5 Elements of Statistical Learning Theory

We now give a more complete exposition of the ideas of statistical learning theory, which we briefly touched on in Chapter 1. We mentioned previously that in order to learn from a small training set, we should try to *explain* the data with a model of *small* capacity; we have not yet justified *why* this is the case, however. This is the main goal of the present chapter.

Overview

We start by revisiting the difference between risk minimization and empirical risk minimization, and illustrating some common pitfalls in machine learning, such as overfitting and training on the test set (Section 5.1). We explain that the motivation for empirical risk minimization is the law of large numbers, but that the classical version of this law is not sufficient for our purposes (Section 5.2). Thus, we need to introduce the statistical notion of *consistency* (Section 5.3). It turns out that consistency of learning algorithms amounts to a law of large numbers, which holds uniformly over all functions that the learning machine can implement (Section 5.4). This crucial insight, due to Vapnik and Chervonenkis, focuses our attention on the set of attainable functions; this set must be restricted in order to have any hope of succeeding. Section 5.5 states probabilistic bounds on the risk of learning machines, and summarizes different ways of characterizing precisely how the set of functions can be restricted. This leads to the notion of *capacity concepts*, which gives us the main ingredients of the typical generalization error bound of statistical learning theory. We do not indulge in a complete treatment; rather, we try to give the main insights to provide the reader with some intuition as to how the different pieces of the puzzle fit together. We end with a section showing an example application of risk bounds for model selection (Section 5.6).

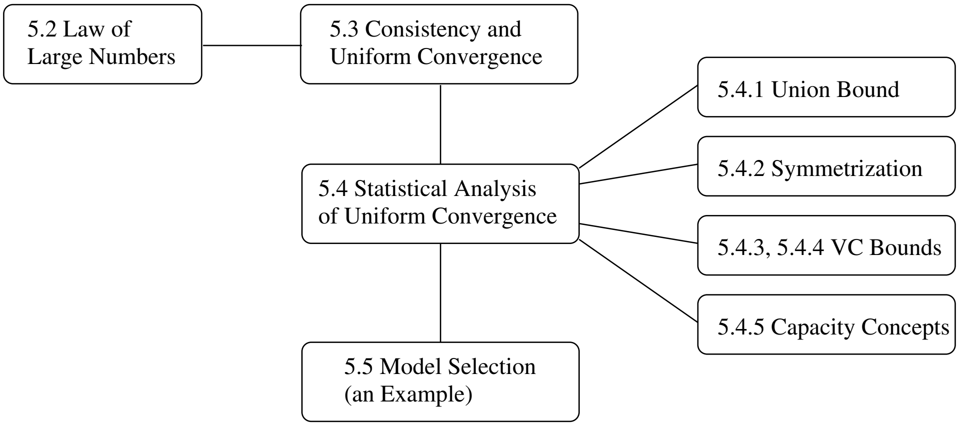
Prerequisites

The chapter attempts to present the material in a fairly non-technical manner, providing intuition wherever possible. Given the nature of the subject matter, however, a limited amount of mathematical background is required. The reader who is not familiar with basic probability theory should first read Section B.1.

5.1 Introduction

Let us start with an example. We consider a regression estimation problem. Suppose we are given empirical observations,

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}, \tag{5.1}$$


5.2 Law of Large Numbers
5.3 Consistency and Uniform Convergence
5.4.1 Union Bound
5.4.2 Symmetrization
5.4 Statistical Analysis of Uniform Convergence
5.4.3, 5.4.4 VC Bounds
5.4.5 Capacity Concepts
5.5 Model Selection (an Example)

Regression Example

where for simplicity we take $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Figure 5.1 shows a plot of such a dataset, along with two possible functional dependencies that could underlie the data. The dashed line represents a fairly complex model, and fits the training data perfectly. The straight line, on the other hand, does not completely "explain" the data, in the sense that there are some residual errors; it is much "simpler," however. A physicist measuring these data points would argue that it cannot be by chance that the measurements almost lie on a straight line, and would much prefer to attribute the residuals to measurement error than to an erroneous model. But is it possible to *characterize* the way in which the straight line is simpler, and why this should imply that it is, in some sense, closer to an underlying true dependency?

Bias-Variance Dilemma

In one form or another, this issue has long occupied the minds of researchers studying the problem of learning. In classical statistics, it has been studied as the *bias-variance dilemma*. If we computed a linear fit for every data set that we ever encountered, then every functional dependency we would ever "discover" would be linear. But this would not come from the data; it would be a *bias* imposed by us. If, on the other hand, we fitted a polynomial of sufficiently high degree to any given data set, we would always be able to fit the data perfectly, but the exact model we came up with would be subject to large fluctuations, depending on

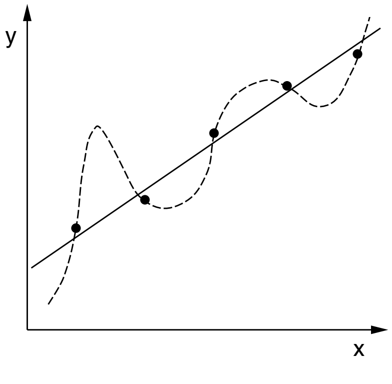


Figure 5.1 Suppose we want to estimate a functional dependence from a set of examples (black dots). Which model is preferable? The complex model perfectly fits all data points, whereas the straight line exhibits residual errors. Statistical learning theory formalizes the role of the *complexity* of the model class, and gives probabilistic guarantees for the validity of the inferred model.

how accurate our measurements were in the first place — the model would suffer from a large *variance*. A related dichotomy is the one between *estimation error* and *approximation error*. If we use a small class of functions, then even the best possible solution will poorly approximate the "true" dependency, while a large class of functions will lead to a large statistical estimation error.

Overfitting

In the terminology of applied machine learning and the design of neural networks, the complex explanation shows *overfitting*, while an overly simple explanation imposed by the learning machine design would lead to *underfitting*. A great deal of research has gone into clever engineering tricks and heuristics; these are used, for instance, to aid in the design of neural networks which will not overfit on a given data set [397]. In neural networks, overfitting can be avoided in a number of ways, such as by choosing a number of hidden units that is not too large, by *stopping* the training procedure early in order not to enforce a perfect explanation of the training set, or by using *weight decay* to limit the size of the weights, and thus of the function class implemented by the network.

Risk

Statistical learning theory provides a solid mathematical framework for studying these questions in depth. As mentioned in Chapters 1 and 3, it makes the assumption that the data are generated by sampling from an unknown underlying distribution $P(x, y)$. The learning problem then consists in minimizing the *risk* (or *expected loss* on the test data, see Definition 3.3),

$$R[f] = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y, f(x))\, dP(x, y). \tag{5.2}$$

Here, c is a loss function. In the case of pattern recognition, where $\mathcal{Y} = \{\pm 1\}$, a common choice is the misclassification error, $c(x, y, f(x)) = \frac{1}{2}|f(x) - y|$.

The difficulty of the task stems from the fact that we are trying to minimize a quantity that we cannot actually evaluate: since we do not know P, we cannot compute the integral (5.2). What we *do* know, however, are the training data (5.1), which are sampled from P. We can thus try to infer a function f from the training sample that is, in some sense, *close* to the one minimizing (5.2). To this end, we need what is called an *induction principle*.

Empirical Risk

One way to proceed is to use the training sample to approximate the integral in (5.2) by a finite sum (see (B.18)). This leads to the empirical risk (Definition 3.4),

$$R_{\text{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i)), \tag{5.3}$$

and the *empirical risk minimization (ERM) induction principle*, which recommends that we choose an f that minimizes (5.3).

Cast in these terms, the fundamental trade-off in learning can be stated as follows: if we allow f to be taken from a very large class of functions $\mathcal{F}$, we can always find an f that leads to a rather small value of (5.3). For instance, if we allow the use of *all* functions f mapping $\mathcal{X} \to \mathcal{Y}$ (in compact notation, $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$), then we can minimize (5.3) yet still be distant from the minimizer of (5.2). Considering a pattern recognition problem, we could set

$$f(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1, \ldots, m \\ 1 & \text{otherwise.} \end{cases} \tag{5.4}$$

This does not amount to any form of learning, however: suppose we are now given a test point drawn from the same distribution, $(x, y) \sim P(x, y)$. If $\mathcal{X}$ is a continuous domain, and we are not in a degenerate situation, the new pattern x will almost never be exactly equal to any of the training inputs x_i. Therefore, the learning machine will almost always predict that $y = 1$. *If we allow* all *functions from $\mathcal{X}$ to $\mathcal{Y}$, then the values of the function at points $x_1, \ldots, x_m$ carry no information about the values at other points.* In this situation, a learning machine cannot do better than chance. This insight lies at the core of the so-called *No-Free-Lunch Theorem* popularized in [608]; see also [254, 48].

The message is clear: if we make no restrictions on the class of functions from which we choose our estimate f, we cannot hope to learn anything. Consequently, machine learning research has studied various ways to implement such restrictions. In statistical learning theory, these restrictions are enforced by taking into account the *complexity* or *capacity* (measured by VC dimension, covering numbers, entropy numbers, or other concepts) of the class of functions that the learning machine can implement.[1]

In the Bayesian approach, a similar effect is achieved by placing *prior distributions* $P(f)$ over the class of functions (Chapter 16). This may sound fundamentally different, but it leads to algorithms which are closely related; and on the theoretical side, recent progress has highlighted intriguing connections [92, 91, 353, 238].

5.2 The Law of Large Numbers

Let us step back and try to look at the problem from a slightly different angle. Consider the case of pattern recognition using the misclassification loss function. Given a fixed function f, then for each example, the loss $\xi_i := \frac{1}{2}|f(x_i) - y_i|$ is either

[1] As an aside, note that the same problem applies to *training on the test set* (sometimes called *data snooping*): sometimes, people optimize tuning parameters of a learning machine by looking at how they change the results on an independent test set. Unfortunately, once one has adjusted the parameter in this way, the test set is not independent anymore. This is identical to the corresponding problem in training on the *training* set: once we have chosen the function to minimize the training error, the latter no longer provides an unbiased estimate of the test error. Overfitting occurs much faster on the training set, however, than it does on the test set. This is usually due to the fact that the number of tuning parameters of a learning machine is much smaller than the total number of parameters, and thus the capacity tends to be smaller. For instance, an SVM for pattern recognition typically has two tuning parameters, and optimizes m weight parameters (for a training set size of m). See also Problem 5.3 and [461].

0 or 1 (provided we have a ±1-valued function f), and all examples are drawn independently. In the language of probability theory, we are faced with *Bernoulli trials*. The $\xi_1, \ldots, \xi_m$ are independently sampled from a random variable

$$\xi := \frac{1}{2}|f(x) - y|. \tag{5.5}$$

Chernoff Bound

A famous inequality due to Chernoff [107] characterizes how the empirical mean $\frac{1}{m}\sum_{i=1}^{m} \xi_i$ converges to the expected value (or expectation) of ξ, denoted by $\mathbf{E}(\xi)$:

$$\mathrm{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} \xi_i - \mathbf{E}(\xi)\right| \geq \epsilon\right\} \leq 2\exp(-2m\epsilon^2) \tag{5.6}$$

Note that the P refers to the probability of getting a sample $\xi_1, \ldots, \xi_m$ with the property $\left|\frac{1}{m}\sum_{i=1}^{m} \xi_i - \mathbf{E}(\xi)\right| \geq \epsilon$. Mathematically speaking, P strictly refers to a so-called *product* measure (cf. (B.11)). We will presently avoid further mathematical detail; more information can be found in Appendix B.

In some instances, we will use a more general bound, due to Hoeffding (Theorem 5.1). Presently, we formulate and prove a special case of the Hoeffding bound, which implies (5.6). Note that in the following statement, the ξ_i are no longer restricted to take values in $\{0, 1\}$.

Hoeffding Bound

Theorem 5.1 (Hoeffding [244]) *Let ξ_i, $i \in [m]$ be m independent instances of a bounded random variable ξ, with values in $[a, b]$. Denote their average by $Q_m = \frac{1}{m}\sum_i \xi_i$. Then for any $\epsilon > 0$,*

$$\left.\begin{array}{l} \mathrm{P}\{Q_m - \mathbf{E}(\xi) \geq \epsilon\} \\ \mathrm{P}\{\mathbf{E}(\xi) - Q_m \geq \epsilon\} \end{array}\right\} \leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right). \tag{5.7}$$

The proof is carried out by using a technique commonly known as Chernoff's bounding method [107]. The proof technique is widely applicable, and generates bounds such as Bernstein's inequality [44] (exponential bounds based on the variance of random variables), as well as concentration-of-measure inequalities (see, e.g., [356, 66]). Readers not interested in the technical details underlying laws of large numbers may want to skip the following discussion.

We start with an auxiliary inequality.

Lemma 5.2 (Markov's Inequality (e.g., [136])) *Denote by ξ a nonnegative random variable with distribution* P*. Then for all $\lambda > 0$, the following inequality holds:*

$$\mathrm{P}\{\xi \geq \lambda\mathbf{E}(\xi)\} \leq \frac{1}{\lambda}. \tag{5.8}$$

Proof Using the definition of $\mathbf{E}(\xi)$, we have

$$\mathbf{E}(\xi) = \int_0^\infty \xi dP(\xi) \geq \int_{\lambda\mathbf{E}(\xi)}^\infty \xi dP(\xi) \geq \lambda\mathbf{E}(\xi)\int_{\lambda\mathbf{E}(\xi)}^\infty dP(\xi) = \lambda\mathbf{E}(\xi)\mathrm{P}\{\xi \geq \lambda\mathbf{E}(\xi)\}.$$

■

Proof of Theorem 5.1. Without loss of generality, we assume that $\mathbf{E}(\xi) = 0$ (otherwise simply define a random variable $\bar{\xi} := \xi - \mathbf{E}(\xi)$ and use the latter in the proof). Chernoff's bounding method consists in transforming a random variable ξ into $\exp(s\xi)$ ($s > 0$), and applying Markov's inequality to it. Depending on ξ, we can obtain different bounds. In our case, we use

$$\mathrm{P}\{\xi \geq \epsilon\} = \mathrm{P}\left\{\exp(s\xi) \geq \exp(s\epsilon)\right\} \leq e^{-s\epsilon}\mathbf{E}\left[\exp(s\xi)\right] \tag{5.9}$$

$$= e^{-s\epsilon}\mathbf{E}\left[\exp\left(\frac{s}{m}\sum_{i=1}^{m}\xi_i\right)\right] \leq e^{-s\epsilon}\prod_{i=1}^{m}\mathbf{E}\left[\exp\left(\frac{s}{m}\xi_i\right)\right]. \tag{5.10}$$

In (5.10), we exploited the fact that for positive random variables $\mathbf{E}\left[\prod_i \xi_i\right] \leq \prod_i \mathbf{E}\left[\xi_i\right]$. Since the inequality holds independent of the choice of s, we may minimize over s to obtain a bound that is as tight as possible. To this end, we transform the expectation over $\exp\left(\frac{s}{m}\xi_i\right)$ into something more amenable. The derivation is rather technical; thus we state without proof [244]: $\mathbf{E}\left[\exp(\frac{s}{m}\xi_i)\right] \leq \exp\left(\frac{s^2(b-a)^2}{8m^2}\right)$. From this, we conclude that the optimal value of s is given by $s = \frac{4m\epsilon}{(b-a)^2}$. Substituting this value into the right hand side of (5.10) proves the bound. ■

Let us now return to (5.6). Substituting (5.5) into (5.6), we have a bound which states how likely it is that for a given function f, the empirical risk is close to the actual risk,

$$\mathrm{P}\{|R_{\mathrm{emp}}[f] - R[f]| \geq \epsilon\} \leq 2\exp(-2m\epsilon^2). \tag{5.11}$$

Using Hoeffding's inequality, a similar bound can be given for the case of regression estimation, provided the loss $c(x, y, f(x))$ is bounded.

For any fixed function, the training error thus provides an unbiased estimate of the test error. Moreover, the convergence (in probability) $R_{\mathrm{emp}}[f] \to R[f]$ as $m \to \infty$ is exponentially fast in the number of training examples.[2] Although this sounds just about as good as we could possibly have hoped, there is one caveat: a crucial property of both the Chernoff and the Hoeffding bound is that they are probabilistic in nature. They state that the probability of a large deviation between test error and training error of f is small; the larger the sample size m, the smaller the probability. Granted, they do not rule out the presence of cases where the deviation is large, and our learning machine will have many functions that it can implement. Could there be a function for which things go wrong? It appears that

[2] *Convergence in probability*, denoted as

$$|R_{\mathrm{emp}}[f] - R[f]| \xrightarrow{\mathrm{P}} 0 \text{ as } m \to \infty,$$

means that for all $\epsilon > 0$, we have

$$\lim_{m\to\infty} \mathrm{P}\{|R_{\mathrm{emp}}[f] - R[f]| > \epsilon\} = 0.$$

we would be very unlucky for this to occur *precisely* for the function f chosen by empirical risk minimization.

At first sight, it seems that empirical risk minimization should work — in contradiction to our lengthy explanation in the last section, arguing that we have to do more than that. What is the catch?

5.3 When Does Learning Work: the Question of Consistency

It turns out that in the last section, we were too sloppy. When we find a function f by choosing it to minimize the training error, we are no longer looking at independent Bernoulli trials. We are actually choosing f such that the mean of the ξ_i is as small as possible. In this sense, we are actively looking for the worst case, for a function which is very atypical, with respect to the average loss (i.e., the empirical risk) that it will produce.

Consistency

We should thus state more clearly what it is that we actually need for empirical risk minimization to work. This is best expressed in terms of a notion that statisticians call *consistency*. It amounts to saying that as the number of examples m tends to infinity, we want the function f^m that minimizes $R_{\text{emp}}[f]$ (note that f^m need not be unique), to lead to a test error which converges to the lowest achievable value. In other words, f^m is asymptotically as good as whatever we could have done if we were able to directly minimize $R[f]$ (which we cannot, as we do not even know it). In addition, consistency requires that asymptotically, the training and the test error of f^m be identical.[3]

It turns out that *without restricting the set of admissible functions*, empirical risk minimization is not consistent. The main insight of VC (Vapnik-Chervonenkis) theory is that actually, the *worst case* over all functions that the learning machine can implement determines the consistency of empirical risk minimization. In other words, we need a version of the law of large numbers which is *uniform* over all functions that the learning machine can implement.

5.4 Uniform Convergence and Consistency

The present section will explain how consistency can be characterized by a uniform convergence condition on the set of functions $\mathcal{F}$ that the learning machine can implement. Figure 5.2 gives a simplified depiction of the question of consistency. Both the empirical risk and the actual risk are drawn as functions of f. For

[3] We refrain from giving a more formal definition of consistency, the reason being that there are some caveats to this classical definition of consistency; these would necessitate a discussion leading us away from the main thread of the argument. For the precise definition of the required notion of "nontrivial consistency," see [561].

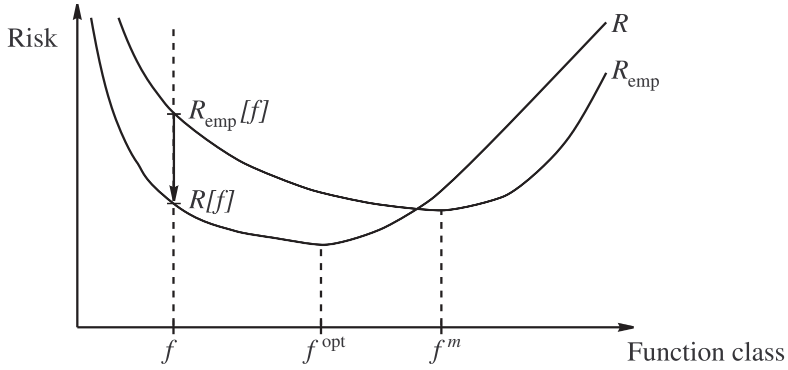


Figure 5.2 Simplified depiction of the convergence of empirical risk to actual risk. The x-axis gives a one-dimensional representation of the function class; the y axis denotes the risk (error). For each *fixed* function f, the law of large numbers tells us that as the sample size goes to infinity, the empirical risk $R_{\text{emp}}[f]$ converges towards the true risk $R[f]$ (indicated by the downward arrow). This does not imply, however, that in the limit of infinite sample sizes, the minimizer of the empirical risk, f^m, will lead to a value of the risk that is as good as the best attainable risk, $R[f^{\text{opt}}]$ *(consistency)*. For the latter to be true, we require convergence of $R_{\text{emp}}[f]$ towards $R[f]$ to be uniform over all functions that the learning machines can implement (see text).

simplicity, we have summarized all possible functions f by a single axis of the plot. Empirical risk minimization consists in picking the f that yields the minimal value of R_{emp}. If it is consistent, then the minimum of R_{emp} converges to that of R in probability. Let us denote the minimizer of R by f^{opt}, satisfying

$$R[f] - R[f^{\text{opt}}] \geq 0 \tag{5.12}$$

for all $f \in \mathcal{F}$. This is the optimal choice that we could make, given complete knowledge of the distribution P.[4] Similarly, since f^m minimizes the empirical risk, we have

$$R_{\text{emp}}[f] - R_{\text{emp}}[f^m] \geq 0, \tag{5.13}$$

for all $f \in \mathcal{F}$. Being true for all $f \in \mathcal{F}$, (5.12) and (5.13) hold in particular for f^m and f^{opt}. If we substitute the former into (5.12) and the latter into (5.13), we obtain

$$R[f^m] - R[f^{\text{opt}}] \geq 0, \tag{5.14}$$

and

$$R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \geq 0. \tag{5.15}$$

[4] As with f^m, f^{opt} need not be unique.

The sum of these two inequalities satisfies

$$\begin{aligned}0 &\leq R[f^m] - R[f^{\text{opt}}] + R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \\ &= R[f^m] - R_{\text{emp}}[f^m] + R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}] \\ &\leq \sup_{f\in\mathcal{F}} \left(R[f] - R_{\text{emp}}[f]\right) + R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}]. \end{aligned} \tag{5.16}$$

Let us first consider the second half of the right hand side. Due to the law of large numbers, we have convergence in probability, i.e., for all $\epsilon > 0$,

$$|R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}]| \xrightarrow{P} 0 \text{ as } m \to \infty. \tag{5.17}$$

This holds true since f^{opt} is a fixed function, which is independent of the training sample (see (5.11)).

Uniform Convergence of Risk

The important conclusion is that if the empirical risk converges to the actual risk one-sided *uniformly*, over all functions that the learning machine can implement,

$$\sup_{f\in\mathcal{F}}(R[f] - R_{\text{emp}}[f]) \xrightarrow{P} 0 \text{ as } m \to \infty, \tag{5.18}$$

then the left hand sides of (5.14) and (5.15) will likewise converge to 0;

$$R[f^m] - R[f^{\text{opt}}] \xrightarrow{P} 0, \tag{5.19}$$

$$R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \xrightarrow{P} 0. \tag{5.20}$$

As argued above, (5.17) is not always true for f^m, since f^m is chosen to minimize R_{emp}, and thus depends on the sample. Assuming that (5.18) holds true, however, then (5.19) and (5.20) imply that in the limit, $R[f^m]$ cannot be larger than $R_{\text{emp}}[f^m]$. One-sided uniform convergence on $\mathcal{F}$ is thus a sufficient condition for consistency of the empirical risk minimization over $\mathcal{F}$.[5]

What about the other way round? Is one-sided uniform convergence also a *necessary* condition? Part of the mathematical beauty of VC theory lies in the fact that this is the case. We cannot go into the necessary details to prove this [571, 561, 562], and only state the main result. Note that this theorem uses the notion of nontrivial consistency that we already mentioned briefly in footnote 3. In a nutshell, this concept requires that the induction principle be consistent even after the "best" functions have been removed. Nontrivial consistency thus rules out, for instance, the case in which the problem is trivial, due to the existence of a function which uniformly does better than all other functions. To understand this, assume that there exists such a function. Since this function is uniformly better than all others, we can already select this function (using ERM) from *one* (arbitrary) data point. Hence the method would be trivially consistent, no matter what the

[5] Note that the onesidedness of the convergence comes from the fact that we only require consistency of empirical risk *minimization*. If we required the same for empirical risk *maximization*, then we would end up with standard uniform convergence, and the parentheses in (5.18) would be replaced with modulus signs.

rest of the function class looks like. Having one function which gets picked as soon as we have seen one data point would essentially void the inherently *asymptotic* notion of consistency.

Theorem 5.3 (Vapnik & Chervonenkis (e.g., [562])) *One-sided uniform convergence in probability,*

$$\lim_{m\to\infty} \mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} = 0, \tag{5.21}$$

for all $\epsilon > 0$, is a necessary and sufficient *condition for nontrivial consistency of empirical risk minimization.*

As explained above, consistency, and thus learning, crucially depends on the set of functions. In Section 5.1, we gave an example where we considered the set of all possible functions, and showed that learning was impossible. The dependence of learning on the set of functions has now returned in a different guise: the condition of uniform convergence will crucially depend on the set of functions for which it must hold.

The abstract characterization in Theorem 5.3 of consistency as a uniform convergence property, whilst theoretically intriguing, is not all that useful in practice. We do not want to check some fairly abstract convergence property every time we want to use a learning machine. Therefore, we next address whether there are properties of learning machines, i.e., of sets of functions, which *ensure* uniform convergence of risks.

5.5 How to Derive a VC Bound

We now take a closer look at the subject of Theorem 5.3; the probability

$$\mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f] - R_{\mathrm{emp}}[f]) > \epsilon\}. \tag{5.22}$$

We give a simplified account, drawing from the expositions of [561, 562, 415, 238]. We do not aim to describe or even develop the theory to the extent that would be necessary to give precise bounds for SVMs, say. Instead, our goal will be to convey central insights rather than technical details. For more complete treatments geared specifically towards SVMs, cf. [562, 491, 24]. We focus on the case of pattern recognition; that is, on functions taking values in $\{\pm 1\}$.

Two tricks are needed along the way: the *union bound* and the method of *symmetrization by a ghost sample*.

5.5.1 The Union Bound

Suppose the set $\mathcal{F}$ consists of two functions, f_1 and f_2. In this case, uniform convergence of risk trivially follows from the law of large numbers, which holds for each of the two. To see this, let

$$C_\epsilon^i := \{(x_1, y_1), \ldots, (x_m, y_m) | (R[f_i] - R_{\text{emp}}[f_i]) > \epsilon\} \tag{5.23}$$

denote the set of samples for which the risks of f_i differ by more than ϵ. Then, by definition, we have

$$\mathrm{P}\{\sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon\} = \mathrm{P}(C_\epsilon^1 \cup C_\epsilon^2). \tag{5.24}$$

The latter, however, can be rewritten as

$$\mathrm{P}(C_\epsilon^1 \cup C_\epsilon^2) = \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2) - \mathrm{P}(C_\epsilon^1 \cap C_\epsilon^2) \leq \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2), \tag{5.25}$$

where the last inequality follows from the fact that P is nonnegative. Similarly, if $\mathcal{F} = \{f_1, \ldots, f_n\}$, we have

$$\mathrm{P}\{\sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon\} = \mathrm{P}(C_\epsilon^1 \cup \ldots \cup C_\epsilon^n) \leq \sum_{i=1}^n \mathrm{P}(C_\epsilon^i). \tag{5.26}$$

Union Bound

This inequality is called the *union bound*. As it is a crucial step in the derivation of risk bounds, it is worthwhile to emphasize that it becomes an equality if and only if all the events involved are *disjoint*. In practice, this is rarely the case, and we therefore lose a lot when applying (5.26). It is a step with a large "slack."

Nevertheless, when $\mathcal{F}$ is finite, we may simply apply the law of large numbers (5.11) for each individual $\mathrm{P}(C_\epsilon^i)$, and the sum in (5.26) then leads to a constant factor n on the right hand side of the bound — it does not change the exponentially fast convergence of the empirical risk towards the actual risk. In the next section, we describe an ingenious trick used by Vapnik and Chervonenkis, to reduce the infinite case to the finite one. It consists of introducing what is sometimes called a *ghost sample*.

5.5.2 Symmetrization

The central observation in this section is that we can bound (5.22) in terms of a probability of an event referring to a *finite* function class. Note first that the empirical risk term in (5.22) effectively refers only to a finite function class: for any given training sample of m points $x_1, \ldots, x_m$, the functions of $\mathcal{F}$ can take at most 2^m different values $y_1, \ldots, y_m$ (recall that the y_i take values only in $\{\pm 1\}$). In addition, the probability that the empirical risk differs from the actual risk by more than ϵ, can be bounded by the twice the probability that it differs from the empirical risk on a *second* sample of size m by more than $\epsilon/2$.

Symmetrization

Lemma 5.4 (Symmetrization (Vapnik & Chervonenkis) (e.g. [559])) *For $m\epsilon^2 \geq 2$, we have*

$$\mathrm{P}\{\sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon\} \leq 2\mathrm{P}\{\sup_{f \in \mathcal{F}}(R_{\text{emp}}[f] - R'_{\text{emp}}[f]) > \epsilon/2\}. \tag{5.27}$$

Here, the first P *refers to the distribution of iid samples of size m, while the second one* *refers to iid samples of size* $2m$*. In the latter case,* R_{emp} *measures the loss on the first half of the sample, and* R'_{emp} *on the second half.*

Although we do not prove this result, it should be fairly plausible: if the empirical error rates on two independent m-samples are close to each other, then they should also be close to the true error rate.

5.5.3 The Shattering Coefficient

The main result of Lemma 5.4 is that it implies, for the purpose of bounding (5.22), that the function class $\mathcal{F}$ is effectively finite: restricted to the $2m$ points appearing on the right hand side of (5.27), it has *at most* 2^{2m} elements. This is because only the outputs of the functions on the patterns of the sample count, and there are $2m$ patterns with two possible outputs, ± 1. The number of effectively different functions can be smaller than 2^{2m}, however; and for our purposes, this is the case that will turn out to be interesting.

Let $Z_{2m} := \big((x_1, y_1), \ldots, (x_{2m}, y_{2m})\big)$ be the given $2m$-sample. Denote by $\mathcal{N}(\mathcal{F}, Z_{2m})$ the cardinality of $\mathcal{F}$ when restricted to $\{x_1, \ldots, x_{2m}\}$, that is, the number of functions from $\mathcal{F}$ that can be distinguished from their values on $\{x_1, \ldots, x_{2m}\}$. Let us, moreover, denote the maximum (over all possible choices of a $2m$-sample) number of functions that can be distinguished in this way as $\mathcal{N}(\mathcal{F}, 2m)$.

Shattering Coefficient

The function $\mathcal{N}(\mathcal{F}, m)$ is referred to as the *shattering coefficient*, or in the more general case of regression estimation, the *covering number* of $\mathcal{F}$.[6] In the case of pattern recognition, which is what we are currently looking at, $\mathcal{N}(\mathcal{F}, m)$ has a particularly simple interpretation: it is the number of different outputs $(y_1, \ldots, y_m)$ that the functions in $\mathcal{F}$ can achieve on samples of a given size.[7] In other words, it simply measures the *number of ways that the function class can separate the patterns into two classes*. Whenever $\mathcal{N}(\mathcal{F}, m) = 2^m$, all possible separations can be implemented by functions of the class. In this case, the function class is said to *shatter* m points. Note that this means that there *exists* a set of m patterns which can be separated in all possible ways — it does not mean that this applies to *all* sets of m patterns.

Shattering

5.5.4 Uniform Convergence Bounds

Let us now take a closer look at the probability that for a $2m$-sample Z_{2m} drawn iid from P, we get a one-sided uniform deviation larger than $\epsilon/2$ (cf. (5.27)),

$$\mathrm{P}\{\sup_{f\in\mathcal{F}}(R_{\text{emp}}[f] - R'_{\text{emp}}[f]) > \epsilon/2\}. \tag{5.28}$$

[6] In regression estimation, the covering number also depends on the accuracy within which we are approximating the function class, and on the loss function used; see Section 12.4 for more details.

[7] Using the zero-one loss $c(x, y, f(x)) = 1/2|f(x) - y| \in \{0, 1\}$, it also equals the number of different loss vectors $(c(x_1, y_1, f(x_1)), \ldots, c(x_m, y_m, f(x_m)))$.

The basic idea now is to pick a maximal set of functions $\{f_1, \ldots, f_{\mathcal{N}(\mathcal{F}, Z_{2m})}\}$ that can be distinguished based on their values on Z_{2m}, then use the union bound, and finally bound each term using the Chernoff inequality. However, the fact that the f_i depend on the sample Z_{2m} will make things somewhat more complicated. To deal with this, we have to introduce an auxiliary step of randomization, using a uniform distibution over permutations σ of the $2m$-sample Z_{2m}.

Let us denote the empirical risks on the two halves of the sample after the permutation σ by $R^{\sigma}_{\text{emp}}[f]$ and $R'^{\sigma}_{\text{emp}}[f]$, respectively. Since the $2m$-sample is iid, the permutation does not affect (5.28). We may thus instead consider

$$\mathrm{P}_{Z_{2m},\sigma}\{\sup_{f \in \mathcal{F}}(R^{\sigma}_{\text{emp}}[f] - R'^{\sigma}_{\text{emp}}[f]) > \epsilon/2\}, \tag{5.29}$$

where the subscripts of P were added to clarify what the distribution refers to. We next rewrite this as

$$\int_{(\mathcal{X}\times\{\pm 1\})^{2m}} \mathrm{P}_{\sigma|Z_{2m}}\{\sup_{f \in \mathcal{F}|_{Z_{2m}}} (R^{\sigma}_{\text{emp}}[f] - R'^{\sigma}_{\text{emp}}[f]) > \epsilon/2 \mid Z_{2m}\}\, d\mathrm{P}(Z_{2m}). \tag{5.30}$$

We can now express the event $C_\epsilon := \{\sigma | \sup_{f \in \mathcal{F}|_{Z_{2m}}} (R^{\sigma}_{\text{emp}}[f] - R'^{\sigma}_{\text{emp}}[f]) > \epsilon/2\}$ as

$$C_\epsilon = \bigcup_{n=1}^{\mathcal{N}(\mathcal{F}, Z_{2m})} C_\epsilon(f_n), \tag{5.31}$$

where the events $C_\epsilon(f_n) := \{\sigma | (R^{\sigma}_{\text{emp}}[f_n] - R'^{\sigma}_{\text{emp}}[f_n]) > \epsilon/2\}$ refer to individual functions f_n chosen such that $\left(\bigcup_n \{f_n\}\right)|_{Z_{2m}} = \mathcal{F}|_{Z_{2m}}$. Note that the functions f_n may be considered as fixed, since we have conditioned on Z_{2m}.

We are now in a position to appeal to the classical law of large numbers. Our random experiment consists of drawing σ from the uniform distribution over all permutations of $2m$-samples. This turns our sequence of losses $\xi^{\sigma}_i = \frac{1}{2}|f(x^{\sigma}_i) - y^{\sigma}_i|$ ($i = 1, \ldots, 2m$) into an iid sequence of independent Bernoulli trials. We then apply a modified Chernoff inequality to bound the probability of each event $C_\epsilon(f_n)$. It states that given a $2m$-sample of Bernoulli trials, we have (see Problem 5.4)

$$\mathrm{P}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i \geq \epsilon\right\} \leq 2\,\exp\left(-\frac{m\epsilon^2}{2}\right). \tag{5.32}$$

For our present problem, we thus obtain

$$\mathrm{P}_{\sigma|Z_{2m}}(C_\epsilon(f_n)) \leq 2\,\exp\left(-\frac{m\epsilon^2}{8}\right), \tag{5.33}$$

independent of f_n. We next use the union bound to get a bound on the probability of the event C_ϵ defined in (5.31). We obtain a sum over $\mathcal{N}(\mathcal{F}, Z_{2m})$ identical terms of the form (5.33). Hence (5.30) (and (5.29)) can be bounded from above by

$$\begin{aligned}&\int_{(\mathcal{X}\times\{\pm 1\})^{2m}} \mathcal{N}(\mathcal{F}, Z_{2m})\, 2\,\exp\left(-\frac{m\epsilon^2}{8}\right) d\mathrm{P}(Z_{2m})\\ &= 2\,\mathbf{E}\left[\mathcal{N}(\mathcal{F}, Z_{2m})\right]\exp\left(-\frac{m\epsilon^2}{8}\right),\end{aligned} \tag{5.34}$$

where the expectation is taken over the random drawing of Z_{2m}. The last step is to combine this with Lemma 5.4, to obtain

$$\begin{aligned} \mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f]-R_{\mathrm{emp}}[f])>\epsilon\} &\leq 4\,\mathbf{E}\,[\mathcal{N}(\mathcal{F},Z_{2m})]\,\exp\left(-\frac{m\epsilon^2}{8}\right) \\ &= 4\,\exp\left(\ln\mathbf{E}\,[\mathcal{N}(\mathcal{F},Z_{2m})]-\frac{m\epsilon^2}{8}\right). \qquad (5.35) \end{aligned}$$

Inequality of Vapnik-Chervonenkis Type

We conclude that provided $\mathbf{E}\,[\mathcal{N}(\mathcal{F},Z_{2m})]$ does not grow exponentially in m (i.e., $\ln\mathbf{E}\,[\mathcal{N}(\mathcal{F},Z_{2m})]$ grows sublinearly), it is actually possible to make nontrivial statements about the *test* error of learning machines.

The above reasoning is essentially the VC style analysis. Similar bounds can be obtained using a strategy which is more common in the field of empirical processes, first proving that $\sup_f(R[f]-R_{\mathrm{emp}}[f])$ is concentrated around its mean [554, 14].

5.5.5 Confidence Intervals

It is sometimes useful to rewrite (5.35) such that we specify the probability with which we want the bound to hold, and then get the confidence interval, which tells us how close the risk should be to the empirical risk. This can be achieved by setting the right hand side of (5.35) equal to some $\delta>0$, and then solving for ϵ. As a result, we get the statement that with a probability at least $1-\delta$,

Risk Bound

$$R[f]\leq R_{\mathrm{emp}}[f]+\sqrt{\frac{8}{m}\left(\ln\mathbf{E}\,[\mathcal{N}(\mathcal{F},Z_{2m})]+\ln\frac{4}{\delta}\right)}. \qquad (5.36)$$

Note that this bound holds independent of f; in particular, it holds for the function f^m minimizing the empirical risk. This is not only a strength, but also a weakness in the bound. It is a strength since many learning machines do not truly minimize the empirical risk, and the bound thus holds for them, too. It is a weakness since by taking into account more information on which function we are interested in, one could hope to get more accurate bounds. We will return to this issue in Section 12.1.

Bounds like (5.36) can be used to justify induction principles different from the empirical risk minimization principle. Vapnik and Chervonenkis [569, 559] proposed minimizing the right hand side of these bounds, rather than just the empirical risk. The confidence term, in the present case, $\sqrt{\frac{8}{m}\left(\ln\mathbf{E}\,[\mathcal{N}(\mathcal{F},Z_{2m})]+\ln\frac{4}{\delta}\right)}$, then ensures that the chosen function, denoted f_*, not only leads to a small risk, but also comes from a function class with small capacity.

The capacity term is a property of the function class $\mathcal{F}$, and not of any individual function f. Thus, the bound cannot simply be minimized over choices of f. Instead, we introduce a so-called *structure* on $\mathcal{F}$, and minimize over the choice of the structure. This leads to an induction principle called *structural risk minimization*. We leave out the technicalities involved [559, 136, 562]. The main idea is depicted in Figure 5.3.

Structural Risk Minimization

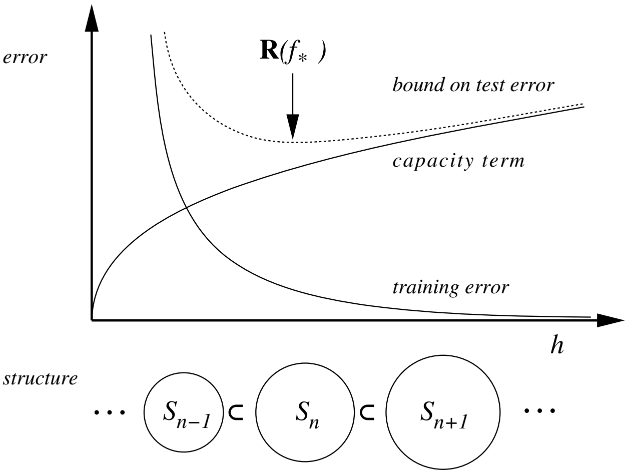


Figure 5.3 Graphical depiction of the structural risk minimization (SRM) induction principle. The function class is decomposed into a nested sequence of subsets of increasing size (and thus, of increasing capacity). The SRM principle picks a function f_* which has small training error, and comes from an element of the structure that has low capacity h, thus minimizing a risk bound of type (5.36).

For practical purposes, we usually employ bounds of the type (5.36) as a guideline for coming up with risk functionals (see Section 4.1). Often, the risk functionals form a compromise between quantities that *should* be minimized from a statistical point of view, and quantities that *can* be minimized efficiently (cf. Problem 5.7).

There exists a large number of bounds similar to (5.35) and its alternative form (5.36). Differences occur in the constants, both in front of the exponential and in its exponent. The bounds also differ in the exponent of ϵ — in some cases, by a factor greater than 2. For instance, if a training error of zero is achievable, we can use Bernstein's inequality instead of Chernoff's result, which leads to ϵ rather than ϵ^2. For further details, cf. [136, 562, 492, 238]. Finally, the bounds differ in the way they measure capacity. So far, we have used covering numbers, but this is not the only method.

5.5.6 The VC Dimension and Other Capacity Concepts

So far, we have formulated the bounds in terms of the so-called *annealed entropy* $\ln \mathbf{E}\left[\mathcal{N}(\mathcal{F}, Z_{2m})\right]$. This led to statements that depend on the distribution and thus can take into account characteristics of the problem at hand. The downside is that they are usually difficult to evaluate; moreover, in most problems, we do not have knowledge of the underlying distribution. However, a number of different capacity concepts, with different properties, can take the role of the term $\ln(\mathbf{E}\left[\mathcal{N}(\mathcal{F}, Z_{2m})\right])$ in (5.36).

- Given an example (x, y), $f \in \mathcal{F}$ causes a loss that we denote by $c(x, y, f(x)) := \frac{1}{2}|f(x) - y| \in \{0, 1\}$. For a larger sample $(x_1, y_1) \ldots, (x_m, y_m)$, the different functions

VC Entropy

$f \in \mathcal{F}$ lead to a *set* of loss vectors $\boldsymbol{\xi}_f = (c(x_1, y_1, f(x_1)), \ldots, c(x_m, y_m, f(x_m)))$, whose cardinality we denote by $\mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)$. The *VC entropy* is defined as

$$H_{\mathcal{F}}(m) = \mathbf{E}\left[\ln \mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)\right], \tag{5.37}$$

where the expectation is taken over the random generation of the m-sample $(x_1, y_1) \ldots, (x_m, y_m)$ from P.

One can show [562] that the convergence

$$\lim_{m \to \infty} H_{\mathcal{F}}(m)/m = 0, \tag{5.38}$$

is equivalent to uniform (two-sided) convergence of risk,

$$\lim_{m \to \infty} \mathrm{P}\{\sup_{f \in \mathcal{F}} |R[f] - R_{\mathrm{emp}}[f]| > \epsilon\} = 0, \tag{5.39}$$

for all $\epsilon > 0$. By Theorem 5.3, (5.39) thus implies consistency of empirical risk minimization.

Annealed Entropy

■ If we exchange the expectation $\mathbf{E}$ and the logarithm in (5.37), we obtain the annealed entropy used above,

$$H_{\mathcal{F}}^{\mathrm{ann}}(m) = \ln \mathbf{E}\left[\mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)\right]. \tag{5.40}$$

Since the logarithm is a concave function, the annealed entropy is an upper bound on the VC entropy. Therefore, whenever the annealed entropy satisfies a condition of the form (5.38), the same automatically holds for the VC entropy.

One can show that the convergence

$$\lim_{m \to \infty} H_{\mathcal{F}}^{\mathrm{ann}}(m)/m = 0, \tag{5.41}$$

implies exponentially fast convergence [561],

$$P\{\sup_{f \in \mathcal{F}} |R[f] - R_{\mathrm{emp}}[f]| > \epsilon\} \leq 4\ \exp(((H_{\mathcal{F}}^{\mathrm{ann}}(2m)/m) - \epsilon^2) \cdot m). \tag{5.42}$$

It has recently been proven that in fact (5.41) is not only sufficient, but also necessary for this [66].

Growth Function

■ We can obtain an upper bound on both entropies introduced so far, by taking a supremum over all possible samples, instead of the expectation. This leads to the *growth function*,

$$G_{\mathcal{F}}(m) = \max_{(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}} \ln \mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right). \tag{5.43}$$

Note that by definition, the growth function is the logarithm of the shattering coefficient, $G_{\mathcal{F}}(m) = \ln \mathcal{N}(\mathcal{F}, m)$.

The convergence

$$\lim_{m \to \infty} G_{\mathcal{F}}(m)/m = 0, \tag{5.44}$$

is necessary and sufficient for exponentially fast convergence of risk *for all underlying distributions* P.

■ The next step will be to summarize the main behavior of the growth function with a single number. If $\mathcal{F}$ is as rich as possible, so that for any sample of size m, the points can be chosen such that by using functions of the learning machine, they can be separated in all 2^m possible ways (i.e., they can be shattered), then

$$G_{\mathcal{F}}(m) = m \cdot \ln(2). \tag{5.45}$$

VC Dimension

In this case, the convergence (5.44) does not take place, and learning will not generally be successful. What about the other case? Vapnik and Chervonenkis [567, 568] showed that either (5.45) holds true for all m, or there exists some *maximal* m for which (5.45) is satisfied. This number is called the *VC dimension* and is denoted by h. If the maximum does not exist, the VC dimension is said to be infinite.

By construction, the VC dimension is thus the maximal number of points which can be shattered by functions in $\mathcal{F}$. It is possible to prove that for $m > h$ [568],

$$G_{\mathcal{F}}(m) \leq h\left(\ln \frac{m}{h} + 1\right). \tag{5.46}$$

This means that up to $m = h$, the growth function increases linearly with the sample size. Thereafter, it only increases logarithmically, i.e., *much* more slowly. This is the regime where learning can succeed.

Although we do not make use of it in the present chapter, it is worthwhile to also introduce the *VC dimension of a class of real-valued functions* $\{f_{\mathbf{w}} | \mathbf{w} \in \Lambda\}$ at this stage. It is defined to equal the VC dimension of the class of indicator functions

$$\left\{\operatorname{sgn}(f_{\mathbf{w}} - \beta) | \mathbf{w} \in \Lambda, \beta \in \left(\inf_x f_{\mathbf{w}}(x), \sup_x f_{\mathbf{w}}(x)\right)\right\}. \tag{5.47}$$

VC Dimension for Real-Valued Functions

In summary, we get a succession of capacity concepts,

$$H_{\mathcal{F}}(m) \leq H_{\mathcal{F}}^{\text{ann}}(m) \leq G_{\mathcal{F}}(m) \leq h\left(\ln \frac{m}{h} + 1\right). \tag{5.48}$$

From left to right, these become less precise. The entropies on the left are distribution-dependent, but rather difficult to evaluate (see, e.g., [430, 391]). The growth function and VC dimension are distribution-independent. This is less accurate, and does not always capture the essence of a given problem, which might have a much more benign distribution than the worst case; on the other hand, we want the learning machine to work for all distributions. If we knew the distribution beforehand, then we would not need a learning machine anymore.

VC Dimension Example

Let us look at a simple example of the VC dimension. As a function class, we consider hyperplanes in $\mathbb{R}^2$, i.e.,

$$f(x) = \operatorname{sgn}(a + b[x]_1 + c[x]_2), \quad \text{with parameters } a, b, c \in \mathbb{R}. \tag{5.49}$$

Suppose we are given three points x_1, x_2, x_3 which are not collinear. No matter how they are labelled (that is, independent of our choice of $y_1, y_2, y_3 \in \{\pm 1\}$), we can always find parameters $a, b, c \in \mathbb{R}$ such that $f(x_i) = y_i$ for all i (see Figure 1.4 in the introduction). In other words, there exist three points that we can shatter. This shows that the VC dimension of the set of hyperplanes in $\mathbb{R}^2$ satisfies $h \geq 3$. On the other hand, we can never shatter *four* points. It follows from simple geometry that given any four points, there is always a set of labels such that we cannot realize the corresponding classification. Therefore, the VC dimension is $h = 3$. More generally, for hyperplanes in $\mathbb{R}^N$, the VC dimension can be shown to be $h = N + 1$. For a formal derivation of this result, as well as of other examples, see [523].

VC Dimension of Hyperplanes

How does this fit together with the fact that SVMs can be shown to correspond to hyperplanes in feature spaces of possibly infinite dimension? The crucial point is that SVMs correspond to *large margin* hyperplanes. Once the margin enters, the capacity can be much smaller than the above general VC dimension of hyperplanes. For simplicity, we consider the case of hyperplanes containing the origin.

VC Dimension of Margin Hyperplanes

Theorem 5.5 (Vapnik [559]) *Consider hyperplanes $\langle \mathbf{w}, \mathbf{x} \rangle = 0$, where $\mathbf{w}$ is normalized such that they are in canonical form w.r.t. a set of points $X^* = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$; i.e.,*

$$\min_{i=1,\ldots,r} |\langle \mathbf{w}, \mathbf{x}_i \rangle| = 1. \tag{5.50}$$

The set of decision functions $f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn} \langle \mathbf{x}, \mathbf{w} \rangle$ defined on X^*, and satisfying the constraint $\|\mathbf{w}\| \leq \Lambda$, has a VC dimension satisfying

$$h \leq R^2 \Lambda^2. \tag{5.51}$$

Here, R is the radius of the smallest sphere centered at the origin and containing X^*.

Before we give a proof, several remarks are in order.

- The theorem states that we can control the VC dimension *irrespective of the dimension of the space* by controlling the length of the weight vector $\|\mathbf{w}\|$. Note, however, that this needs to be done a priori, by choosing a value for Λ. It therefore does not strictly motivate what we will later see in SVMs, where $\|\mathbf{w}\|$ is minimized in order to control the capacity. Detailed treatments can be found in the work of Shawe-Taylor et al. [491, 24, 125].
- There exists a similar result for the case where R is the radius of the smallest sphere (not necessarily centered at the origin) enclosing the data, and where we allow for the possibility that the hyperplanes have a nonzero offset b [562]. In this case, we give a simple visualization in figure Figure 5.4, which shows it is plausible that enforcing a large margin amounts to reducing the VC dimension.
- Note that the theorem talks about functions defined on X^*. To extend it to the case where the functions are defined on all of the input domain $\mathcal{X}$, it is best to state it for the *fat shattering dimension*. For details, see [24].

The proof [24, 222, 559] is somewhat technical, and can be skipped if desired.

Proof Let us assume that $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are shattered by canonical hyperplanes with $\|\mathbf{w}\| \leq \Lambda$. Consequently, for all $y_1, \ldots, y_r \in \{\pm 1\}$, there exists a $\mathbf{w}$ with $\|\mathbf{w}\| \leq \Lambda$, such that

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \text{ for all } i = 1, \ldots, r. \tag{5.52}$$

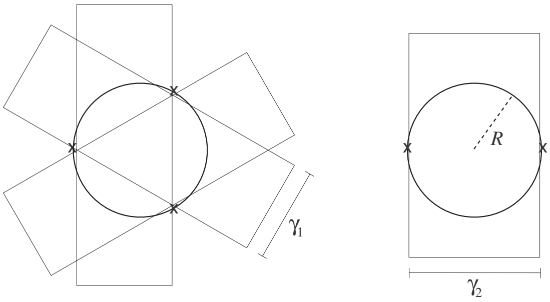


Figure 5.4 Simple visualization of the fact that enforcing a large margin of separation amounts to limiting the VC dimension. Assume that the data points are contained in a ball of radius R (cf. Theorem 5.5). Using hyperplanes with margin γ_1, it is possible to separate three points in all possible ways. Using hyperplanes with the larger margin γ_2, this is only possible for *two* points, hence the VC dimension in that case is two rather than three.

The proof proceeds in two steps. In the first part, we prove that the more points we want to shatter (5.52), the larger $\|\sum_{i=1}^r y_i\mathbf{x}_i\|$ must be. In the second part, we prove that we can upper bound the size of $\|\sum_{i=1}^r y_i\mathbf{x}_i\|$ in terms of R. Combining the two gives the desired condition, which tells us the maximum number of points we can shatter.

Summing (5.52) over $i = 1, \ldots, r$ yields

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^r y_i\mathbf{x}_i\right)\right\rangle \geq r. \tag{5.53}$$

By the Cauchy-Schwarz inequality, on the other hand, we have

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^r y_i\mathbf{x}_i\right)\right\rangle \leq \|\mathbf{w}\| \left\|\sum_{i=1}^r y_i\mathbf{x}_i\right\| \leq \Lambda \left\|\sum_{i=1}^r y_i\mathbf{x}_i\right\|. \tag{5.54}$$

Here, the second inequality follows from $\|\mathbf{w}\| \leq \Lambda$.

Combining (5.53) and (5.54), we get the desired lower bound,

$$\frac{r}{\Lambda} \leq \left\|\sum_{i=1}^r y_i\mathbf{x}_i\right\|. \tag{5.55}$$

We now move on to the second part. Let us consider independent random labels $y_i \in \{\pm 1\}$ which are uniformly distributed, sometimes called *Rademacher variables*. Let $\mathbf{E}$ denote the expectation over the choice of the labels. Exploiting the linearity of $\mathbf{E}$, we have

$$\begin{aligned}\mathbf{E}\left\|\sum_{i=1}^r y_i\mathbf{x}_i\right\|^2 &= \sum_{i=1}^r \mathbf{E}\left\langle y_i\mathbf{x}_i, \sum_{j=1}^r y_j\mathbf{x}_j\right\rangle \\ &= \sum_{i=1}^r \mathbf{E}\left\langle y_i\mathbf{x}_i, \left(\left(\sum_{j\neq i} y_j\mathbf{x}_j\right) + y_i\mathbf{x}_i\right)\right\rangle\end{aligned}$$

$$= \sum_{i=1}^{r} \left(\left(\sum_{j \neq i} \mathbf{E} \left\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \right\rangle \right) + \mathbf{E} \left\langle y_i \mathbf{x}_i, y_i \mathbf{x}_i \right\rangle \right)$$

$$= \sum_{i=1}^{r} \mathbf{E} \|y_i \mathbf{x}_i\|^2, \tag{5.56}$$

where the last equality follows from the fact that the Rademacher variables have zero mean and are independent. Exploiting the fact that $\|y_i \mathbf{x}_i\| = \|\mathbf{x}_i\| \leq R$, we get

$$\mathbf{E} \left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\|^2 \leq r R^2. \tag{5.57}$$

Since this is true for the expectation over the random choice of the labels, there must be at least one set of labels for which it also holds true. We have so far made no restrictions on the labels, hence we may now use this specific set of labels. This leads to the desired upper bound,

$$\left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\|^2 \leq r R^2. \tag{5.58}$$

Combining the upper bound with the lower bound (5.55), we get

$$\frac{r^2}{\Lambda^2} \leq r R^2; \tag{5.59}$$

hence,

$$r \leq R^2 \Lambda^2. \tag{5.60}$$

In other words, if the r points are shattered by a canonical hyperplane satisfying the assumptions we have made, then r is constrained by (5.60). The VC dimension h also satisfies (5.60), since it corresponds to the maximum number of points that can be shattered. ■

In the next section, we give an application of this theorem. Readers only interested in the theoretical background of learning theory may want to skip this section.

5.6 A Model Selection Example

In the following example, taken from [470], we use a bound of the form (5.36) to predict which kernel would perform best on a character recognition problem (USPS set, see Section A.1). Since the problem is essentially separable, we disregard the empirical risk term in the bound, and choose the parameters of a polynomial kernel by minimizing the second term. Note that the second term is a monotonic function of the capacity. As a capacity measure, we use the upper bound on the VC dimension described in Theorem 5.5, which in turn is an upper bound on the logarithm of the covering number that appears in (5.36) (by the arguments put forward in Section 5.5.6).

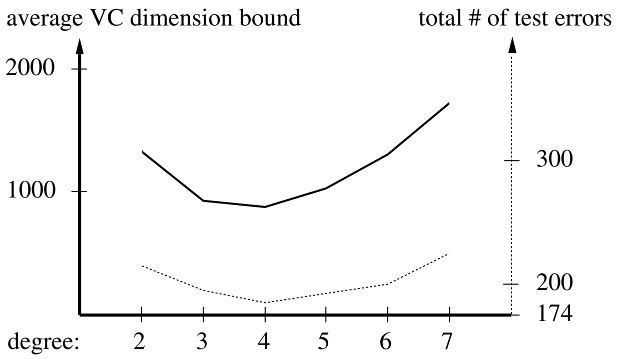


Figure 5.5 Average VC dimension (solid), and total number of test errors, of ten two-class-classifiers (dotted) with polynomial degrees 2 through 7, trained on the USPS set of handwritten digits. The baseline 174 on the error scale, corresponds to the total number of test errors of the ten *best* binary classifiers, chosen from degrees 2 through 7. The graph shows that for this problem, which can essentially be solved with zero training error for all degrees greater than 1, the VC dimension allows us to predict that degree 4 yields the best overall performance of the two-class-classifier on the test set (from [470, 467]).

We employ a version of Theorem 5.5, which uses the radius of the smallest sphere containing the data in a feature space $\mathcal{H}$ associated with the kernel k [561]. The radius was computed by solving a quadratic program [470, 85] (cf. Section 8.3). We formulate the problem as follows:

Computing the Enclosing Sphere in $\mathcal{H}$

$$\begin{aligned} \underset{R\geq 0,\mathbf{x}^*\in\mathcal{H}}{\text{minimize}} \quad & R^2 \\ \text{subject to} \quad & \|\mathbf{x}_i - \mathbf{x}^*\|^2 \leq R^2, \end{aligned} \tag{5.61}$$

where $\mathbf{x}^*$ is the center of the sphere, and is found in the course of the optimization. Employing the tools of constrained optimization, as briefly described in Chapter 1 (for details, see Chapter 6), we construct a Lagrangian,

$$R^2 - \sum_{i=1}^{m} \lambda_i (R^2 - (\mathbf{x}_i - \mathbf{x}^*)^2), \tag{5.62}$$

and compute the derivatives with respect to $\mathbf{x}^*$ and R, to get

$$\mathbf{x}^* = \sum_{i=1}^{m} \lambda_i \mathbf{x}_i, \tag{5.63}$$

and the Wolfe dual problem:

$$\underset{\boldsymbol{\lambda}\in\mathbb{R}^m}{\text{maximize}} \quad \sum_{i=1}^{m} \lambda_i \cdot \langle \mathbf{x}_i, \mathbf{x}_i \rangle - \sum_{i,j=1}^{m} \lambda_i \lambda_j \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \tag{5.64}$$

$$\text{subject to} \quad \sum_{i=1}^{m} \lambda_i = 1, \ \lambda_i \geq 0, \tag{5.65}$$

where $\boldsymbol{\lambda}$ is the vector of all Lagrange multipliers $\lambda_i, i = 1, \ldots, m$.

As in the Support Vector algorithm, this problem has the property that the $\mathbf{x}_i$

appear only in dot products, so we can again compute the dot products in feature space, replacing $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ by $k(x_i, x_j)$ (where the x_i belong to the input domain $\mathcal{X}$, and the $\mathbf{x}_i$ in the feature space $\mathcal{H}$).

As Figure 5.5 shows, the VC dimension bound, using the radius R computed in this way, gives a rather good prediction of the error on an independent test set.

5.7 Summary

In this chapter, we introduced the main ideas of statistical learning theory. For learning processes utilizing empirical risk minimization to be successful, we need a version of the law of large numbers that holds uniformly over all functions the learning machine can implement. For this uniform law to hold true, the capacity of the set of functions that the learning machine can implement has to be "well-behaved." We gave several capacity measures, such as the VC dimension, and illustrated how to derive bounds on the test error of a learning machine, in terms of the training error and the capacity. We have, moreover, shown how to bound the capacity of margin classifiers, a result which will later be used to motivate the Support Vector algorithm. Finally, we described an application in which a uniform convergence bound was used for model selection.

Whilst this discussion of learning theory should be sufficient to understand most of the present book, we will revisit learning theory at a later stage. In Chapter 12, we will present some more advanced material, which applies to kernel learning machines. Specifically, we will introduce another class of generalization error bound, building on a concept of *stability* of algorithms minimizing regularized risk functionals. These bounds are proven using concentration-of-measure inequalities, which are themselves generalizations of Chernoff and Hoeffding type bounds. In addition, we will discuss *leave-one-out* and *PAC-Bayesian* bounds.

5.8 Problems

5.1 (No Free Lunch in Kernel Choice ••) *Discuss the relationship between the "no-free-lunch Theorem" and the statement that there is no free lunch in kernel choice.*

5.2 (Error Counting Estimate [136] •) *Suppose you are given a test set with n elements to assess the accuracy of a trained classifier. Use the Chernoff bound to quantify the probability that the mean error on the test set differs from the true risk by more than $\epsilon > 0$. Argue that the test set should be as large as possible, in order to get a reliable estimate of the performance of a classifier.*

5.3 (The Tainted Die ••) *A con-artist wants to taint a die such that it does not generate any '6' when cast. Yet he does not know exactly how. So he devises the following scheme:*

he makes some changes and subsequently rolls the die 20 times to check that no '6' occurs. Unless pleased with the outcome, he changes more things and repeats the experiment.

How long will it take on average, until, even with a perfect die, he will be convinced that he has a die that never generates a '6'? What is the probability that this already happens at the first trial? Can you improve the strategy such that he can be sure the die is 'well' tainted (hint: longer trials provide increased confidence)?

5.4 (Chernoff Bound for the Deviation of Empirical Means ●●) *Use (5.6) and the triangle inequality to prove that*

$$\mathrm{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i\right| \geq \epsilon\right\} \leq 4\ \exp\left(-\frac{m\epsilon^2}{2}\right). \tag{5.66}$$

Next, note that the bound (5.66) is symmetric in how it deals with the two halves of the sample. Therefore, since the two events

$$\left\{\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i \geq \epsilon\right\} \tag{5.67}$$

and

$$\left\{\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i \leq -\epsilon\right\} \tag{5.68}$$

are disjoint, argue that (5.32) holds true. See also Corollary 6.34 below.

5.5 (Consistency and Uniform Convergence ●●) *Why can we not get a bound on the generalization error of a learning algorithm by applying (5.11) to the outcome of the algorithm? Argue that since we do not know in advance which function the learning algorithm returns, we need to consider the worst possible case, which leads to uniform convergence considerations.*

Speculate whether there could be restrictions on learning algorithms which imply that effectively, *empirical risk minimization only leads to a subset of the set of all possible functions. Argue that this amounts to restricting the capacity. Consider as an example neural networks with back-propagation: if the training algorithm always returns a local minimum close to the starting point in weight space, then the network effectively does not explore the whole weight (i.e., function) space.*

5.6 (Confidence Interval and Uniform Convergence ●) *Derive (5.36) from (5.35).*

5.7 (Representer Algorithms for Minimizing VC Bounds ○○○) *Construct kernel algorithms that are more closely aligned with VC bounds of the form (5.36). Hint: in the risk functional, replace the standard SV regularizer $\|\mathbf{w}\|^2$ with the second term of (5.36), bounding the shattering coefficient with the VC dimension bound (Theorem 5.5). Use the representer theorem (Section 4.2) to argue that the minimizer takes the form of a kernel expansion in terms of the training examples. Find the optimal expansion coefficients by minimizing the modified risk functional over the choice of expansion coefficients.*

5.8 (Bounds in Terms of the VC Dimension •) *From (5.35) and (5.36), derive bounds in terms of the growth function and the VC dimension, using the results of Section 5.5.6. Discuss the conditions under which they hold.*

5.9 (VC Theory and Decision Theory •••) *(i) Discuss the relationship between minimax estimation (cf. footnote 7 in Chapter 1) and VC theory. Argue that the VC bounds can be made "worst case" over distributions by picking suitable capacity measures. However, they only bound the difference between empirical risk and true risk, thus they are only "worst case" for the variance term, not for the bias (or empirical risk). The minimization of an upper bound on the risk of the form (5.36) as performed in SRM is done in order to construct an induction principle rather than to make a minimax statement. Finally, note that the minimization is done with respect to a structure on the set of functions, while in the minimax paradigm one takes the minimum directly over (all) functions.*

(ii) Discuss the following folklore statement: "VC statisticians do not care about doing the optimal thing, as long as they can guarantee how well they are doing. Bayesians do not care how well they are doing, as long as they are doing the optimal thing."

5.10 (Overfitting on the Test Set •••) *Consider a learning algorithm which has a free parameter C. Suppose you randomly pick n values $C_1, \ldots, C_n$, and for each n, you train your algorithm. At the end, you pick the value for C which did best on the test set. How would you expect your misjudgment of the true test error to scale with n?*

How does the situation change if the C_i are not picked randomly, but by some adaptive scheme which proposes new values of C by looking at how the previous ones did, and guessing which change of C would likely improve the performance on the test set?

5.11 (Overfitting the Leave-One-Out Error ••) *Explain how it is possible to overfit the leave-one-out error. I.e., consider a learning algorithm that minimizes the leave-one-out error, and argue that it is possible that this algorithm will overfit.*

5.12 (Learning Theory for Differential Equations ○○○) *Can you develop a statistical theory of estimating differential equations from data? How can one suitably restrict the "capacity" of differential equations?*

Note that without restrictions, already ordinary differential equations may exhibit behavior where the capacity is infinite, as exemplified by Rubel's universal differential equation [447]

$$\begin{aligned} 3y'^4y''y''''^2 - 4y'^4y'''^2y'''' + 6y'^3y''^2y'''y'''' + 24y'^2y''^4y'''' \\ -12y'^3y''y'''^3 - 29y'^2y''^3y'''^2 + 12y''^7 \quad = 0. \end{aligned} \tag{5.69}$$

Rubel proved that given any continuous function $f : \mathbb{R} \to \mathbb{R}$ and any positive continuous function $\varepsilon : \mathbb{R} \to \mathbb{R}^+$, there exists a C^∞ solution y of (5.69) such that $|y(t) - f(t)| < \varepsilon(t)$ for all $t \in \mathbb{R}$. Therefore, all continuous functions are uniform limits of sequences of solutions of (5.69). Moreover, y can be made to agree with f at a countable number of distinct points (t_i). Further references of interest to this problem include [61, 78, 63].