot utilize this optimal predictor $f_{\mathcal{D}}$. What the learner does have access to is the training sample. We can now present the formal definition of agnostic PAC learnability, which is a natural extension of the definition of PAC learnability to the more realistic, non-realizable, learning setup we have just discussed.

Clearly, we cannot hope that the learning algorithm will find a hypothesis whose error is smaller than the minimal possible error, that of the Bayes predictor. Furthermore, as we shall prove later, once we make no prior assumptions about the data-generating distribution, no algorithm can be guaranteed to find a predictor which is as good as the Bayes optimal one. Instead, we require that the learning algorithm will find a predictor whose error is not much larger than the best possible error of a predictor in some given benchmark hypothesis class. Of course, the strength of such a requirement depends on the choice of that hypothesis class.

DEFINITION 3.3 (agnostic PAC learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

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

Clearly, if the realizability assumption holds, agnostic PAC learning provides the same guarantee as PAC learning. In that sense, agnostic PAC learning generalizes the definition of PAC learning. When the realizability assumption does not hold, no learner can guarantee an arbitrarily small error. Nevertheless, under the definition of agnostic PAC learning, a learner can still declare success if its error is not much larger than the best error achievable by a predictor from the class $\mathcal{H}$. This is in contrast to PAC learning in which the learner is required to achieve a small error in absolute term and not relative to the best error achievable by the hypothesis class.

3.2.2 The scope of learning problems modeled

We next extend our model so that it could be applied to a wide variety of learning tasks. Let us consider some examples of different learning tasks.

- **Multiclass classification** Our classification does not have to be binary. Take for example the task of document classification: We wish to design a program that will be able to classify given documents according to topics (e.g., News, Sports, Biology, Medicine, etc.). A learning algorithm for such a task, will have access to examples of correctly classified documents, and, based on these examples, should output a program that can take as input a new document and output a topic classification for that document. Here, the *domain set* is the set of all potential documents. Once again, we would usually represent documents by a set of *features* which could include counts of different key words in the document, as well as other possibly relevant features like the size of the document or its origin. The *label set* in this task would be the set of possible document topics (so $\mathcal{Y}$ will be some large finite set). Once we determine our domain and label sets, the other components of our framework look exactly the same as in the Papaya tasting example; Our *training sample* will be a finite sequence of (feature vector, label) pairs, the learner's output will be a function from the domain set to the label set, and, finally, for our measure of success, we can use the probability, over (document, topic) pairs, of the event that our predictor suggests a wrong label.
- **Regression** In this task, one wishes to find some simple *pattern* in the data - a functional relationship between the $\mathcal{X}$ and $\mathcal{Y}$ components of the data. For example, one wishes to find a linear function that best predicts a baby's birth weight based on ultrasound measures of his head circumference, abdominal circumference, and femur length. Here, our domain set $\mathcal{X}$ is some subset of $\mathbb{R}^3$ (the three ultrasound measurements) and the set of "labels", $\mathcal{Y}$, is the the set of real numbers (the weight in grams). In this context, it is more adequate to call $\mathcal{Y}$ the *target* set. Our training data as well as the learner's output, are as before (a finite sequence of (x, y) pairs, and a function from $\mathcal{X}$ to $\mathcal{Y}$, respectively). However, our measure of success is dif-

ferent. We may evaluate the quality of a hypothesis function, $h : \mathcal{X} \to \mathcal{Y}$ by the *expected square difference* between the true labels and their predicted values. Namely,

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}(h(x) - y)^2 \,. \tag{3.2}$$

To accommodate a wide range of learning tasks we generalize our formalism of the measure of success as follows:

Generalized Loss Functions

Given any set $\mathcal{H}$ (that plays the role of our hypotheses, or models) and some domain Z let ℓ be any function from $\mathcal{H} \times Z$ to the set of non-negative real numbers, $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$. We call such functions *loss functions*.

Note that for prediction problems, we have that $Z = \mathcal{X} \times \mathcal{Y}$. However, our notion of the loss function is generalized beyond prediction tasks, and therefore it allows Z to be any domain of examples (for instance, in unsupervised learning tasks such as the one described in Chapter 22, Z is not a product of an instance domain and a label domain).

We now define the *risk function* to be the expected loss of a classifier, $h \in \mathcal{H}$, with respect to a probability distribution D over Z, namely,

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h, z)]. \tag{3.3}$$

That is, we consider the expectation of the loss of h over objects z picked randomly according to $\mathcal{D}$. Similarly, we define the *empirical risk* to be the expected loss over a given sample $S = (z_1, \ldots, z_m) \in Z^m$, namely,

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i). \tag{3.4}$$

The loss functions used in the above examples of classification and regression tasks are as follows:

- $0-1$ **loss:** Here, our random variable z ranges over the set of pairs $\mathcal{X} \times \mathcal{Y}$ and the loss function is

$$\ell_{0-1}(h, (x, y)) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases}$$

  This loss function is used in binary or multiclass classification problems. One should note that, for a random variable, α, taking the values $\{0, 1\}$, $\mathbb{E}_{\alpha\sim D}[\alpha] = \mathbb{P}_{\alpha\sim D}[\alpha = 1]$. Consequently, for this loss function, the definitions of $L_{\mathcal{D}}(h)$ given in Equation (3.3) and Equation (3.1) coincides.
- **Square loss:** Here, our random variable z ranges over the set of pairs $\mathcal{X} \times \mathcal{Y}$ and the loss function is

$$\ell_{\text{sq}}(h, (x, y)) \stackrel{\text{def}}{=} (h(x) - y)^2 \,.$$

  This loss function is used in regression problems.

We will later see more examples of useful instantiations of loss functions.

To summarize, we formally define agnostic PAC learnability for general loss functions.

DEFINITION 3.4 (agnostic PAC learnability for general loss functions) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

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

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$.

Remark 3.1 (A note about measurability*) In the aforementioned definition, for every $h \in \mathcal{H}$, we view the function $\ell(h, \cdot) : Z \to \mathbb{R}_+$ as a random variable, and define $L_{\mathcal{D}}(h)$ to be the expected value of this random variable. For that, we need to require that the function $\ell(h, \cdot)$ is measurable. Formally, we assume that there is a σ-algebra of subsets of Z, over which the probability $\mathcal{D}$ is defined, and that the pre-image of every initial segment in $\mathbb{R}_+$ is in this σ-algebra. In the specific case of binary classification with the $0-1$ loss, the σ-algebra is over $\mathcal{X} \times \{0,1\}$ and our assumption on ℓ is equivalent to the assumption that for every h, the set $\{(x, h(x)) : x \in \mathcal{X}\}$ is in the σ-algebra.

Remark 3.2 (Proper vs. representation-independent learning*) In the above definition, we required that the algorithm will return a hypothesis from $\mathcal{H}$. In some situations, $\mathcal{H}$ is a subset of a set $\mathcal{H}'$, and the loss function can be naturally extended to be a function from $\mathcal{H}' \times Z$ to the reals. In this case, we may allow the algorithm to return a hypothesis $h' \in \mathcal{H}'$, as long as it satisfies the requirement $L_{\mathcal{D}}(h') \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$. Allowing the algorithm to output a hypothesis from $\mathcal{H}'$ is called *representation independent* learning, while proper learning is when the algorithm must output a hypothesis from $\mathcal{H}$. Representation independent learning is sometimes called "improper learning", although there is nothing improper in representation independent learning.

3.3 Summary

In this chapter we defined our main formal learning model — PAC learning. The basic model relies on the realizability assumption, while the agnostic variant does not impose any restrictions on the underlying distribution over the examples. We also generalized the PAC model to arbitrary loss functions. We will sometime

refer to the most general model simply as PAC learning, omitting the "agnostic" prefix and letting the reader infer what is the underlying loss function from the context. When we would like to emphasize that we are dealing with the original PAC setting we mention that the realizability assumption holds. In Chapter 7 we will discuss other notions of learnability.

3.4 Bibliographic Remarks

Our most general definition of agnostic PAC learning with general loss functions follows the works of Vladimir Vapnik and Alexey Chervonenkis (Vapnik & Chervonenkis 1971). In particular, we follow Vapnik's general setting of learning (Vapnik 1982, Vapnik 1992, Vapnik 1995, Vapnik 1998).

PAC learning was introduced by Valiant (1984). Valiant has been named the winner of the 2010 Turing Award for the introduction of the PAC model. Valiant's definition requires that the sample complexity will be polynomial in $1/\epsilon$ and in $1/\delta$, as well as in the representation size of hypotheses in the class (see also Kearns & Vazirani (1994)). As we will see in Chapter 6, if a problem is at all PAC learnable then the sample complexity depends polynomially on $1/\epsilon$ and $\log(1/\delta)$. Valiant's definition also requires that the *runtime* of the learning algorithm would be polynomial in these quantities. In contrast, we chose to distinguish between the statistical aspect of learning and the computational aspect of learning. We will elaborate on the computational aspect later on in Chapter 8, where we introduce the full PAC learning model of Valiant. For expository reasons, we use the term PAC learning even when we ignore the runtime aspect of learning. Finally, the formalization of agnostic PAC learning is due to Haussler (1992).

3.5 Exercises

1. **Monotonicity of sample complexity:** Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot,\cdot)$. Show that $m_{\mathcal{H}}$ is monotonically non-increasing in each of its parameters. That is, show that given $\delta \in (0,1)$, and given $0 < \epsilon_1 \leq \epsilon_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon_1,\delta) \geq m_{\mathcal{H}}(\epsilon_2,\delta)$. Similarly, show that given $\epsilon \in (0,1)$, and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon,\delta_1) \geq m_{\mathcal{H}}(\epsilon,\delta_2)$.
2. Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{\text{Singleton}} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, where, for each $z \in \mathcal{X}$, h_z is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely $\forall x \in X$, $h^-(x) = 0$. The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.
   1. Describe an algorithm which implements the ERM rule for learning $\mathcal{H}_{\text{Singleton}}$ in the realizable setup.

2. Show that $\mathcal{H}_{\text{Singleton}}$ is PAC learnable. Provide an upper bound on the sample complexity.
3. Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane, i.e., $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$. Prove that $\mathcal{H}$ is PAC learnable (assume realizability), and its sample complexity is bounded by

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

4. In this question, we study the hypothesis class of *boolean conjunctions* defined as follows. The instance space is $\mathcal{X} = \{0, 1\}^d$ and the label set is $\mathcal{Y} = \{0, 1\}$. A literal over the variables $x_1, \ldots, x_d$ is a simple boolean function which takes the form $f(\mathbf{x}) = x_i$, for some $i \in [d]$, or $f(\mathbf{x}) = 1 - x_i$ for some $i \in [d]$. We use the notation $\bar{x}_i$ as a shorthand for $1 - x_i$. A conjunction is any product of literals. In boolean logic, the product is denoted using the $\wedge$ sign. For example, the function $h(\mathbf{x}) = x_1 \cdot (1 - x_2)$ is written as $x_1 \wedge \bar{x}_2$.

   We consider the hypothesis class of all conjunctions of literals over the d variables. The empty conjunction is interpreted as the all-positive hypothesis (namely, the function that returns $h(\mathbf{x}) = 1$ for all $\mathbf{x}$). The conjunction $x_1 \wedge \bar{x}_1$ (and similarly any conjunction involving a literal and its negation) is allowed and interpreted as the all-negative hypothesis (namely, the conjunction that returns $h(\mathbf{x}) = 0$ for all $\mathbf{x}$). We assume realizability, namely, we assume that there exists a boolean conjunction which generates the labels. Thus, each example $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ consists of an assignment to the d boolean variables $x_1, \ldots, x_d$, and its truth value (0 for false and 1 for true).

   For instance, let $d = 3$ and suppose that the true conjunction is $x_1 \wedge \bar{x}_2$. Then, the training set S might contain the following instances:

$$((1, 1, 1), 0), ((1, 0, 1), 1), ((0, 1, 0), 0)((1, 0, 0), 1) .$$

   Prove that the hypothesis class of all conjunctions over d variables is PAC learnable and bound its sample complexity. Propose an algorithm that implements the ERM rule, whose runtime is polynomial in $d \cdot m$.
5. Let $\mathcal{X}$ be a domain and let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ be a sequence of distributions over $\mathcal{X}$. Let $\mathcal{H}$ be a finite class of binary classifiers over $\mathcal{X}$ and let $f \in \mathcal{H}$. Suppose we are getting a sample S of m examples, such that the instances are independent but are not identically distributed; the i^{th} instance is sampled from $\mathcal{D}_i$ and then y_i is set to be $f(\mathbf{x}_i)$. Let $\bar{D}_m$ denote the average, that is, $\bar{\mathcal{D}}_m = (\mathcal{D}_1 + \ldots + \mathcal{D}_m)/m$.

   Fix an accuracy parameter $\epsilon \in (0, 1)$. Show that

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } L_{(\bar{\mathcal{D}}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right] \leq |\mathcal{H}| e^{-\epsilon m} .$$

   Hint: Use the geometric-arithmetic mean inequality.

6. Let $\mathcal{H}$ be a hypothesis class of binary classifiers. Show that if $\mathcal{H}$ is agnostic PAC learnable, then $\mathcal{H}$ is PAC learnable as well. Furthermore, if A is a successful agnostic PAC learner for $\mathcal{H}$, then A is also a successful PAC learner for $\mathcal{H}$.
7. **(*) The Bayes optimal predictor:** Show that for every probability distribution $\mathcal{D}$, the Bayes optimal predictor $f_{\mathcal{D}}$ is optimal, in the sense that for every classifier g from $\mathcal{X}$ to $\{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.
8. **(*)** We say that a learning algorithm A *is better than* B *with respect to* some probability distribution, $\mathcal{D}$, if

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(B(S))$$

for all samples $S \in (\mathcal{X} \times \{0,1\})^m$. We say that a learning algorithm A *is better than* B, if it is better than B with respect to all probability distributions $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$.

1. A probabilistic label predictor is a function that assigns to every domain point x a probability value, $h(x) \in [0,1]$, which determines the probability to predict the label 1. That is, given such an h and an input, x, the label for x is predicted by tossing a coin with bias $h(x)$ towards Heads and predicting 1 iff the coin comes up Heads. Formally, we define a probabilistic label predictor as a function, $h : \mathcal{X} \to [0,1]$. The loss of such h on an example (x,y) is defined to be $|h(x) - y|$, which is exactly the probability that the prediction of h will not equal to y. Note that if h is deterministic, i.e. returns values in $\{0,1\}$, then $|h(x) - y| = \mathbb{1}_{[h(x) \neq y]}$.
   Prove that for every data-generating distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, the Bayes optimal predictor has the smallest risk (w.r.t. the loss function $\ell(h,(x,y)) = |h(x) - y|$, among all possible label predictors, including probabilistic ones).
2. Let $\mathcal{X}$ be a domain and $\{0,1\}$ be a set of labels. Prove that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, there exists a learning algorithm $A_{\mathcal{D}}$ that is better than any other learning algorithm with respect to $\mathcal{D}$.
3. Prove that for every learning algorithm A there exists a probability distribution, $\mathcal{D}$, and a learning algorithm B such that A is not better than B w.r.t. $\mathcal{D}$.

9. Consider a variant of the PAC model in which there are two example oracles: one which generates positive examples, and one which generates negative examples, both according to the underlying distribution $\mathcal{D}$ on $\mathcal{X}$. Formally, given a target function $f : \mathcal{X} \to \{0,1\}$, let $\mathcal{D}^+$ be the distribution over $\mathcal{X}^+ = \{x \in \mathcal{X} : f(x) = 1\}$ defined by $\mathcal{D}^+(A) = \mathcal{D}(A)/\mathcal{D}(\mathcal{X}^+)$, for every $A \subset \mathcal{X}^+$. Similarly, $\mathcal{D}^-$ is the distribution over $\mathcal{X}^-$ induced by $\mathcal{D}$.

   The definition of PAC-learnability in the two-oracle model is the same as the standard definition of PAC learnability except that here the learner has access to $m^+_{\mathcal{H}}(\epsilon,\delta)$ i.i.d. examples from $\mathcal{D}^+$, and $m^-(\epsilon,\delta)$ i.i.d. examples from $\mathcal{D}^-$. The learner's goal is to output h s.t. with probability at least

$1 - \delta$ (over the choice of the two training sets, and possibly over the nondeterministic decisions made by the learning algorithm), both $L_{(D^+,f)}(h) \leq \epsilon$ and $L_{(D^-,f)}(h) \leq \epsilon$.

1. (*) Show that if $\mathcal{H}$ is PAC learnable (in the standard one-oracle model), then $\mathcal{H}$ is PAC learnable in the two-oracle model.
2. (**) Define h^+ to be the always-plus hypothesis and h^- to be the always-minus hypothesis. Assume that $h^+, h^- \in \mathcal{H}$. Show that if $\mathcal{H}$ is PAC learnable in the two-oracle model, then $\mathcal{H}$ is PAC learnable in the standard one-oracle model.

4 Learning via Uniform Convergence

The first formal learning model that we have discussed was the PAC model. In Chapter 2 we have shown that under the realizability assumption, any finite hypothesis class is PAC learnable. In this chapter we will develop a general tool, *uniform convergence*, and apply it to show that any finite class is learnable in the agnostic PAC model with general loss functions, as long as the range loss function is bounded.

4.1 Uniform Convergence is Sufficient for Learnability

The idea behind the learning condition discussed in this chapter is very simple. Recall that, given a hypothesis class, $\mathcal{H}$, the ERM learning paradigm works as follows: upon receiving a training sample, S, the learner evaluates the risk (or error) of each h in $\mathcal{H}$ on the given sample, and outputs a member of $\mathcal{H}$ that minimizes this empirical risk. The hope is that an h that minimizes the empirical risk with respect to S, is a risk minimizer (or has risk close to the minimum) with respect to the true data probability distribution as well. For that, it suffices to ensure that the empirical risks of all members of $\mathcal{H}$ are good approximations of their true risk. Put another way, we need that uniformly over all hypotheses in the hypothesis class, the empirical risk will be close to the true risk, as formalized below.

DEFINITION 4.1 (ϵ-representative sample) A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \ .$$

The next simple lemma states that whenever the sample is $(\epsilon/2)$-representative, the ERM learning rule is guaranteed to return a good hypothesis.

LEMMA 4.2 *Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then, any output of* $\mathrm{ERM}_{\mathcal{H}}(S)$, *namely any $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$, satisfies*

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

Proof For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon \ ,$$

where the first and third inequalities are due to the assumption that S is $\frac{\epsilon}{2}$-representative (Definition 4.1) and the second inequality holds since h_S is an ERM predictor. □

The above lemma implies that to ensure that the ERM rule is an agnostic PAC learner, it suffices to show that with probability of at least $1 - \delta$ over the random choice of a training set, it will be an ϵ-representative training set. The uniform convergence condition formalizes this requirement.

DEFINITION 4.3 (uniform convergence) We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\text{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ϵ-representative.

Similar to the definition of sample complexity for PAC learning, the function $m_{\mathcal{H}}^{\text{UC}}$ measures the (minimal) sample complexity of obtaining the uniform convergence property, namely, how many examples we need to ensure that with probability of at least $1 - \delta$ the sample would be ϵ-representative.
The term "uniform" here refers to having a fixed sample size that works for all members of $\mathcal{H}$ and over all possible probability distributions over the domain.

The following corollary follows directly from Lemma 4.2 and the definition of uniform convergence.

COROLLARY 4.4 *If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the* $\text{ERM}_{\mathcal{H}}$ *paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

4.2 Finite Classes are Agnostic PAC Learnable

In view of Corollary 4.4, the claim that every finite hypothesis class is agnostic PAC learnable will follow once we establish that uniform convergence holds for a finite hypothesis class.

To show that uniform convergence holds we follow a two step argument, similar to the derivation in Chapter 2. The first step applies the union bound while the second step employs a measure concentration inequality. We now explain these two steps in detail.

Fix some ϵ, δ. We need to find a sample size m which guarantees that for any $\mathcal{D}$, with probability of at least $1 - \delta$ of the choice of $S = (z_1, \ldots, z_m)$ sampled

i.i.d. from $\mathcal{D}$ we have that for all $h \in \mathcal{H}$, $|L_S(h) - L_\mathcal{D}(h)| \leq \epsilon$. That is,

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| \leq \epsilon\}) \geq 1 - \delta \ .$$

Equivalently, we need to show that

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) < \delta \ .$$

Writing

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| > \epsilon\} = \cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_\mathcal{D}(h)| > \epsilon\} \ ,$$

and applying the union bound (Lemma 2.2) we obtain

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) \ . \tag{4.1}$$

Our second step will be to argue that each summand of the right-hand side of the above is small enough (for a sufficiently large m). That is, we will show that for any fixed hypothesis, h, (which is chosen in advance prior to the sampling of the training set), the gap between the true and empirical risks, $|L_S(h) - L_\mathcal{D}(h)|$, is likely to be small.

Recall that $L_\mathcal{D}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ and that $L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i)$. Since each z_i is sampled i.i.d. from $\mathcal{D}$, the expected value of the random variable $\ell(h, z_i)$ is $L_\mathcal{D}(h)$. By the linearity of expectation, it follows that $L_\mathcal{D}(h)$ is also the expected value of $L_S(h)$. Hence, the quantity $|L_\mathcal{D}(h) - L_S(h)|$ is the deviation of the random variable $L_S(h)$ from its expectation. We therefore need to show that the measure of $L_S(h)$ is *concentrated* around its expected value.

A basic statistical fact, the *law of large numbers*, states that when m goes to infinity, empirical averages converge to their true expectation. This is true for $L_S(h)$, since it is the empirical average of m i.i.d random variabels. However, since the law of large numbers is only an asymptotic result, it provides no information about the gap between the empirically estimated error and its true value for any given, finite, sample size.

Instead, we will use a measure concentration inequality due to Hoeffding, which quantifies the gap between empirical averages and their expected value.

LEMMA 4.5 (Hoeffding's inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu\right| > \epsilon\right] \ \leq \ 2\exp\left(-2m\epsilon^2/(b-a)^2\right) \ .$$

The proof can be found in Appendix B.

Getting back to our problem, let θ_i be the random variable $\ell(h, z_i)$. Since h is fixed and $z_1, \ldots, z_m$ are sampled i.i.d., it follows that $\theta_1, \ldots, \theta_m$ are also i.i.d. random variables. Furthermore, $L_S(h) = \frac{1}{m}\sum_{i=1}^m \theta_i$ and $L_\mathcal{D}(h) = \mu$. Let us

further assume that the range of ℓ is $[0,1]$ and therefore $\theta_i \in [0,1]$. We therefore obtain that

$$\mathcal{D}^m(\{S : |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right) . \tag{4.2}$$

Combining this with Equation (4.1) yields

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) &\leq \sum_{h\in\mathcal{H}} 2\exp\left(-2m\epsilon^2\right) \\ &= 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right) .\end{aligned}$$

Finally, if we choose

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

then

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) \leq \delta .$$

COROLLARY 4.6 *Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity*

$$m_\mathcal{H}^{UC}(\epsilon,\delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_\mathcal{H}(\epsilon,\delta) \leq m_\mathcal{H}^{UC}(\epsilon/2,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

Remark 4.1 (The "discretization trick") While the above corollary only applies to finite hypothesis classes, there is a simple trick which allows us to get a very good estimate of the practical sample complexity of infinite hypothesis classes. Consider a hypothesis class which is parametrized by d parameters. For example, let $\mathcal{X} = \mathbf{R}$, $\mathcal{Y} = \{\pm 1\}$, and the hypothesis class, $\mathcal{H}$, being all functions of the form $h_\theta(x) = \text{sign}(x-\theta)$. That is, each hypothesis is parameterized by one parameter, $\theta \in \mathbb{R}$, and the hypothesis outputs 1 for all instances larger than θ and outputs -1 for instances smaller than θ. This is a hypothesis class of an infinite size. However, if we are going to learn this hypothesis class in practice, using a computer, we would probably maintain real numbers using floating point representation, say of 64 bits. It follows that in practice, our hypothesis class is parametrized by the set of scalars that can be represented using a 64 bits floating point number. There are at most 2^{64} such numbers, hence the actual size of our hypothesis class is at most 2^{64}. More generally, if our hypothesis class is parameterized by d numbers, in practice we learn a hypothesis class of size at most 2^{64d}. Applying Corollary 4.6 we obtain that the sample complexity of such classes is bounded

by $\frac{128d+2\log(2/\delta)}{\epsilon^2}$. This upper bound on the sample complexity has the deficiency of being dependent on the specific representation of real numbers used by our machine. In Chapter 6 we will introduce a rigorous way to analyze the sample complexity of infinite size hypothesis classes. Nevertheless, the discretization trick can be used to get a rough estimate of the sample complexity in many practical situations.

Summary

If the uniform convergence property holds for a hypothesis class $\mathcal{H}$ then in most cases the empirical risks of hypotheses in $\mathcal{H}$ will faithfully represent their true risks. Uniform convergence suffices for agnostic PAC learnability using the ERM rule. We have shown that finite hypothesis classes enjoy the uniform convergence property and are hence agnostic PAC learnable.

4.3 Bibliographic Remarks

Classes of functions for which the uniform convergence property holds are also called Glivenko-Cantelli classes, named after Valery Ivanovich Glivenko and Francesco Paolo Cantelli, who proved the first uniform convergence result in the 1930s. See (Dudley, Gine & Zinn 1991). The relation between uniform convergence and learnability was thoroughly studied by Vapnik — see (Vapnik 1992, Vapnik 1995, Vapnik 1998). In fact, as we will see later in Chapter 6, the fundamental theorem of learning theory states that in binary classification problems, uniform convergence is not only a sufficient condition for learnability but it is also a necessary condition. This is not the case for more general learning problems (see (Shalev-Shwartz, Shamir, Srebro & Sridharan 2010)).

4.4 Exercises

1. In this exercise, we show that the (ϵ, δ) requirement on the convergence of errors in our definitions of PAC learning, is, in fact, quite close to a simpler looking requirement about averages (or expectations). Prove that the following two statements are equivalent (for any learning algorithm A, any probability distribution $\mathcal{D}$, and any loss function whose range is $[0,1]$):
   1. For every $\epsilon, \delta > 0$, there exists $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$
   $$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] < \delta$$
   2.
   $$\lim_{m\to\infty} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$$

(where $\mathbb{E}_{S\sim\mathcal{D}^m}$ denotes the expectation over samples S of size m).

2. **Bounded loss functions:** In Corollary 4.6 we assumed that the range of the loss function is $[0,1]$. Prove that if the range of the loss function is $[a,b]$ then the sample complexity satisfies:

$$m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil .$$

5 The Bias-Complexity Tradeoff

In Chapter 2 we saw that without being careful, the training data can mislead the learner, and result in overfitting. To overcome this problem, we restricted the search space to some hypothesis class $\mathcal{H}$. Such a hypothesis class can be viewed as reflecting some prior knowledge that the learner has about the task - a belief that one of the members of the class $\mathcal{H}$ is a low-error model for the task. For example, in our papayas taste problem, based on our previous experience with other fruits, we may assume that some rectangle in the color-hardness plane predicts (at least approximately) the papaya's tastiness.

Is such prior knowledge really necessary for the success of learning? Maybe there exists some kind of universal learner? That is, a learner which has no prior knowledge about a certain task, and is ready to be challenged by any task. Let us elaborate on this point. A specific learning task is defined by an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where the goal of the learner is to find a predictor $h : \mathcal{X} \to \mathcal{Y}$, whose risk, $L_{\mathcal{D}}(h)$, is small enough. The question is therefore whether there exists a learning algorithm A and a training set size m, such that for every distribution $\mathcal{D}$, if A receives m i.i.d. examples from $\mathcal{D}$, there's a high chance it outputs a predictor h which has a low risk.

The first part of this chapter addresses this question formally. The No-Free-Lunch theorem states that no such universal learner exists. To be more precise, the theorem states that for binary classification prediction tasks, for every learner there exists a distribution on which it fails. We say that the learner fails if, upon receiving i.i.d. examples from that distribution, its output hypothesis is likely to have a large risk, say, ≥ 0.3, whereas for the same distribution, there exists another learner that will output a hypothesis with a small risk. In other words, the theorem states that no learner can succeed on all learnable tasks - every learner has tasks on which it fails while other learners succeed.

Therefore, when approaching a particular learning problem, defined by some distribution $\mathcal{D}$, we should have some prior knowledge on $\mathcal{D}$. One type of such prior knowledge is that $\mathcal{D}$ comes from some specific parametric family of distributions. We will study learning under such assumptions later on in Chapter 24. Another type of a prior knowledge on $\mathcal{D}$, which we assumed when defining the PAC learning model, is that there exists h in some predefined hypothesis class $\mathcal{H}$, such that $L_{\mathcal{D}}(h) = 0$. A softer type of prior knowledge on $\mathcal{D}$ is assuming that $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ is small. In a sense, this weaker assumption on $\mathcal{D}$ is a prerequisite

for using the agnostic PAC model, in which we require that the risk of the output hypothesis will not be much larger than $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

In the second part of this chapter we study the benefits and pitfalls of using a hypothesis class as means of formalizing prior knowledge. We decompose the error of an ERM algorithm over a class $\mathcal{H}$ into two components. The first component reflects the quality of our prior knowledge, measured by the minimal risk of a hypothesis in our hypothesis class, $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$. This component is also called the *approximation error*, or the *bias* of the algorithm toward choosing a hypothesis from $\mathcal{H}$. The second component is the error due to overfitting, which depends on the size or the complexity of the class $\mathcal{H}$, and is called the *estimation error*. These two terms imply a tradeoff between choosing a more complex $\mathcal{H}$ (which can decrease the bias but increases the risk of overfitting) or a less complex $\mathcal{H}$ (which might increase the bias but decreases the potential overfitting).

5.1 The No-Free-Lunch theorem

In this part we prove that there is no universal learner. We do this by showing that no learner can succeed on all learning tasks, as formalized in the following theorem:

THEOREM 5.1 (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

This theorem states that for every learner, there exists a task on which it fails, even though that task can be successfully learnt by another learner. Indeed, a trivial successful learner in this case would be an ERM learner with the hypothesis class $\mathcal{H} = \{f\}$, or more generally, ERM with respect to any finite hypothesis class that contains f and whose size satisfies the equation $m \geq 8 \log(7|\mathcal{H}|/6)$ (see Corollary 2.3).

Proof Let C be a subset of $\mathcal{X}$ of size $2m$. The intuition of the proof is that any learning algorithm which observes only half of the instances in C has no information on what should be the labels of the rest of the instances in C. Therefore, there exists a "reality", that is some target function f, which would contradict the labels that $A(S)$ predicts on the unobserved instances in C.

Note that there are $T = 2^{2m}$ possible functions from C to $\{0,1\}$. Denote these functions by $f_1, \ldots, f_T$. For each such function, let $\mathcal{D}_i$ be a distribution over

$C \times \{0,1\}$ defined by

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} 1/|C| & \text{if } y = f_i(x) \\ 0 & \text{otherwise} \end{cases} .$$

That is, the probability to choose a pair (x,y) is $1/|C|$ if the label y is indeed the true label according to f_i, and the probability is 0 if $y \neq f_i(x)$. Clearly, $L_{\mathcal{D}_i}(f_i) = 0$.

We will show that for every algorithm, A, that receives a training set of m examples from $C \times \{0,1\}$ and returns a function $A(S) : C \to \{0,1\}$, it holds that

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim \mathcal{D}_i^m} [L_{\mathcal{D}_i}(A(S))] \geq 1/4 . \tag{5.1}$$

Clearly, this means that for every algorithm A', that receives a training set of m examples from $\mathcal{X} \times \{0,1\}$, there exists a function $f : \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, such that $L_{\mathcal{D}}(f) = 0$ and

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} [L_{\mathcal{D}}(A'(S))] \geq 1/4 . \tag{5.2}$$

It is easy to verify that the above suffices for showing that $\mathbb{P}[L_{\mathcal{D}}(A'(S)) \geq 1/8] \geq 1/7$, which is what we need to prove (see Exercise 1).

We now turn to prove that Equation (5.1) holds. There are $k = (2m)^m$ possible sequences of m examples from C. Denote these sequences by $S_1, \ldots, S_k$. Also, if $S_j = (x_1, \ldots, x_m)$ we denote by S_j^i the sequence containing the instances in S_j labeled by the function f_i, namely, $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$. If the distribution is $\mathcal{D}_i$ then the possible training sets A can receive are $S_1^i, \ldots, S_k^i$, and all these training sets have the same probability to be sampled. Therefore,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}_i^m} [L_{\mathcal{D}_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) . \tag{5.3}$$

Using the facts that "maximum" is larger than "average" and that "average" is larger than "minimum", we have

$$\begin{aligned} \max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \\ &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \\ &\geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) . \end{aligned} \tag{5.4}$$

Next, fix some $j \in [k]$. Denote $S_j = (x_1, \ldots, x_m)$ and let $v_1, \ldots, v_p$ be the examples in C that do not appear in S_j. Clearly, $p \geq m$. Therefore, for every

function $h : C \to \{0,1\}$ and every i we have

$$\begin{aligned} L_{\mathcal{D}_i}(h) &= \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[h(x) \neq f_i(x)]} \\ &\geq \frac{1}{2m} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \\ &\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \ . \end{aligned} \tag{5.5}$$

Hence,

$$\begin{aligned} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} \\ &= \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} \\ &\geq \frac{1}{2} \cdot \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} \ . \end{aligned} \tag{5.6}$$

Next, fix some $r \in [p]$. We can partition all the functions in $f_1, \ldots, f_T$ into $T/2$ disjoint pairs, where for a pair $(f_i, f_{i'})$ we have that for every $c \in C$, $f_i(c) \neq f_{i'}(c)$ if and only if $c = v_r$. Since for such a pair we must have $S_j^i = S_j^{i'}$, it follows that

$$\mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} + \mathbb{1}_{[A(S_j^{i'})(v_r) \neq f_{i'}(v_r)]} = 1 \ ,$$

which yields

$$\frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2} \ .$$

Combining this with Equation (5.6), Equation (5.4), and Equation (5.3), we obtain that Equation (5.1) holds, which concludes our proof. □

5.1.1 No-Free-Lunch and prior knowledge

How does the No-Free-Lunch result relate to the need of prior knowledge? Let us consider an ERM predictor over the hypothesis class $\mathcal{H}$ of all the functions f from X to $\{0,1\}$. This class represents lack of prior knowledge: every possible function from the domain to the label set is considered a good candidate. According to the No-Free-Lunch theorem, any algorithm which chooses its output from hypotheses in $\mathcal{H}$, and in particular the ERM predictor, will fail on some learning task. Therefore, this class is not PAC-learnable, as formalized in the following corollary:

COROLLARY 5.2 *Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathcal{H}$ is not PAC learnable.*

Proof Assume, by way of contradiction, that the class is learnable. Choose some $\epsilon < 1/8$ and $\delta < 1/7$. By the definition of PAC learnability, there must be some learning algorithm A and an integer $m = m(\epsilon, \delta)$, such that for any data-generating distribution over $\mathcal{X} \times \{0,1\}$, if for some function $f : \mathcal{X} \to \{0,1\}$, $L_{\mathcal{D}}(f) = 0$, then with probability greater than $1-\delta$ when A is applied to samples S of size m, generated i.i.d. by $\mathcal{D}$, $L_{\mathcal{D}}(A(S)) \leq \epsilon$. However, applying the above No-Free-Lunch theorem, since $|\mathcal{X}| > 2m$, for every learning algorithm (and in particular for the algorithm A), there exists a distribution $\mathcal{D}$ such that with probability greater than $1/7 > \delta$, $L_{\mathcal{D}}(A(S)) > 1/8 > \epsilon$, which leads to the desired contradiction. □

How can we prevent our algorithms from such failures? We can escape the hazards foreseen by the No-Free-Lunch theorem by using our prior knowledge about a specific learning task, to avoid the distributions which will cause us to fail when learning that task. Such prior knowledge can be expressed by restricting our hypothesis class.

But how should we choose a good hypothesis class? On the one hand, we want to believe that this class includes the hypothesis which has no error at all (in the PAC setting), or at least that the smallest error achievable by a hypothesis from this class is indeed rather small (in the agnostic setting). On the other hand, we have just seen that we cannot simply choose the richest class – the class of all functions over the given domain. This tradeoff is discussed in the following section.

5.2 Error decomposition

To answer this question we decompose the error of an $\text{ERM}_{\mathcal{H}}$ predictor into two components as follows. Let h_S be an $\text{ERM}_{\mathcal{H}}$ hypothesis. Then, we can write

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \quad \text{where :} \quad \epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h), \; \epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}} \; . \tag{5.7}$$

- **The approximation error**—the minimum risk achievable by a predictor in the hypothesis class. This term measures how much risk we have because we restrict ourselves to a specific class, namely, how much *inductive bias* we have. The approximation error does not depend on the sample size, and is determined by the hypothesis class chosen. Enlarging the hypothesis class can decrease the approximation error.

  Under the realizability assumption, the approximation error is zero. In the agnostic case, however, the approximation error can be large.[1]

[1] In fact, it always includes the error of the Bayes optimal predictor (see Chapter 3), the minimal yet inevitable error, due to the possible non-determinism of the world in this model. Sometimes in the literature the term *approximation error* refers not to

- **The estimation error**—the difference between the approximation error and the error achieved by the ERM predictor. The estimation error is a result of the empirical risk (i.e. training error) being only an estimate of the true risk, and so the predictor minimizing the empirical risk is only an estimate of the predictor minimizing the true risk.

  The quality of this estimation depends on the training set size and on the size, or complexity, of the hypothesis class. As we have shown, for a finite hypothesis class, ϵ_{est} increases (logarithmically) with $|\mathcal{H}|$ and decreases with m. We can think of the size of $\mathcal{H}$ as a measure of its complexity. In future chapters we will define other complexity measures of hypothesis classes.

Since our goal is to minimize the total risk, we face a tradeoff, called the *bias-complexity tradeoff*. On one hand, choosing $\mathcal{H}$ to be a very rich class decreases the approximation error but at the same time might increase the estimation error, as a rich $\mathcal{H}$ might lead to *overfitting*. On the other hand, choosing $\mathcal{H}$ to be a very small set reduces the estimation error but might increase the approximation error, or in other words, might lead to *underfitting*. Of course, a great choice for $\mathcal{H}$ is the class that contains only one classifier – the Bayes optimal classifier. But, the Bayes optimal classifier depends on the underlying distribution $\mathcal{D}$, which we do not know (indeed, learning would have been unnecessary had we known $\mathcal{D}$).

Learning theory studies how rich we can make $\mathcal{H}$ while still maintaining reasonable estimation error. In many cases, empirical research focuses on designing good hypothesis classes for a certain domain. Here, "good" means classes for which the approximation error would not be excessively high. The idea is that although we are not experts and do not know how to construct the optimal classifier, we still have some prior knowledge on the specific problem at hand which enables us to design hypothesis classes for which both the approximation error and the estimation error are not too large. Getting back to our Papayas example, we do not know how exactly the color and hardness of a Papaya predict its taste, but we do know that Papaya is a fruit and based on previous experience with other fruit we conjecture that a rectangle in the color-hardness space may be a good predictor.

Summary

The No-Free-Lunch theorem states that there is no universal learner. Every learner has to be specified to some task, and use some prior knowledge about that task, in order to succeed. So far we have modeled our prior knowledge by restricting our output hypothesis to be a member of a chosen hypothesis class. When choosing this hypothesis class, we face a tradeoff, between a larger, or

$\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, but rather to the excess error over that of the Bayes optimal predictor, namely $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - \epsilon_{\text{Bayes}}$.

more complex, class which is more likely to have a small approximation error, and a more restricted class which would guarantee that the estimation error will be small. In the next chapter we will study in more detail the behavior of the estimation error. In Chapter 7 we will discuss alternative ways to express prior knowledge.

5.3 Bibliographic remarks

(Wolpert & Macready 1997) proved several no free lunch theorems for optimization, but these are rather different than the theorem we prove here. The theorem we prove here is closely related to lower bounds in VC theory, as we will study in the next chapter.

5.4 Exercise

1. Prove that Equation (5.2) suffices for showing that $\mathbb{P}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$. *Hint:* Let θ be a random variable that receives values in $[0,1]$ and whose expectation satisfies $\mathbb{E}[\theta] \geq 1/4$. Use Lemma B.1 to show that $\mathbb{P}[\theta \geq 1/8] \geq 1/7$.
2. Assume you are asked to design a learning algorithm to predict if patients are going to suffer a heart attack. Relevant patient features the algorithm may have access to include: Blood Pressure (BP), body-mass index (BMI), age (A), level of physical activity (P), and income (I).

   You have to choose between two algorithms; the first picks an axis-alined rectangle in the two dimensional space spanned by the features BP and BMI and the other picks an axis-alined rectangle in the 5 dimensional space spanned by all the above features.
   1. Explain the pros and cons of each of choice.
   2. Explain how will the number of available labeled training samples effect your choice.
3. Prove that if $|\mathcal{X}| \geq km$ for a positive integer $k \geq 2$, then we can replace the lower bound of $1/4$ in the No-Free-Lunch theorem with $\frac{k-1}{2k} = \frac{1}{2} - \frac{1}{2k}$. Namely, let A be a learning algorithm for the task of binary classification. Let m be any number smaller than $|\mathcal{X}|/k$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:
   - There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.
   - $\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{2} - \frac{1}{2k}$.

6 The VC-dimension

In the previous chapter, we decomposed the error of the $\text{ERM}_{\mathcal{H}}$ rule into approximation error and estimation error. The approximation error depends on the fit of our prior knowledge (as reflected by the choice of the hypothesis class $\mathcal{H}$) to the underlying unknown distribution. In contrast, the definition of PAC learnability requires that the estimation error would be bounded uniformly over all distributions.

Our current goal is to figure out which classes $\mathcal{H}$ are PAC learnable, and to exactly characterize the sample complexity of learning a given hypothesis class. So far we have seen that finite classes are learnable, but that the class of all functions (over an infinite size domain) is not. What makes one class learnable and the other unlearnable? Can infinite-size classes be learnable, and if so, what determines their sample complexity?

We begin the chapter by showing that infinite classes can indeed be learnable, and thus, finiteness of the hypothesis class is not a necessary condition for learnability. We then present a remarkably crisp characterization of the family of learnable classes in the setup of binary valued classification with the zero-one loss. This characterization was first discovered by Vladimir Vapnik and Alexey Chervonenkis in 1970 and relies on a combinatorial notion called the Vapnik-Chervonenkis dimension (VC-dimension). We formally define the VC-dimension, provide several examples, and then state the fundamental theorem of statistical learning theory, which integrates the concepts of learnability, VC-dimension, the ERM rule and uniform convergence.

6.1 Infinite-size Classes can be Learnable

In Chapter 4 we saw that finite classes are learnable, and in fact the sample complexity of a hypothesis class is upper bounded by the log of its size. To show that the size of the hypothesis class is not the right characterization of its sample complexity, we first present a simple example of an infinite-size hypothesis class which is learnable.

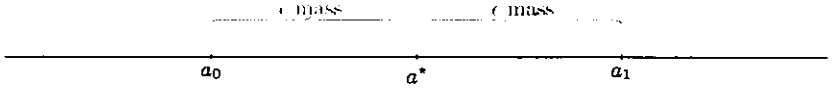
Example 6.1 Let $\mathcal{H}$ be the set of threshold functions over the real line, namely, $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0,1\}$ is a function such that $h_a(x) = \mathbb{1}_{[x<a]}$. To remind the reader, $\mathbb{1}_{[x<a]}$ is 1 if $x < a$ and 0 otherwise. Clearly, $\mathcal{H}$ is of infinite

size. Nevertheless, the following Lemma shows that $\mathcal{H}$ is learnable in the PAC model using the ERM algorithm.

Lemma 6.1 *Let $\mathcal{H}$ be the class of thresholds as defined above. Then, $\mathcal{H}$ is PAC learnable, using the ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta)/\epsilon \rceil$.*

Proof Let a^* be a threshold such that the hypothesis $h^*(x) = \mathbb{1}_{[x<a^*]}$ achieves $L_{\mathcal{D}}(h^*) = 0$. Let $\mathcal{D}_x$ be the marginal distribution over the domain $\mathcal{X}$ and let $a_0 < a^* < a_1$ be such that

$$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^*, a_1)] = \epsilon \ .$$



(If $\mathcal{D}_x(-\infty, a^*) \leq \epsilon$ we set $a_0 = -\infty$ and similarly for a_1). Given a training set S, let $b_0 = \max\{x : (x, 1) \in S\}$ and $b_1 = \min\{x : (x, 0) \in S\}$ (if no example in S is positive we set $b_0 = -\infty$ and if no example in S is negative we set $b_1 = \infty$). Let b_S be a threshold corresponding to an ERM hypothesis, h_S, which implies that $b_S \in (b_0, b_1)$. Therefore, a sufficient condition for $L_{\mathcal{D}}(h_S) \leq \epsilon$ is that both $b_0 \geq a_0$ and $b_1 \leq a_1$. In other words,

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0 \vee b_1 > a_1],$$

and using the union bound we can bound the above by

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \ . \tag{6.1}$$

The event $b_0 < a_0$ happens if and only if all examples in S are not in the interval (a_0, a^*), whose probability mass is defined to he ϵ. Namely,

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall (x, y) \in S, \ x \notin (a_0, a^*)] = (1-\epsilon)^m \leq e^{-\epsilon m} \ .$$

Since we assume $m > \log(2/\delta)/\epsilon$ it follows that the above is at most $\delta/2$. In the same way it is easy to see that $\mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2$. Combining with Equation (6.1) we conclude our proof. □

6.2 The VC-Dimension

We see, therefore, that while finiteness of $\mathcal{H}$ is a sufficient condition for learnability, it is not a necessary condition. As we will show, a property called the VC dimension of a hypothesis class gives the correct characterization of its learnability. To motivate the definition of the VC dimension, let us recall the No-Free-Lunch theorem (Theorem 5.1) and its proof. There, we have shown that without

restricting the hypothesis class, for any learning algorithm, an adversary can construct a distribution for which the learning algorithm will perform poorly, while there is another learning algorithm that will succeed on the same distribution. To do so, the adversary used a finite set $C \subset \mathcal{X}$ and considered a family of distributions that are concentrated on elements of C. Each distribution was derived from a "true" target function from C to $\{0,1\}$. To make any algorithm fail, the adversary used the power of choosing a target function from the set of *all* possible functions from C to $\{0,1\}$.

When considering PAC learnability of a hypothesis class $\mathcal{H}$, the adversary is restricted to construct distributions for which some hypothesis $h \in \mathcal{H}$ achieves a zero risk. Since we are considering distributions that are concentrated on elements of C, we should study how $\mathcal{H}$ behaves on C, which leads to the following definition.

DEFINITION 6.2 (Restriction of $\mathcal{H}$ to C) Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to C is the set of functions from C to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,

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

where we represent each function from C to $\{0,1\}$ as a vector in $\{0,1\}^{|C|}$.

If the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$, then we say that $\mathcal{H}$ *shatters* the set C. Formally:

DEFINITION 6.3 (Shattering) A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$.

Example 6.2 Let $\mathcal{H}$ be the class of threshold functions over $\mathbf{R}$. Take a set $C = \{c_1\}$. Now, if we take $a = c_1 + 1$, then we have $h_a(c_1) = 1$, and if we take $a = c_1 - 1$, then we have $h_a(c_1) = 0$. Therefore, $\mathcal{H}_C$ is the set of all functions from C to $\{0,1\}$, and $\mathcal{H}$ shatters C. Now take a set $C = \{c_1, c_2\}$, where $c_1 \leq c_2$. No $h \in \mathcal{H}$ can account for the labeling $(0,1)$, because any threshold that assigns the label 0 to c_1 must assign the label 0 to c_2 as well. Therefore not all functions from C to $\{0,1\}$ are included in $\mathcal{H}_C$, hence C is not shattered by $\mathcal{H}$.

Getting back to the construction of an adversarial distribution as in the proof of the No-Free-Lunch theorem (Theorem 5.1), we see that whenever some set C is shattered by $\mathcal{H}$, the adversary is not restricted by $\mathcal{H}$, as they can construct a distribution over C based on *any* target function from C to $\{0,1\}$, while still maintaining the realizability assumption. This immediately yields:

COROLLARY 6.4 *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$. Let m be a training set size. Assume that there exists a set $C \subset \mathcal{X}$ of size $2m$ which is shattered by $\mathcal{H}$. Then, for any learning algorithm, A, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $L_{\mathcal{D}}(h) = 0$ but with probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Corollary 6.4 tells us that if $\mathcal{H}$ shatters some set C of size $2m$ then we cannot learn $\mathcal{H}$ using m examples. Intuitively, if a set C is shattered by $\mathcal{H}$, and we receive a sample containing half the instances of C, the labels of these instances give us no information about the labels of the rest of the instances in C—every possible labeling of the rest of the instances can be explained by some hypothesis in $\mathcal{H}$. Philosophically,

If someone can explain every phenomena, his explanations are worthless.

This leads us directly to the definition of the VC dimension.

DEFINITION 6.5 (VC dimension) The VC dimension of a hypothesis class $\mathcal{H}$, denoted $\text{VCdim}(\mathcal{H})$, 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 arbitrarily large size we say that $\mathcal{H}$ has infinite VC dimension.

A direct consequence of Corollary 6.4 is therefore:

THEOREM 6.6 *Let $\mathcal{H}$ be a class of infinite VC dimension. Then, $\mathcal{H}$ is not PAC learnable.*

Proof Since $\mathcal{H}$ has an infinite VC dimension, for any training set size m, there exists a shattered set of size $2m$, and the claim follows by Corollary 6.4. □

We shall see later in this chapter that the converse is also true: a finite VC-dimension guarantees learnability. Hence, the VC dimension characterizes PAC learnability. But before dwelling into more theory, we first show several examples.

6.3 Examples

In this section we calculate the VC dimension of several hypothesis classes. To show that $\text{VCdim}(\mathcal{H}) = d$ we need to show that:

1. There exists a set C of size d which is shattered by $\mathcal{H}$.
2. Every set C of size $d + 1$ is not shattered by $\mathcal{H}$.

6.3.1 Threshold functions

Let $\mathcal{H}$ be the class of threshold functions over $\mathbb{R}$. Recall Example 6.2, where we have shown that for an arbitrary set $C = \{c_1\}$, $\mathcal{H}$ shatters C, therefore $\text{VCdim}(\mathcal{H}) \geq 1$. We have also shown that for an arbitrary set $C = \{c_1, c_2\}$ where $c_1 \leq c_2$, $\mathcal{H}$ does not shatter C. We therefore conclude that $\text{VCdim}(\mathcal{H}) = 1$.

6.3.2 Intervals

Let $\mathcal{H}$ be the class of intervals over $\mathbb{R}$, namely, $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$, where $h_{a,b} : \mathbb{R} \to \{0,1\}$ is a function such that $h_{a,b}(x) = \mathbb{1}_{[x \in (a,b)]}$. Take the set $C = \{1,2\}$. Then, $\mathcal{H}$ shatters C (make sure you understand why) and therefore $\text{VCdim}(\mathcal{H}) \geq 2$. Now take an arbitrary set $C = \{c_1, c_2, c_3\}$ and assume without loss of generality that $c_1 \leq c_2 \leq c_3$. Then, the labeling $(1,0,1)$ cannot be obtained by an interval and therefore $\mathcal{H}$ does not shatter C. We therefore conclude that $\text{VCdim}(\mathcal{H}) = 2$.

6.3.3 Axis aligned rectangles

Let $\mathcal{H}$ be the class of axis aligned rectangles, formally:

$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\}$$

where

$$h_{(a_1,a_2,b_1,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

We shall show below that $\text{VCdim}(\mathcal{H}) = 4$. To prove this we need to find a set of 4 points which are shattered by $\mathcal{H}$, and also show that no set of 5 points can be shattered by $\mathcal{H}$. Finding a set of 4 points that are shattered is easy (see Figure 6.1). Now, consider any set $C \subset \mathbb{R}^2$ of 5 points. In C, take a leftmost point (whose first coordinate is the smallest in C), a rightmost point (first coordinate is the largest), a lowest point (second coordinate is the smallest), and a highest point (second coordinate is the largest). Without loss of generality, denote $C = \{c_1, \ldots, c_5\}$ and let c_5 be the point that was not selected. Now, define the labeling $(1,1,1,1,0)$. It is impossible to obtain this labeling by an axis-aligned rectangle. Indeed, such a rectangle must contain $c_1, \ldots, c_4$; but in this case the rectangle contains c_5 as well, because its coordinates are within the intervals defined by the selected points. So, C is not shattered by H, and therefore $\text{VCdim}(H) = 4$.

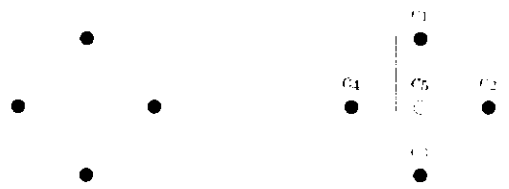


Figure 6.1 Left: 4 points which are shattered by axis-aligned rectangles. Right: Any axis-aligned rectangle cannot label c_5 by 0 and the rest of the points by 1.

6.3.4 Finite classes

Let $\mathcal{H}$ be a finite class. Then, clearly, for any set C we have $|\mathcal{H}_C| \leq |\mathcal{H}|$ and thus C cannot be shattered if $|\mathcal{H}| < 2^{|C|}$. This implies that $\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$. This shows that the PAC learnability of finite classes follows from the more general statement of PAC learnability of classes with finite VC dimension, that we shall see in the next section. Note, however, that the VC dimension of a finite class $\mathcal{H}$ can be significantly smaller than $\log_2(|\mathcal{H}|)$. For example, let $\mathcal{X} = \{1, \ldots, k\}$, for some integer k, and consider the class of threshold functions (as defined in Example 6.2). Then, $|\mathcal{H}| = k$ but $\text{VCdim}(\mathcal{H}) = 1$. Since k can be arbitrarily large, the gap between $\log_2(|\mathcal{H}|)$ and $\text{VCdim}(\mathcal{H})$ can be arbitrarily large.

6.3.5 VC-dimension and the number of parameters

In the previous examples, the VC dimension happened to equal the number of parameters defining the hypothesis class. While this is often the case, it is not always true. Consider for example the domain $\mathcal{X} = \mathbb{R}$, and the hypothesis class $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$ where $h_\theta : \mathcal{X} \to \{0, 1\}$ is defined by $h_\theta(x) = \lceil 0.5 \sin(\theta x) \rceil$. It is possible to prove that $\text{VCdim}(\mathcal{H}) = \infty$, namely, for *every* d, one can find d points which are shattered by $\mathcal{H}$ (see Exercise 8).

6.4 The Fundamental Theorem of PAC learning

We have already shown that a class of infinite VC dimension is not learnable. The converse statement is also true, which leads to the fundamental theorem of statistical learning theory:

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the $0-1$ loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC dimension.*

The proof of the theorem is given in the next section.

Not only does the VC dimension characterize PAC learnability, it even determines the sample complexity.

THEOREM 6.8 (The Fundamental Theorem of Statistical Learning—Quantitative version) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the $0-1$ loss. Assume that $\text{VCdim}(\mathcal{H}) = d < \infty$. Then, there are absolute constants C_1, C_2 such that:*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m^{UC}_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

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

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

The proof of this theorem is given in Chapter 28.

Remark 6.3 We stated the fundamental theorem for binary classification tasks. A similar result holds for some other learning problems such as regression with the absolute loss or the squared loss. However, the theorem does not hold for all learning tasks. In particular, learnability is sometimes possible even though the uniform convergence property does not hold (we will see an example in Chapter 13, Exercise 2). Furthermore, in some situations, the ERM rule fails but learnability is possible with other learning rules.

6.5 Proof of Theorem 6.7

We have already seen that $1 \to 2$ in Chapter 4. The implications $2 \to 3$ and $3 \to 4$ are trivial and so is $2 \to 5$. The implications $4 \to 6$ and $5 \to 6$ follow from the No-Free-Lunch theorem. The difficult part is to show that $6 \to 1$. The proof is based on two main claims:

- If $\text{VCdim}(\mathcal{H}) = d$, then even though $\mathcal{H}$ might be infinite, when restricting it to a finite set $C \subset \mathcal{X}$, its "effective" size, $|\mathcal{H}_C|$, is only $O(|C|^d)$. That is, the size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$. This claim is often referred to as *Sauer's lemma*, but it has also been stated and proved independently by Shelah and by Perles. The formal statement is given in Section 6.5.1 below.
- In Section 4 we have shown that finite hypothesis classes enjoy the uniform convergence property. In Section 6.5.2 below we generalize this result and show that uniform convergence holds whenever the hypothesis class has a "small effective size". By "small effective size" we mean classes for which $|\mathcal{H}_C|$ grows polynomially with $|C|$.

6.5.1 Sauer's Lemma and the growth function

We defined the notion of *shattering*, by considering the restriction of $\mathcal{H}$ to a finite set of instances. The growth function measures the maximal "effective" size of $\mathcal{H}$ on a set of m examples. Formally:

DEFINITION 6.9 (Growth function) Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as

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

In words, $\tau_H(m)$ is the number of different functions from a set C of size m to $\{0,1\}$ that can be obtained by restricting $\mathcal{H}$ to C.

Obviously, if $\text{VCdim}(\mathcal{H}) = d$ then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$. In such cases, $\mathcal{H}$ induces all possible functions from C to $\{0,1\}$. The following beautiful lemma, proposed independently by Sauer, Shelah, and Perles, shows that when m becomes larger than the VC dimension, the growth function increases polynomially rather than exponentially with m.

LEMMA 6.10 (Sauer-Shelah-Perles) *Let $\mathcal{H}$ be a hypothesis class with* $\text{VCdim}(\mathcal{H}) \leq d < \infty$. *Then, for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.*

Proof of Sauer's Lemma *

To prove the lemma it suffices to prove the following stronger claim: For any $C = \{c_1, \ldots, c_m\}$ we have

$$\forall \mathcal{H}, \quad |\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| . \tag{6.3}$$

The reason why Equation (6.3) is sufficient to prove the lemma is because if $\text{VCdim}(\mathcal{H}) \leq d$ then no set whose size is larger than d is shattered by $\mathcal{H}$ and therefore

$$|\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i} .$$

When $m > d+1$ the right-hand side of the above is at most $(em/d)^d$ (see Lemma A.5 in Appendix A).

We are left with proving Equation (6.3) and we do it using an inductive argument. For $m = 1$, no matter what $\mathcal{H}$ is, either both sides of Equation (6.3) equal 1 or both sides equal 2 (the empty set is always considered to be shattered by $\mathcal{H}$). Assume Equation (6.3) holds for sets of size $k < m$ and let us prove it for sets of size m. Fix $\mathcal{H}$ and $C = \{c_1, \ldots, c_m\}$. Denote $C' = \{c_2, \ldots, c_m\}$ and in addition, define the following two sets:

$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\} ,$$

and

$$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\} .$$

It is easy to verify that $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Additionally, since $Y_0 = \mathcal{H}_{C'}$, using the induction assumption (applied on $\mathcal{H}$ and C') we have that

$$|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}| .$$

Next, define $\mathcal{H}' \subseteq \mathcal{H}$ to be:

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } (1 - h'(c_1), h'(c_2), \ldots, h'(c_m)) = (h(c_1), h(c_2), \ldots, h(c_m))\} .$$

Namely, $\mathcal{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1. Using this definition, it is clear that if $\mathcal{H}'$ shatters a set $B \subseteq C'$ then it also shatters the set $B \cup \{c_1\}$ and vice versa. Combining this with the fact that $Y_1 = \mathcal{H}'_{C'}$ and using the inductive assumption (now applied on $\mathcal{H}'$ and C') we obtain that,

$$\begin{aligned} |Y_1| &= |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B\}| = |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \\ &= |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}' \text{ shatters } B\}| \leq |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}| . \end{aligned}$$

Overall, we have shown that

$$\begin{aligned} |\mathcal{H}_C| &= |Y_0| + |Y_1| \\ &\leq |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| , \end{aligned}$$

which concludes our proof.

6.5.2 Uniform convergence for classes of small effective size

In this section we prove that if $\mathcal{H}$ has small effective size then it enjoys the uniform convergence property. Formally,

THEOREM 6.11 *Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every $\mathcal{D}$ and every $\delta \in (0,1)$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ we have*

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta \sqrt{2m}} .$$

Before proving the theorem, let us first conclude the proof of Theorem 6.7.

Proof of Theorem 6.7 It suffices to prove that if the VC dimension is finite then the uniform convergence property holds. We will prove that

$$m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta) \leq 4 \frac{16d}{(\delta\epsilon)^2} \log\left(\frac{16d}{(\delta\epsilon)^2}\right) + \frac{16\, d \log(2e/d)}{(\delta\epsilon)^2} .$$

From Sauer's lemma we have that for $m > d$, $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$. Combining this with Theorem 6.11 we obtain that with probability of at least $1 - \delta$,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{d \log(2em/d)}}{\delta \sqrt{2m}} .$$

For simplicity assume that $\sqrt{d \log(2em/d)} \geq 4$, hence,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta} \sqrt{\frac{2d \log(2em/d)}{m}} .$$

To ensure that the above is at most ϵ we need that

$$m \geq \frac{2d\log(m)}{(\delta\epsilon)^2} + \frac{2\,d\log(2e/d)}{(\delta\epsilon)^2} \ .$$

Standard algebraic manipulations (see Lemma A.2 in Appendix A) show that a sufficient condition for the above to hold is that

$$m \geq 4\,\frac{2d}{(\delta\epsilon)^2}\,\log\left(\frac{2d}{(\delta\epsilon)^2}\right) + \frac{4\,d\log(2e/d)}{(\delta\epsilon)^2} \ .$$

□

Remark 6.4 The upper bound on $m_{\mathcal{H}}^{\mathrm{UC}}$ we derived in the proof Theorem 6.7 is not the tightest possible. A tighter analysis that yields the bounds given in Theorem 6.8 can be found in Chapter 28.

Proof of Theorem 6.11 *

We will start by showing that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}} \ . \tag{6.4}$$

Since the random variable $\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|$ is non-negative, the proof of the theorem follows directly from the above using Markov's inequality (see Section B.1).

To bound the left-hand side of Equation (6.4) we first note that for every $h \in \mathcal{H}$, we can rewrite $L_{\mathcal{D}}(h) = \mathbb{E}_{S'\sim\mathcal{D}^m}[L_{S'}(h)]$, where $S' = z'_1, \ldots, z'_m$ is an additional i.i.d. sample. Therefore,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\left|\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m} L_{S'}(h) - L_S(h)\right|\right] \ .$$

A generalization of the triangle inequality yields

$$\left|\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}[L_{S'}(h) - L_S(h)]\right| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}|L_{S'}(h) - L_S(h)| \ ,$$

and the fact that supermum of expectation is smaller than expectation of supremum yields

$$\sup_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}|L_{S'}(h) - L_S(h)| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\sup_{h\in\mathcal{H}}|L_{S'}(h) - L_S(h)| \ .$$

Formally, the above two inequalities follow from Jensen's inequality. Combining all the above we obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] &\leq \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{S'}(h) - L_S(h)|\right] \\ &= \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^{m}(\ell(h,z'_i) - \ell(h,z_i))\right|\right] \ .\end{aligned} \tag{6.5}$$

The expectation on the right-hand side is over a choice of two i.i.d. samples

$S = z_1, \ldots, z_m$ and $S' = z'_1, \ldots, z'_m$. Since all of these $2m$ vectors are chosen i.i.d., nothing would change if we replace the name of the random vector z_i with the name of the random vector z'_i. If we do it, instead of the term $(\ell(h, z'_i) - \ell(h, z_i))$ in Equation (6.5) we will have the term $-(\ell(h, z'_i) - \ell(h, z_i))$. It follows that for every $\boldsymbol{\sigma} \in \{\pm 1\}^m$ we have that Equation (6.5) equals to

$$\mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right]$$

Since this holds for every $\boldsymbol{\sigma} \in \{\pm 1\}^m$, it also holds if we sample each component of $\boldsymbol{\sigma}$ uniformly at random from the uniform distribution over $\{\pm 1\}$, denoted $U_{\pm}$. Hence, Equation (6.5) also equals to

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right] ,$$

and by the linearity of expectation it also equals to

$$\mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right] .$$

Next, fix S and S', and let C be the instances appearing in S and S'. Then, we can take the supremum only over $h \in \mathcal{H}_C$. Therefore,

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U_{\pm}^m} \left[\max_{h \in \mathcal{H}_C} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i)) \right| \right] . \end{aligned}$$

Fix some $h \in \mathcal{H}_C$ and denote $\theta_h = \frac{1}{m} \sum_{i=1}^{m} \sigma_i (\ell(h, z'_i) - \ell(h, z_i))$. Since $\mathbb{E}[\theta_h] = 0$ and θ_h is an average of independent variables, each of which take values in $[-1, 1]$, we have by Hoeffding's inequality that for every $\rho > 0$,

$$\mathbb{P}[|\theta_h| > \rho] \leq 2 \exp\left(-2\, m\, \rho^2\right) .$$

Applying the union bound over $h \in \mathcal{H}_C$, we obtain that for any $\rho > 0$,

$$\mathbb{P}\left[\max_{h \in \mathcal{H}_C} |\theta_h| > \rho \right] \leq 2\, |\mathcal{H}_C| \exp\left(-2\, m\, \rho^2\right) .$$

Finally, Lemma A.4 in Appendix A tells us that the above implies

$$\mathbb{E}\left[\max_{h \in \mathcal{H}_C} |\theta_h| \right] \leq \frac{4 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{2m}} .$$

Combining all the above with the definition of $\tau_{\mathcal{H}}$, we have shown that

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \right] \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}} .$$

Summary

The fundamental theorem of learning theory characterizes PAC learnability of classes of binary classifiers using VC dimension. The VC-dimension of a class is a combinatorial property which denotes the maximal sample size which can be shattered by the class. The fundamental theorem states that a class is PAC learnable if and only if its VC-dimension is finite, and specifies the sample complexity required for PAC learning. The theorem also shows that if a problem is at all learnable, then uniform convergence holds and therefore the problem is learnable using the ERM rule.

6.6 Bibliographic remarks

The definition of VC dimension and its relation to learnability and to uniform convergence is due to the seminal work of Vapnik & Chervonenkis (1971). The relation to the definition of PAC learnability is due to Blumer, Ehrenfeucht, Haussler & Warmuth (1989).

Several generalizations of the VC dimension have been proposed. For example, the fat-shattering dimension characterizes learnability of some regression problems (Kearns, Schapire & Sellie 1994, Alon, Ben-David, Cesa-Bianchi & Haussler 1997, Bartlett, Long & Williamson 1994, Anthony & Bartlet 1999), and the Natarajan dimension characterizes learnability of some multiclass learning problems (Natarajan 1989). However, in general, there is no equivalence between learnability and uniform convergence. See (Shalev-Shwartz, Shamir, Srebro & Sridharan 2010, Daniely, Sabato, Ben-David & Shalev-Shwartz 2011).

Sauer's lemma has been proved by Sauer in response to a problem of Erdos (Sauer 1972). Shelah (with Perles) proved it as a useful lemma for Shelah's theory of stable models (Shelah 1972). Gil Kalai tells[1] that at some later time, Benjy Weiss asked Perles about such a result in the context of ergodic theory and Perles, who forgot that he had proved it once, proved it again. Vapnik and Chervonenkis proved the lemma in the context of statistical learning theory.

6.7 Exercises

1. Show the following monotonicity property of VC-dimension: for every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$ then $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$.
2. Given some finite domain set, $\mathcal{X}$, and a number $k \leq |\mathcal{X}|$, figure out the VC dimension of each of the following classes (and prove your claims):
   1. $\mathcal{H}^{\mathcal{X}}_{=k} = \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k\}$. That is, the set of all functions that assign the value 1 to exactly k elements of $\mathcal{X}$.

[1] http://gilkalai.wordpress.com/2008/09/28/extremal-combinatorics-iii-some-basic-theorems

2. $\mathcal{H}_{at-most-k} = \{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| \leq k \text{ or } |\{x : h(x) = 0\}| \leq k\}$.
3. Let $\mathcal{X}$ be the boolean hypercube $\{0,1\}^n$. For a set $I \subseteq \{1, 2, \ldots, n\}$ we define a *parity function* h_I as follows. On a binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$,

$$h_I(\mathbf{x}) = \left(\sum_{i \in I} x_i\right) \bmod 2 .$$

(That is, h_I computes parity of bits in I.) What is the VC-dimension of the class of all such parity functions, $\mathcal{H}_{n\text{-parity}} = \{h_I : I \subseteq \{1, 2, \ldots, n\}\}$?

4. We proved Sauer's lemma by proving that for every class $\mathcal{H}$ of finite VC-dimension d, and every subset A of the domain,

$$|\mathcal{H}_A| \leq |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{|A|}{i} .$$

Show that there are cases in which the above two inequalities are strict (namely, the $\leq$ can be replaced by $<$) and cases in which they can be replaced by equalities. Demonstrate all four combinations of $=$ and $<$.

5. **VC-dimension of axis-aligned rectangles in $\mathbf{R}^d$:** Let $\mathcal{H}^d_{\text{rec}}$ be the class of axis-aligned rectangles in $\mathbf{R}^d$. We have already seen that $\text{VCdim}(\mathcal{H}^2_{\text{rec}}) = 4$. Prove that in general, $\text{VCdim}(\mathcal{H}^d_{\text{rec}}) = 2d$.
6. **VC-dimension of boolean conjunctions:** Let $\mathcal{H}^d_{con}$ be the class of boolean conjunctions over the variables $x_1, \ldots, x_d$ ($d \geq 2$). We already know that this class is finite and thus (agnostic) PAC-learnable. In this question we calculate $\text{VCdim}(\mathcal{H}^d_{con})$.
   1. Show that $|\mathcal{H}^d_{con}| \leq 3^d + 1$.
   2. Conclude that $\text{VCdim}(\mathcal{H}) \leq d \log 3$.
   3. Show that $\mathcal{H}^d_{con}$ shatters the set of unit vectors $\{\mathbf{e}_i : i \leq d\}$.
   4. (**) Show that $\text{VCdim}(\mathcal{H}^d_{con}) \leq d$.
      Hint: Assume by contradiction that there exists a set $C = \{c_1, \ldots, c_{d+1}\}$ that is shattered by $\mathcal{H}^d_{con}$. Let $h_1, \ldots, h_{d+1}$ be hypotheses in $\mathcal{H}^d_{con}$ which satisfy

$$\forall i, j \in [d+1],\ h_i(c_j) = \begin{cases} 0 & i = j \\ 1 & \text{otherwise} . \end{cases}$$

For each $i \in [d+1]$, h_i (or more accurately, the conjunction which corresponds to h_i) contains some literal ℓ_i which is false on c_i and true on c_j for each $j \neq i$. Use the Pigeonhole principle to show that there must be a pair $i < j \leq d+1$ such that ℓ_i and ℓ_j use the same x_k and use that fact to derive a contradiction to the requirements from the conjunctions h_i, h_j.
   5. Consider the class $\mathcal{H}^d_{mcon}$ of monotone boolean conjunctions over $\{0,1\}^d$. Monotonicity here means that the conjunctions do not contain negations.

As in $\mathcal{H}^d_{con}$, the empty conjunction is interpreted as the all-positive hypothesis. We augment $\mathcal{H}^d_{mcon}$ with the all-negative hypothesis h^-. Show that $\text{VCdim}(\mathcal{H}^d_{mcon}) = d$.

7. We have shown that for a finite hypothesis class $\mathcal{H}$, $\text{VCdim}(\mathcal{H}) \leq \lfloor \log(|\mathcal{H}|) \rfloor$. However, this is just an upper bound. The VC-dimension of a class can be much lower than that:
   1. Find an example of a class $\mathcal{H}$ of functions over the real interval $\mathcal{X} = [0, 1]$ such that $\mathcal{H}$ is infinite while $\text{VCdim}(\mathcal{H}) = 1$.
   2. Give an example of a finite hypothesis class $\mathcal{H}$ over the domain $\mathcal{X} = [0, 1]$, where $\text{VCdim}(\mathcal{H}) = \lfloor \log_2(|\mathcal{H}|) \rfloor$.
8. (*) It is often the case that the VC-dimension of a hypothesis class equals (or can be bounded above by) the number of parameters one needs to set in order to define each hypothesis in the class. For instance, if $\mathcal{H}$ is the class of axis-aligned rectangles in $\mathbf{R}^d$, then $\text{VCdim}(\mathcal{H}) = 2d$ which is equal to the number of parameters used to define a rectangle in $\mathbf{R}^d$. Here is some example which shows that this is not always the case. We will see that a hypothesis class might be very complex and even not learnable, although it has a small number of parameters.

   Consider the domain $\mathcal{X} = \mathbf{R}$, and the hypothesis class

   $$\mathcal{H} = \{x \mapsto \lceil \sin(\theta x) \rceil : \theta \in \mathbf{R}\}$$

   (here, we take $\lceil -1 \rceil = 0$). Prove that $\text{VCdim}(\mathcal{H}) = \infty$.

   Hint: There is more than one way to prove the required result. One option is by applying the following lemma: If $0.x_1x_2x_3\ldots$, is the binary expansion of $x \in (0, 1)$, then for any natural number m, $\lceil \sin(2^m \pi x) \rceil = (1 - x_m)$, provided that $\exists k \geq m$ s.t. $x_k = 1$.
9. Let $\mathcal{H}$ be the class of signed intervals, i.e.,
   $\mathcal{H} = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$ where

   $$h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}$$

   Calculate $\text{VCdim}(\mathcal{H})$.
10. Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0, 1\}$,
    1. Prove that if $\text{VCdim}(\mathcal{H}) \geq d$, for any d, then for some probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, for every sample size, m,

       $$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{d - m}{2d}$$

       Hint: Use Exercise 3 in Chapter 5.
    2. Prove that for every $\mathcal{H}$ which is PAC learnable, $\text{VCdim}(\mathcal{H}) < \infty$. (Note that this is the implication $3 \to 6$ in Theorem 6.7).
11. **VC of union:** Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = \max_i \text{VCdim}(\mathcal{H}_i)$ and assume for simplicity that $d \geq 3$.

1. Prove that,

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

Hint: Take a set of k examples and assume that they are shattered by the union class. Therefore, the union class can produce all 2^k possible labelings on these examples. Use Sauer's lemma to show that the union class cannot produce more than rk^d labelings. Therefore, $2^k < rk^d$. Now use Lemma A.2.

2. **(*)** Prove that for $r = 2$ it holds that

$$\text{VCdim}\left(\mathcal{H}_1 \cup \mathcal{H}_2\right) \leq 2d + 1 .$$

12. **Dudley classes:** In this question we discuss an algebraic framework for defining concept classes over $\mathbb{R}^n$ and show a connection between the VC dimension of such classes and their algebraic properties. Given a function $f : \mathbb{R}^n \to \mathbb{R}$ we define the corresponding function, $POS(f)(x) = \mathbb{1}_{[f(x)>0]}$. For a class $\mathcal{F}$ of real valued functions we define a corresponding class of functions $POS(\mathcal{F}) = \{POS(f) : f \in \mathcal{F}\}$. We say that a family, $\mathcal{F}$, of real valued functions is *linearly closed* if for all $f, g \in \mathcal{F}$ and $r \in \mathbb{R}$, $(f + rg) \in \mathcal{F}$ (where addition and scalar multiplication of functions is defined point wise, namely, for all $x \in \mathbb{R}^n$, $(f + rg)(x) = f(x) + rg(x)$). Note that if a family of functions is linearly closed then we can view it as a vector space over the reals. For a function $g : \mathbb{R}^n \to \mathbb{R}$ and a family of functions $\mathcal{F}$, let $\mathcal{F} + g \stackrel{\text{def}}{=} \{f + g : f \in \mathcal{F}\}$. Hypothesis classes that have a representation as $POS(\mathcal{F} + g)$ for some vector space of functions $\mathcal{F}$ and some function g are called *Dudley classes*.
    1. Show that for every $g : \mathbb{R}^n \to \mathbb{R}$ and every vector space of functions $\mathcal{F}$ as above, $\text{VCdim}(POS(\mathcal{F} + g)) = \text{VCdim}(POS(\mathcal{F}))$.
    2. **(**)** For every linearly closed family of real-valued functions $\mathcal{F}$, the VC-dimension of the corresponding class $POS(\mathcal{F})$ equals the linear dimension of $\mathcal{F}$ (as a vector space). *Hint:* Let $f_1, \ldots, f_d$ be a basis for the vector space $\mathcal{F}$. Consider the mapping $x \mapsto (f_1(x), \ldots, f_d(x))$ (from $\mathbb{R}^n$ to $\mathbb{R}^d$). Note that this mapping induces a matching between functions over $\mathbb{R}^n$ of the form $POS(f)$ and homogeneous linear half-spaces in $\mathbb{R}^d$ (the VC-dimension of the class of homogeneous linear half-spaces is analyzed in Chapter 9).
    3. Show that each of the following classes can be represented as a Dudley class:
        1. The class HS_n of halfspaces over $\mathbb{R}^n$ (see Chapter 9).
        2. The class HHS_n of all homogeneous halfspaces over $\mathbb{R}^n$ (see Chapter 9).
        3. The class B_d of all functions defined by (open) balls in $\mathbb{R}^d$. Use the Dudley representation to figure out the VC-dimension of this class.
        4. Let P_n^d denote the class of functions defined by polynomial inequalities of degree $\leq d$. Namely,

$$P_n^d = \{h_p : p \text{ is a polynomial of degree } \leq d \text{ in the variables } x_1, \ldots, x_n\},$$

where, for $\mathbf{x} = (x_1, \ldots, x_n)$, $h_p(\mathbf{x}) = \mathbb{1}_{[p(\mathbf{x}) \geq 0]}$ (the degree of a multivariable polynomial is the maximal sum of variable exponents over all of its terms. For example, the degree of $p(\mathbf{x}) = 3x_1^3x_2^2 + 4x_3x_7^2$ is 5).

1. Use the Dudley representation to figure out the VC-dimension of the class P_1^d - the class of all d-degree polynomials over $\mathbb{R}$.
2. Prove that the class of all polynomial classifiers over $\mathbb{R}$ has infinite VC-dimension.
3. Use the Dudley representation to figure out the VC-dimension of the class P_n^d (as a function of d and n).