4

Generalisation Theory

The introduction of kernels greatly increases the expressive power of the learning machines while retaining the underlying linearity that will ensure that learning remains tractable. The increased flexibility, however, increases the risk of overfitting as the choice of separating hyperplane becomes increasingly ill-posed due to the number of degrees of freedom.

In Chapter 1 we made several references to the reliability of the statistical inferences inherent in the learning methodology. Successfully controlling the increased flexibility of kernel-induced feature spaces requires a sophisticated theory of generalisation, which is able to precisely describe which factors have to be controlled in the learning machine in order to guarantee good generalisation. Several learning theories exist that can be applied to this problem. The theory of Vapnik and Chervonenkis (VC) is the most appropriate to describe SVMs, and historically it has motivated them, but it is also possible to give a Bayesian interpretation, among others.

In this chapter we review the main results of VC theory that place reliable bounds on the generalisation of linear classifiers and hence indicate how to control the complexity of linear functions in kernel spaces. Also, we briefly review results from Bayesian statistics and compression schemes that can also be used to describe such systems and to suggest which parameters to control in order to improve generalisation.

4.1 Probably Approximately Correct Learning

The model we will now introduce is known under a number of different names depending on the discipline concerned. Within statistics it would be known as the study of rates of uniform convergence, or frequentist inference, but within computer science it is generally referred to as the probably approximately correct or *pac* model, although Vapnik and Chervonenkis applied this style of analysis to statistical inference many years before it became popular in machine learning. The reason for the name will become apparent when we describe the components of the model.

The key assumption on which the model is based is that the data used in training and testing are generated independently and identically (i.i.d.) according

to an unknown but fixed distribution $\mathscr{D}$. We assume this is a distribution over input/output pairings $(\mathbf{x}, y) \in X \times \{-1, 1\}$, an approach which subsumes the case where the output y is determined by a fixed *target* function t of the input $y = t(\mathbf{x})$. Adaptations of the model have considered the case where the distribution changes over time, or where there is not full independence in the generation of the examples in the training set, as might be expected in for instance a sequence of examples generated as a time series. The model also ignores the possibility that the learner might be able to influence which examples are chosen, an ingredient that is studied in the query model of learning. We will ignore all these refinements and consider only the i.i.d. case.

Since the test examples are generated according to the distribution $\mathscr{D}$, the natural measure of error in the classification case is the probability that a randomly generated example is misclassified. Again consideration can be made of unequal costs for misclassification of positive and negative examples, but this question will be ignored in our initial analysis. We therefore define the error $\mathrm{err}_{\mathscr{D}}(h)$ of a classification function h in distribution $\mathscr{D}$ to be

$$\underset{\mathscr{D}}{\mathrm{err}}(h) = \mathscr{D}\{(\mathbf{x}, y) : h(\mathbf{x}) \neq y\}.$$

Such a measure is also referred to as a *risk functional*, as its measure the expected error rate. The aim of the analysis will be to assert bounds on this error in terms of several quantities. Perhaps the most crucial is the number of training examples used. Frequently pac results have been presented as bounds on the number of examples required to obtain a particular level of error. This is also known as the *sample complexity* of the learning problem. We prefer bounding the error in terms of the number of examples as this error can then be used directly as a criterion for choosing between different classes, the so-called model selection problem.

Consider a fixed inference rule for selecting a hypothesis h_S from the class H of classification rules at the learner's disposal based on a set

$$S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$$

of ℓ training examples chosen i.i.d. according to $\mathscr{D}$. In this setting we can view the generalisation error $\mathrm{err}_{\mathscr{D}}(h_S)$ as a random variable depending on the random selection of the training set. The statistical mechanical approach to analysing generalisation aims to bound the expected generalisation error, where the expectation is taken over the random selection of training sets of a particular size. There are situations where such an estimate can be unreliable, as a particular error may fall far from its expectation. An example of such an unreliable estimator is that given by cross-validation. The pac model of learning requires a generalisation error bound that is unlikely to fail. It therefore bounds the tail of the distribution of the generalisation error random variable $\mathrm{err}_{\mathscr{D}}(h_S)$. The size of the tail is determined by a further parameter δ specified by the learner. Hence, a pac bound has the form $\varepsilon = \varepsilon(\ell, H, \delta)$, and asserts that with probability at least $1 - \delta$ over randomly generated training sets S, the generalisation error of the

selected hypothesis h_S will be bounded by

$$\operatorname*{err}_{\mathscr{D}}(h_S) \leq \varepsilon(\ell, H, \delta), \tag{4.1}$$

or in other words is *probably approximately correct (pac)*. This is equivalent to asserting that the probability that the training set gives rise to a hypothesis with large error is small:

$$\mathscr{D}^{\ell}\left\{S : \operatorname*{err}_{\mathscr{D}}(h_S) > \varepsilon(\ell, H, \delta)\right\} < \delta. \tag{4.2}$$

The pac approach has the flavour of a statistical test in the sense that it asserts that the probability the data have misled us is small. This corresponds to saying that a result is significant at the δ level, or in other words that the probability a test has misled us is at most δ. In this sense it provides a hypothesis with a statistical validation similar to those employed in developing the experimental sciences.

One of the key ingredients of the pac approach is that unlike many statistical tests the bound on the error should not depend on the distribution $\mathscr{D}$. This means that the bounds must hold whatever the distribution generating the examples, a property sometimes referred to as *distribution free*. It is not surprising that some distributions make learning harder than others, and so a theory that holds for all distributions must inevitably be pessimistic in many cases. We will see later that the large margin approach breaks this worst case deadlock and is able to take advantage of benign distributions. First, however, we will introduce the analysis of the distribution free case.

4.2 Vapnik Chervonenkis (VC) Theory

For a finite set of hypotheses it is not hard to obtain a bound in the form of inequality (4.1). Assume we use the inference rule that selects any hypothesis h that is consistent with the training examples in S. The probability that all ℓ of the independent examples are consistent with a hypothesis h for which $\operatorname{err}_{\mathscr{D}}(h) > \varepsilon$, is bounded by

$$\mathscr{D}^{\ell}\{S : h \text{ consistent and } \operatorname*{err}_{\mathscr{D}}(h) > \varepsilon\} \leq (1-\varepsilon)^{\ell} \leq \exp(-\varepsilon\ell),$$

where the second inequality is a simple mathematical bound. Now, even if we assume that all $|H|$ of the hypotheses have large error, the probability that one of them is consistent with S is at most

$$|H|\exp(-\varepsilon\ell),$$

by the union bound on the probability that one of several events occurs. This bounds the probability that a consistent hypothesis h_S has error greater than ε, as given in inequality (4.2),

$$\mathscr{D}^{\ell}\{S : h_S \text{ consistent and } \operatorname*{err}_{\mathscr{D}}(h_S) > \varepsilon\} < |H|\exp(-\varepsilon\ell).$$

In order to ensure the right hand side is less that δ, we set

$$\varepsilon = \varepsilon(\ell, H, \delta) = \frac{1}{\ell} \ln \frac{|H|}{\delta}.$$

This simple bound already shows how the complexity of the function class H measured here by its cardinality has a direct effect on the error bound. Clearly choosing H too large can lead to overfitting. The result also shows that a property relating the true error to the empirical error holds for all hypotheses in the set H. For this reason it is said to demonstrate *uniform convergence*. Learning theory relies on bounding the difference between empirical and true estimates of error uniformly over the set of hypotheses and conditions that can arise. As we have seen this is not difficult when the number of hypotheses is finite. The major contribution of the theory developed by Vapnik and Chervonenkis was to extend such an analysis to infinite sets of hypotheses, as for example in the case when we consider linear learning machines indexed by real-valued weight vectors.

We assume an inference rule that delivers any consistent hypothesis and denote by $\text{err}_S(h)$ the number of errors made by hypothesis h on the set S of examples. The key to bounding over an infinite set of functions is to bound the probability of inequality (4.2) by twice the probability of having zero error on the training examples but high error on a second random sample $\hat{S}$:

$$\begin{aligned} \mathscr{D}^{\ell} \left\{ S : \exists h \in H : \underset{S}{\text{err}}(h) = 0, \underset{\mathscr{D}}{\text{err}}(h) > \varepsilon \right\} \\ \leq 2\mathscr{D}^{2\ell} \left\{ S\hat{S} : \exists h \in H : \underset{S}{\text{err}}(h) = 0, \underset{\hat{S}}{\text{err}}(h) > \varepsilon\ell/2 \right\}. \end{aligned} \tag{4.3}$$

This relation follows from an application of Chernoff bounds provided $\ell > 2/\varepsilon$. The quantity on the right hand side is bounded by fixing the 2ℓ sample and counting different orders in which the points might have been chosen while still keeping all the errors in the second sample. Since each order or permutation is equally likely the fraction of orders that satisfy the property is an upper bound on its probability. By only considering permutations that swap corresponding points from the first and second sample, we can bound the fraction of such permutations by $2^{-\varepsilon\ell/2}$, independently of the particular set of 2ℓ sample points. The advantage of considering errors over a finite set of 2ℓ sample points is that the hypothesis space has effectively become finite, since there cannot be more than $2^{2\ell}$ classification functions on 2ℓ points. In order to obtain a union bound on the overall probability of the right hand side of inequality (4.3), all that is required is a bound on the size of the hypothesis space when restricted to 2ℓ points, a quantity known as the *growth function*

$$B_H(\ell) = \max_{(\mathbf{x}_1, \ldots, \mathbf{x}_\ell) \in X^\ell} |\{(h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_\ell)) : h \in H\}|,$$

where $|A|$ denotes the cardinality of the set A. The first observation about this quantity is that it cannot exceed 2^ℓ since the sets over which the maximum is

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

sought are all subsets of the set of binary sequences of length ℓ. A set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ for which the set

$$\{(h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_\ell)) : h \in H\} = \{-1, 1\}^\ell$$

is said to be *shattered* by H. If there are sets of any size which can be shattered then the growth function is equal to 2^ℓ for all ℓ. The final ingredient in the Vapnik Chervonenkis theory is an analysis of the case when there is a finite d which is the largest size of shattered set. In this case the growth function can be bounded as follows for $\ell \geq d$

$$B_H(\ell) \leq \left(\frac{e\ell}{d}\right)^d,$$

giving polynomial growth with exponent d. The value d is known as the Vapnik Chervonenkis (VC) dimension of the class H, denoted by VCdim(H). These quantities measure the richness or flexibility of the function class, something that is also often referred to as its capacity. Controlling the capacity of a learning system is one way of improving its generalisation accuracy. Putting the above bound on the growth function together with the observation about the fraction of permutations for which the first half of the sample is able to mislead the learner, we obtain the following bound on the left hand side of inequality (4.2):

$$\mathscr{D}^\ell \left\{ S : \exists h \in H : \operatorname*{err}_S(h) = 0, \operatorname*{err}_{\mathscr{D}}(h) > \varepsilon \right\} \leq 2 \left(\frac{2e\ell}{d}\right)^d 2^{-\varepsilon\ell/2},$$

resulting in a pac bound for any consistent hypothesis h of

$$\operatorname*{err}_{\mathscr{D}}(h) \leq \varepsilon(\ell, H, \delta) = \frac{2}{\ell}\left(d \log \frac{2e\ell}{d} + \log \frac{2}{\delta}\right),$$

where $d = \text{VCdim}(H)$. Hence we have shown the following *fundamental theorem of learning.*

Theorem 4.1 *(Vapnik and Chervonenkis) Let H be a hypothesis space having VC dimension d. For any probability distribution $\mathscr{D}$ on $X \times \{-1, 1\}$, with probability $1 - \delta$ over ℓ random examples S, any hypothesis $h \in H$ that is consistent with S has error no more than*

$$\operatorname*{err}_{\mathscr{D}}(h) \leq \varepsilon(\ell, H, \delta) = \frac{2}{\ell}\left(d \log \frac{2e\ell}{d} + \log \frac{2}{\delta}\right),$$

provided $d \leq \ell$ and $\ell > 2/\varepsilon$.

Remark 4.2 The theorem shows that for infinite sets of hypotheses the problem of overfitting is avoidable and the measure of complexity that should be used is precisely the VC dimension. The size of training set required to ensure good generalisation scales linearly with this quantity in the case of a consistent hypothesis.

VC theory provides a distribution free bound on the generalisation of a consistent hypothesis, but more than that it can be shown that the bound is in fact tight up to log factors, as the following theorem makes clear.

Theorem 4.3 *Let H be a hypothesis space with finite VC dimension $d \geq 1$. Then for any learning algorithm there exist distributions such that with probability at least δ over ℓ random examples, the error of the hypothesis h returned by the algorithm is at least*

$$\max\left(\frac{d-1}{32\ell}, \frac{1}{\ell}\ln\frac{1}{\delta}\right)$$

Remark 4.4 The theorem states that for a hypothesis class with high VC dimension there exist input probability distributions which will force the learner to require a large training set to obtain good generalisation. We can therefore see that finite VC dimension characterises learnability in the pac sense – we can bound the error as a function of a finite VCdim(H), while for unbounded VC dimension learning is impossible in the distribution free sense. Note, however, that the lower bound does not hold for all distributions. It is possible that a class with high VC dimension is learnable if the distribution is benign. Indeed this fact is essential for the performance of SVMs which are designed to take advantage of such benign distributions. This will be discussed further in the next section.

In order to apply the theory to linear learning machines, we must compute the VCdim($\mathscr{L}$) of a linear function class $\mathscr{L}$ in $\mathbb{R}^n$ in terms of the dimension n, that is determine what is the largest number d of examples that can be classified in all 2^d possible classifications by different linear functions, that is that can be shattered by $\mathscr{L}$. The following proposition characterises when this can be done.

Proposition 4.5 *Let $\mathscr{L}$ be the class of linear learning machines over $\mathbb{R}^n$.*

1. *Given any set S of $n+1$ training examples in general position (not lying in an $n-1$ dimensional affine subspace), there exists a function in $\mathscr{L}$ that consistently classifies S, whatever the labelling of the training points in S.*
2. *For any set of $\ell > n+1$ inputs there is at least one classification that cannot be realised by any function in $\mathscr{L}$.*

Theorem 4.3 and Proposition 4.5 imply that learning in very high dimensional feature spaces is not possible. An extreme example would be the use of a Gaussian kernel when we are effectively employing an infinite dimensional feature space. We must conclude that according to a distribution free pac analysis the Support Vector Machine approach to learning cannot succeed. The fact that SVMs can learn must therefore derive from the fact that the distribution generating the examples is not worst case as required for the lower bound of Theorem 4.3. In the next section we will sketch a more refined pac analysis that shows that the margin of a classifier provides a measure of how helpful the distribution is in

identifying the target concept, resulting in a generalisation error bound of the form

$$\operatorname*{err}_{\mathscr{D}}(h) \leq \varepsilon(\ell, \mathscr{L}, \delta, \gamma)$$

that will not involve the dimension of the feature space. Hence, the SVM learning strategy is able to exploit collusions between distribution and target concept when they occur, as is frequently the case in real-world learning problems. Bounds of this type that involve quantities measured as a result of the training process will be referred to as *data dependent*.

The theory we have sketched so far only applies when the hypothesis is consistent with the training data. If there is noise in the data or the class of hypotheses is unable to capture the full richness of the target function, it may not be possible or advisable to aim for full consistency. The theory can be adapted to allow for a number of errors on the training set by counting the permutations which leave no more errors on the left hand side. The resulting bound on the generalisation error is given in the following theorem.

Theorem 4.6 *Let H be a hypothesis space having VC dimension d. For any probability distribution $\mathscr{D}$ on $X \times \{-1, 1\}$, with probability $1-\delta$ over ℓ random examples S, any hypothesis $h \in H$ that makes k errors on the training set S has error no more than*

$$\operatorname*{err}_{\mathscr{D}}(h) \leq \varepsilon(\ell, H, \delta) = \frac{2k}{\ell} + \frac{4}{\ell}\left(d \log \frac{2e\ell}{d} + \log \frac{4}{\delta}\right),$$

provided $d \leq \ell$.

The theorem suggests that a learning algorithm for a hypothesis class H should seek to minimise the number of training errors, since everything else in the bound has been fixed by the choice of H. This inductive principle is known as empirical risk minimisation, since it seeks to minimise the empirically measured value of the risk functional. The theorem can also be applied to a nested sequence of hypothesis classes

$$H_1 \subset H_2 \subset \ldots \subset H_i \subset \ldots \subset H_M$$

by using δ/M, hence making the probability of any one of the bounds failing to hold to be less than δ. If a hypothesis h_i with minimum training error is sought in each class H_i, then the number of errors k_i that it makes on the fixed training set S will satisfy

$$k_1 \geq k_2 \geq \ldots \geq k_i \geq \ldots \geq k_M,$$

while the VC dimensions $d_i = \mathrm{VCdim}(H_i)$ form a non-decreasing sequence. The bound of Theorem 4.6 can be used to choose the hypothesis h_i for which the bound is minimal, that is the reduction in the number of errors (first term) outweighs the increase in capacity (second term). This induction strategy is known as structural risk minimisation.

4.3 Margin-Based Bounds on Generalisation

Recall that the definition of the margin of a classifier was given in Definition 2.2. We now generalise the definitions to an arbitrary class of real-valued functions.

Definition 4.7 Consider using a class $\mathscr{F}$ of real-valued functions on an input space X for classification by thresholding at 0. We define the *margin of an example* $(\mathbf{x}_i, y_i) \in X \times \{-1, 1\}$ *with respect to a function* $f \in \mathscr{F}$ to be the quantity

$$\gamma_i = y_i f(\mathbf{x}_i).$$

Note that $\gamma_i > 0$ implies correct classification of $(\mathbf{x}_i, y_i)$. The *margin distribution of* f *with respect to a training set* S is the distribution of the margins of the examples in S. We sometimes refer to the minimum of the margin distribution as the *margin* $m_S(f)$ *of* f *with respect to the training set* S. This quantity will be positive if f correctly classifies S. Finally, the *margin of a training set* S with respect to the class $\mathscr{F}$ is the maximum margin over all $f \in \mathscr{F}$.

The following three subsections will consider bounds involving different measures of the margin distribution

$$M_S(f) = \{\gamma_i = y_i f(\mathbf{x}_i) : i = 1, \ldots, \ell\},$$

over a training set $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ for the real-valued function f. If we are considering a linear function class we assume that the margins are geometric (see Definition 2.2), or in other words that the weight vector has unit norm. We begin by considering the margin $m_S(f)$ or $\min M_S(f)$.

4.3.1 Maximal Margin Bounds

When proving Theorem 4.1, we reduced the probability over an infinite set of hypotheses to the finite set of functions that can be realised on a 2ℓ sample. A large margin γ can reduce the effective size of the function space still further because the generalisation performance can be approximated by a function whose output is within $\gamma/2$ on the points of the double sample. In many cases the size of a set of functions that approximate the behaviour of the whole class to within $\gamma/2$ on a fixed set of ℓ points is much smaller than the size of the growth function of the thresholded class. The estimation of the size of such a representative sample of functions requires some extra machinery and notation.

Definition 4.8 Let $\mathscr{F}$ be a class of real-valued functions on a domain X. A γ-*cover* of $\mathscr{F}$ with respect to a sequence of inputs

$$S = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell)$$

is a finite set of functions B such that for all $f \in \mathscr{F}$, there exists $g \in B$, such that $\max_{1 \le i \le \ell} (|f(\mathbf{x}_i) - g(\mathbf{x}_i)|) < \gamma$. The size of the smallest such cover is denoted by $\mathscr{N}(\mathscr{F}, S, \gamma)$, while the *covering numbers* of $\mathscr{F}$ are the values

$$\mathscr{N}(\mathscr{F}, \ell, \gamma) = \max_{S \in X^\ell} \mathscr{N}(\mathscr{F}, S, \gamma).$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

We now show how for a hypothesis f with margin $m_S(f) = \gamma$ on the training set S, Theorem 4.1 can be reformulated in terms of the covering numbers of the underlying real-valued function class. We assume that there is a fixed threshold and that $\mathrm{err}_S(f)$ counts the number of errors the thresholded output of f makes on the sample S. Similarly, for $\mathrm{err}_{\mathscr{D}}(f)$ on a point generated randomly according to $\mathscr{D}$. We have

$$\begin{aligned} & \mathscr{D}^{\ell}\left\{S : \exists f \in \mathscr{F} : \underset{S}{\mathrm{err}}(f) = 0, m_S(f) \geq \gamma, \underset{\mathscr{D}}{\mathrm{err}}(f) > \varepsilon\right\} \\ \leq \; & 2\mathscr{D}^{2\ell}\left\{S\hat{S} : \exists f \in \mathscr{F} : \underset{S}{\mathrm{err}}(f) = 0, m_S(f) \geq \gamma, \underset{\hat{S}}{\mathrm{err}}(f) > \frac{\varepsilon\ell}{2}\right\}. \end{aligned} \tag{4.4}$$

Consider a $(\gamma/2)$-cover B of $\mathscr{F}$ with respect to the sequence $S\hat{S}$. Let $g \in B$, be within $\gamma/2$ of f. It follows that g has $\mathrm{err}_S(g) = 0, m_S(g) > \gamma/2$, while if f made an error on some point $\mathbf{x} \in \hat{S}$ then g must have margin less than $\gamma/2$ on $\mathbf{x}$. Hence, if $(\gamma/2)$-$\mathrm{err}_{\hat{S}}(g)$ denotes the number of points in $\hat{S}$ for which g has margin less that $\gamma/2$, we can bound the right hand side of inequality (4.4) by

$$\begin{aligned} & 2\mathscr{D}^{2\ell}\left\{S\hat{S} : \exists f \in \mathscr{F} : \underset{S}{\mathrm{err}}(f) = 0, m_S(f) \geq \gamma, \underset{\hat{S}}{\mathrm{err}}(f) > \varepsilon\ell/2\right\} \\ \leq \; & 2\mathscr{D}^{2\ell}\left\{S\hat{S} : \exists g \in B : \underset{S}{\mathrm{err}}(g) = 0, m_S(g) > \gamma/2, (\gamma/2)\text{-}\underset{\hat{S}}{\mathrm{err}}(g) > \varepsilon\ell/2\right\} \\ \leq \; & 2|B|2^{-\varepsilon\ell/2} \leq 2\mathscr{N}(\mathscr{F}, 2\ell, \gamma/2)2^{-\varepsilon\ell/2}, \end{aligned}$$

by a similar permutation argument and union bound. We have therefore demonstrated the following preliminary result.

Theorem 4.9 *Consider thresholding a real-valued function space $\mathscr{F}$ and fix $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathscr{D}$ on $X \times \{-1, 1\}$, with probability $1 - \delta$ over ℓ random examples S, any hypothesis $f \in \mathscr{F}$ that has margin $m_S(f) \geq \gamma$ on S has error no more than*

$$\underset{\mathscr{D}}{\mathrm{err}}(f) \leq \varepsilon(\ell, \mathscr{F}, \delta, \gamma) = \frac{2}{\ell}\left(\log \mathscr{N}(\mathscr{F}, 2\ell, \gamma/2) + \log\frac{2}{\delta}\right),$$

provided $\ell > 2/\varepsilon$.

Remark 4.10 The theorem shows how the generalisation error can be bounded in terms of the quantity $m_S(f)$ which is observed as a result of training. We expect that for larger values of γ, the size of $\log \mathscr{N}(\mathscr{F}, 2\ell, \gamma/2)$ will get smaller. This quantity can be viewed as an effective VC dimension and so we can expect that observing a large margin will result in good generalisation from a small sample. Notice that the VC dimension of $\mathscr{F}$ does not enter into the bound. We will see examples later where the VC dimension is in fact infinite, but this effective VC dimension is still finite and hence learning can take place. This does not contradict the lower bound of Theorem 4.3, as the observation of a large margin indicates that the distribution is benign. Even though the bound holds for all distributions it will only be useful when the distribution is benign.

Remark 4.11 The theorem only applies for a fixed value of γ specified before the learning began. In order to be able to use the theorem for the observed value of γ after training, we must apply the theorem for a range of values, ensuring that there will be one close to any realistic outcome of training. The fact that the δ enters into the bound inside a log factor makes it possible to perform such a uniform application over a finite set without appreciable loss in the quality of the bound. The precise details of how we make the choice of different values of γ and obtain a uniform result become rather technical, but add little to the main message. We therefore will not go further into this question (see Section 4.8 for pointers to papers with more detail), but rather turn our attention to how we can bound $\log \mathcal{N}(\mathcal{F}, 2\ell, \gamma/2)$, the critical quantity if we want to make use of the result.

The bound on $\log \mathcal{N}(\mathcal{F}, \ell, \gamma)$ represents a generalisation of the bound on the growth function required for the VC theory. In that case the critical measure was the VC dimension d and the growth function was shown to grow polynomially with degree d. The corresponding quantity we shall use to bound the covering numbers will be a real-valued generalisation of the VC dimension known as the fat-shattering dimension.

Definition 4.12 Let $\mathcal{F}$ be a class of real-valued functions defined on a domain X. We say a set of points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \in X^\ell$ is *γ-shattered* by $\mathcal{F}$, if there exist real numbers r_i, $i = 1, \ldots, \ell$, such that for every binary classification $\mathbf{b} \in \{-1, 1\}^\ell$, there exists $f_\mathbf{b} \in \mathcal{F}$, such that

$$f_\mathbf{b}(\mathbf{x}_i) \begin{cases} \geq r_i + \gamma, \text{ if } \mathbf{b}_i = 1, \\ < r_i - \gamma, \text{ if } \mathbf{b}_i = -1. \end{cases}$$

The *fat-shattering dimension* $\text{fat}_\mathcal{F}(\gamma)$ at scale γ is the size of the largest γ-shattered subset of X.

The dimension is also referred to as the scale-sensitive VC dimension. The real numbers r_i can be viewed as individual thresholds for each point while γ-shattering implies that we can realise every classification with margin γ relative to the chosen thresholds. Clearly the larger the value of γ, the smaller the size of set that can be shattered since the restrictions placed on the functions that can be used become stricter. If the thresholds r_i are required to be the same for all the points, the dimension is known as level fat-shattering. The freedom to choose individual thresholds appears to introduce extra flexibility, but in the case of linear functions it does not in fact increase the size of sets that can be shattered. We will return to the class of linear functions after considering the following bound on the covering numbers in terms of the fat-shattering dimension.

Lemma 4.13 *Let $\mathcal{F}$ be a class of functions $X \to [a, b]$ and $\mathcal{D}$ a distribution over X. Choose $0 < \gamma < 1$ and let $d = \text{fat}_\mathcal{F}(\gamma/4)$. Then for $\ell \geq d$*

$$\log \mathcal{N}(\mathcal{F}, \ell, \gamma) \leq 1 + d \log \frac{2e\ell(b-a)}{d\gamma} \log \frac{4\ell(b-a)^2}{\gamma^2}.$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

The bound on $\mathcal{N}(\mathscr{F}, \ell, \gamma)$ is therefore slightly greater than polynomial, but if we ignore the log factors the dependency of $\log \mathcal{N}(\mathscr{F}, \ell, \gamma)$ on the fat-shattering dimension of $\mathscr{F}$ exactly mimics the dependence of $\log B_H(\ell)$ on the VC dimension of H. We can therefore think of the fat-shattering dimension at scale $\gamma/8$ as the *effective* VC dimension when we observe a margin of γ. Indeed if we use Lemma 4.13 in Theorem 4.9 for a fixed value of γ, we obtain the following corollary.

Corollary 4.14 Consider thresholding a real-valued function space $\mathscr{F}$ with range $[-R, R]$ and fix $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathscr{D}$ on $X \times \{-1, 1\}$, with probability $1 - \delta$ over ℓ random examples S, any hypothesis $f \in \mathscr{F}$ that has margin $m_S(f) \geq \gamma$ on S has error no more than

$$\operatorname*{err}_{\mathscr{D}}(f) \leq \varepsilon(\ell, \mathscr{F}, \delta, \gamma) = \frac{2}{\ell}\left(d \log \frac{16e\ell R}{d\gamma} \log \frac{128\ell R^2}{\gamma^2} + \log \frac{4}{\delta}\right),$$

provided $\ell > 2/\varepsilon$, $d < \ell$, where $d = \operatorname{fat}_{\mathscr{F}}(\gamma/8)$.

Remark 4.15 Notice that if one ignores log factors, the role of the fat-shattering dimension in this bound is analogous to that of the VC dimension in Theorem 4.1, but the actual value of this quantity depends on the observed margin, hence the expression effective VC dimension.

Again, if we wish to take account of larger ranges and different values of γ extra technical analysis is necessary, but these details will not alter the overall shape of the result and will not be covered here. For more details see references in Section 4.8. We can view stratifying the result over different values of γ as assigning hypotheses to classes of different complexity depending on their margin. Hence, the classes are data dependent in contrast to classical structural risk minimisation when they must be specified before seeing the data. For this reason this type of result is sometimes referred to as data dependent structural risk minimisation.

We now turn our attention to the question of bounding the fat-shattering dimension for linear function classes, the final ingredient required if we wish to use the bounds for SVMs.

Theorem 4.16 *Suppose that X is the ball of radius R in an inner product space $\mathbb{H}$, $X = \{\mathbf{x} \in \mathbb{H} : \|\mathbf{x}\|_{\mathbb{H}} \leq R\}$, and consider the class of functions*

$$\mathscr{L} = \{\mathbf{x} \mapsto \langle \mathbf{w} \cdot \mathbf{x} \rangle : \|\mathbf{w}\|_{\mathbb{H}} \leq 1, \mathbf{x} \in X\}.$$

Then

$$\operatorname*{fat}_{\mathscr{L}}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2.$$

The proof of this theorem follows from two intermediate results. The first states that if $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ is γ-shattered by $\mathscr{L}$, then every subset $S_0 \subseteq S$ satisfies

$$\left\| \sum S_0 - \sum (S - S_0) \right\|_{\mathbb{H}} \geq \ell\gamma, \tag{4.5}$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

where the sum of a set means the sum of the vectors contained in that set. This result follows from considering the inner product of the vector inside the norm with the weight vector realising the classification determined by S_0 with margin γ. The second intermediate result computes the expected left hand side norm squared under a random choice of the subset S_0. If s is the $\{-1,1\}$ vector indicating membership in S_0, we choose s uniformly at random and must estimate

$$\begin{aligned} E\left\|\sum S_0 - \sum(S - S_0)\right\|_{\mathbb{H}}^2 &= E\left\|\sum_{i=1}^{\ell} s_i \mathbf{x}_i\right\|_{\mathbb{H}}^2 \\ &= E\sum_{i=1}^{\ell} s_i^2 \|\mathbf{x}_i\|_{\mathbb{H}}^2 + 2E\sum_{i\neq j} s_i s_j \langle \mathbf{x}_i \cdot \mathbf{x}_j\rangle_{\mathbb{H}} \\ &= E\sum_{i=1}^{\ell} \|\mathbf{x}_i\|_{\mathbb{H}}^2 \leq R^2\ell. \end{aligned}$$

Since there must be at least one S_0 with value less than or equal to the expectation, there exists at least one classification for which

$$\left\|\sum S_0 - \sum(S - S_0)\right\|_{\mathbb{H}} \leq R\sqrt{\ell}.$$

This inequality, together with inequality (4.5), shows that $R\sqrt{\ell} \geq \ell\gamma$, and the result follows.

Remark 4.17 Note that the bound on the fat-shattering dimension of linear learning machines is analogous to the mistake bound for the perceptron algorithm given in Theorem 2.3.

We are now in a position to quote an error bound for Support Vector Machines.

Theorem 4.18 *Consider thresholding real-valued linear functions $\mathscr{L}$ with unit weight vectors on an inner product space X and fix $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathscr{D}$ on $X \times \{-1,1\}$ with support in a ball of radius R around the origin, with probability $1-\delta$ over ℓ random examples S, any hypothesis $f \in \mathscr{L}$ that has margin $m_S(f) \geq \gamma$ on S has error no more than*

$$\operatorname*{err}_{\mathscr{D}}(f) \leq \varepsilon(\ell, \mathscr{L}, \delta, \gamma) = \frac{2}{\ell}\left(\frac{64R^2}{\gamma^2}\log\frac{e\ell\gamma}{4R}\log\frac{128\ell R^2}{\gamma^2} + \log\frac{4}{\delta}\right),$$

provided $\ell > 2/\varepsilon$ and $64R^2/\gamma^2 < \ell$.

The important qualitative aspect of this result is that the dimension of the input space does not appear, indeed the result also applies to infinite dimensional spaces. This type of result is sometimes said to be *dimension free*, as it suggests that the bound may overcome the curse of dimensionality. Based on the observations at the end of Section 4.2 we conclude that avoidance of the curse of

dimensionality will only be possible if the distribution generating the examples is sufficiently benign and renders the task of identifying the particular target function correspondingly easier. In such cases the bound gives an assurance that with high probability we will make few errors on randomly chosen test examples. It is in this sense that we can view γ as providing a measure of how benign the distribution is and therefore how well we can expect to generalise. It is also possible to give a more refined estimate of the probability of misclassification in terms of the distance of the test point from the hyperplane – see Section 4.8 for pointers to references.

Theorem 4.18 becomes trivial and hence gives no information for the case where the data are non-separable or noise in the data causes the margin to be very small. The next two subsections discuss two methods which can handle these situations by taking a different measure of the margin distribution.

4.3.2 Margin Percentile Bounds

The next measure of the distribution of margin values that can be used to bound generalisation is a general percentile. This measure has the significant advantage that it includes the case when a hypothesis is not fully consistent with the training data. If we order the values in the margin distribution

$$M_S(f) = \{\gamma_i = y_i f(\mathbf{x}_i)\}$$

so that $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_\ell$ and fix a number $k < \ell$, the k/ℓ percentile $M_{S,k}(f)$ of $M_S(f)$ is γ_k. The following theorem provides a bound on the generalisation error in terms of k/ℓ and $M_{S,k}(f)$.

Theorem 4.19 *Consider thresholding real-valued linear functions $\mathcal{L}$ with unit weight vectors on an inner product space X and fix $\gamma \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $X \times \{-1, 1\}$ with support in a ball of radius R around the origin, with probability $1-\delta$ over ℓ random examples S, any hypothesis $f \in \mathcal{L}$ has error no more than*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \frac{k}{\ell} + \sqrt{\frac{c}{\ell}\left(\frac{R^2}{M_{S,k}(f)^2}\log^2 \ell + \log\frac{1}{\delta}\right)},$$

for all $k < \ell$.

The proof of this theorem follows a similar pattern to that of Theorem 4.9 except that the counting of permutations of the double sample is complicated by the need to allow some mistakes on the left hand side.

The theorem suggests that we can obtain the best generalisation performance by minimising the number of margin errors, where we define a training point to be a *γ-margin error* if it has margin less than γ. The bound will be able to handle cases where there are a few outliers either causing the training set to be non-separable or forcing a very small margin. In these cases the margins of the difficult points can be ignored and the margin of the remaining points used. The

cost incurred by moving to this larger margin is two-fold. Firstly the extra term k/ℓ takes into account the fact that we have ignored that fraction of the training set, while the second cost is the additional square root of the complexity term as compared with Theorem 4.18. The next subsection describes a bound in terms of the margin distribution that does not include the square root but rather depends on a measure of the size of the margin errors.

4.3.3 Soft Margin Bounds

We begin this subsection by giving a precise definition of the margin slack variables. This generalises Definition 2.6 to general function classes. For the case where $\mathscr{F}$ is a class of linear functions the definition reduces back to Definition 2.6. The motivation is given by considering a target margin γ, and asking by how much individual points fail to meet this target. For points with margin larger than γ, this amount is zero, while for points that are misclassified the slack variable is greater than γ.

Definition 4.20 Consider using a class $\mathscr{F}$ of real-valued functions on an input space X for classification by thresholding at 0. We define the *margin slack variable* of an example $(\mathbf{x}_i, y_i) \in X \times \{-1, 1\}$ with respect to a function $f \in \mathscr{F}$ and target margin γ to be the quantity (see Figure 2.4)

$$\xi((\mathbf{x}_i, y_i), f, \gamma) = \xi_i = \max(0, \gamma - y_i f(\mathbf{x}_i)).$$

Note that $\xi_i > \gamma$ implies incorrect classification of $(\mathbf{x}_i, y_i)$. The *margin slack vector* $\boldsymbol{\xi}(S, f, \gamma)$ of a training set

$$S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$$

with respect to a function f and target margin γ contains the margin slack variables

$$\boldsymbol{\xi} = \boldsymbol{\xi}(S, f, \gamma) = (\xi_1, \ldots, \xi_\ell),$$

where the dependence on S, f, and γ is dropped when it is clear from the context.

We can think of the slack variables as measures of noise in the data which has caused individual training points to have smaller or even negative margins. The approach derived from taking account of the slack variables is suitable for handling noisy data.

We will derive a bound on the generalisation of the hypothesis f in terms of the target margin γ and different norms of the corresponding margin slack vector. The trick is to move the points that fail to meet the target margin outwards by embedding the input space into a larger space where a function can be found which increases their margins. The cost of performing the move can be measured by the norm of the margin slack vector. For an input space X, we

use the auxiliary inner product space

$$L(X) = \left\{ f \in \mathbb{R}^X : \text{supp}(f) \text{ is countable and } \sum_{\mathbf{x} \in \text{supp}(f)} f(\mathbf{x})^2 < \infty \right\},$$

where for $f, g \in L(X)$, the inner product is given by

$$\langle f \cdot g \rangle = \sum_{\mathbf{x} \in \text{supp}(f)} f(\mathbf{x}) g(\mathbf{x}).$$

We now embed the input space into the space $X \times L(X)$, using the mapping

$$\tau : \mathbf{x} \mapsto (\mathbf{x}, \delta_{\mathbf{x}})$$

where

$$\delta_{\mathbf{x}}(\mathbf{z}) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{z}, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, a general function in $g \in L(X)$ that maps input $\mathbf{x}_i$ to a value $c_i \in \mathbb{R}$, $i = 1, 2, \ldots$, can be written

$$g = \sum_{i=1}^{\infty} c_i \delta_{\mathbf{x}_i}.$$

Since $L(X)$ is an inner product space we can view elements of $L(X)$ as linear functions on $L(X)$. Hence, for a function $(f, g) \in \mathcal{F} \times L(X)$, we define the action of (f, g) on $(\mathbf{x}, \phi) \in X \times L(X)$ by

$$(f, g)(\mathbf{x}, \phi) = f(\mathbf{x}) + \langle g \cdot \phi \rangle,$$

so that the action of (f, g) on $\tau(\mathbf{x})$ is given by

$$(f, g)(\tau(\mathbf{x})) = f(\mathbf{x}) + \langle g \cdot \delta_{\mathbf{x}} \rangle.$$

Our strategy is to show that a judicious choice of $g \in L(X)$ causes the margin of the combined function to be γ, while the covering numbers of the expanded function class can be bounded in terms of the norm of the margin slack vector. We first show how to choose g so that the data are separated with margin γ. We define the following auxiliary function $g_f = g(S, f, \gamma) \in L(X)$:

$$g_f = \sum_{i=1}^{\ell} \xi_i y_i \delta_{\mathbf{x}_i}.$$

Now for $(\mathbf{x}_i, y_i) \in S$,

$$\begin{aligned} y_i(f, g_f)(\tau(\mathbf{x}_i)) &= y_i f(\mathbf{x}_i) + y_i \langle g_f \cdot \delta_{\mathbf{x}_i} \rangle \\ &= y_i f(\mathbf{x}_i) + y_i \sum_{j=1}^{\ell} \xi_j y_j \langle \delta_{\mathbf{x}_j} \cdot \delta_{\mathbf{x}_i} \rangle \\ &= y_i f(\mathbf{x}_i) + \xi_i y_i^2 \\ &= y_i f(\mathbf{x}_i) + \xi_i \geq \gamma. \end{aligned}$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

Hence, the augmented function does indeed have margin γ on the training set. For a point $\mathbf{x}$ not in the training set its action is identical to f,

$$\begin{aligned}(f, g_f)(\tau(\mathbf{x})) &= f(\mathbf{x}) + \sum_{j=1}^{\ell} \xi_j y_j \langle \delta_{\mathbf{x}_j} \cdot \delta_{\mathbf{x}} \rangle \\ &= f(\mathbf{x}),\end{aligned}$$

and so the generalisation performance of the two functions is identical. We therefore have the following result.

Theorem 4.21 *Consider thresholding a real-valued function space $\mathscr{F}$ and fix a subspace $L \subseteq L(X)$, and $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathscr{D}$ on $X \times \{-1, 1\}$, with probability $1 - \delta$ over ℓ random examples S, any hypothesis $f \in \mathscr{F}$ for which $g(S, f, \gamma) \in L$ has error no more than*

$$\operatorname*{err}_{\mathscr{D}}(f) \leq \varepsilon(\ell, \mathscr{F}, \delta, \gamma) = \frac{2}{\ell}\left(\log \mathscr{N}(\mathscr{F}, 2\ell, \gamma/4) + \log \mathscr{N}(L, 2\ell, \gamma/4) + \log \frac{2}{\delta}\right),$$

provided $\ell > 2/\varepsilon$ and there is no discrete probability on misclassified training points.

Proof By the above we have that $(f, g(S, f, \gamma))$ has margin γ on the training set and equals f for points outside the training set. We can therefore apply Theorem 4.9 to bound the generalisation error for points not in the training set. It remains to place an appropriate bound on $\log \mathscr{N}(\mathscr{F} \times L, 2\ell, \gamma/2)$. Let A be a cover of $\mathscr{F}$ and B a cover of L each at scale $\gamma/4$ with respect to the 2ℓ points $\mathbf{x}_1, \ldots, \mathbf{x}_{2\ell}$. Then $A \times B$ is a $\gamma/2$ cover of $\mathscr{F} \times L$ with respect to the same points, since for general $(f, g) \in \mathscr{F} \times L$, we can find $f' \in A$, such that

$$|f(\mathbf{x}_i) - f'(\mathbf{x}_i)| \leq \gamma/4,\ i = 1, \ldots, 2\ell,$$

and $g' \in B$, such that

$$|g(\delta_{\mathbf{x}_i}) - g'(\delta_{\mathbf{x}_i})| \leq \gamma/4,\ i = 1, \ldots, 2\ell,$$

whence we have that

$$\begin{aligned}|(f, g)(\tau(\mathbf{x}_i)) - (f', g')(\tau(\mathbf{x}_i))| &\leq |f(\mathbf{x}_i) - f'(\mathbf{x}_i)| + |g(\delta_{\mathbf{x}_i}) - g'(\delta_{\mathbf{x}_i})| \\ &\leq \gamma/2,\ i = 1, \ldots, 2\ell.\end{aligned}$$

We conclude that

$$\mathscr{N}(\mathscr{F} \times L, 2\ell, \gamma/2) \leq \mathscr{N}(\mathscr{F}, 2\ell, \gamma/4)\mathscr{N}(L, 2\ell, \gamma/4),$$

and the result follows. □

In order to apply this result we must choose appropriate sequences of subspaces of $L(X)$,

$$L_1 \subset L_2 \subset \ldots \subset L_k \subset \ldots \subset L(X),$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

and apply the theorem for each subspace, subsequently choosing the smallest L_k that contains $g(S,f,\gamma)$, for a given γ, S, and f. The two sequences that we will consider are those defined in terms of the 2-norm and 1-norm of the functions. In the case of the 2-norm, we have an inner produce space and can apply Theorem 4.16 and Lemma 4.13 to bound the covering numbers and obtain the following result.

Theorem 4.22 *Consider thresholding real-valued linear functions $\mathcal{L}$ with unit weight vectors on an inner product space X and fix $\gamma \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $X \times \{-1,1\}$ with support in a ball of radius R around the origin, with probability $1-\delta$ over ℓ random examples S, any hypothesis $f \in \mathcal{L}$ has error no more than*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{R^2 + \|\xi\|_2^2}{\gamma^2}\log^2 \ell + \log\frac{1}{\delta}\right),$$

where $\xi = \xi(f,S,\gamma)$ is the margin slack vector with respect to f and γ.

Remark 4.23 An analogous bound for the number of mistakes made in the first iteration of the perceptron algorithm is given in Theorem 2.7.

If the sequence L_k is defined in terms of the 1-norm then the bound obtained depends on this value together with an additional log factor.

Theorem 4.24 *Consider thresholding real-valued linear functions $\mathcal{L}$ with unit weight vectors on an inner product space X and fix $\gamma \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $X \times \{-1,1\}$ with support in a ball of radius R around the origin, with probability $1-\delta$ over ℓ random examples S, any hypothesis $f \in \mathcal{L}$ has error no more than*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{R^2 + \|\xi\|_1^2 \log(1/\gamma)}{\gamma^2}\log^2 \ell + \log\frac{1}{\delta}\right),$$

where $\xi = \xi(f,S,\gamma)$ is the margin slack vector with respect to f and γ.

The conclusion to be drawn from Theorems 4.22 and 4.24 is that the generalisation error bound takes into account the amount by which points fail to meet a target margin γ. The bound is in terms of a norm of the slack variable vector suggesting that this quantity should be minimised in order to optimise performance. The bound does not rely on the training points being linearly separable and hence can also handle the case when the data are corrupted by noise or the function class cannot capture the full complexity of the decision rule. Optimising the norm of the margin slack vector does not necessarily mean minimising the number of misclassifications. Hence, the inductive principle suggested by the theorems does not correspond to empirical risk minimisation. This fact will be important, as we shall see that minimising the number of misclassifications

appears to be computationally more demanding than optimising the norms of the margin slack vector.

Optimising the norms of the margin slack vector has a diffuse effect on the margin. For this reason it is referred to as a *soft margin* in contrast to the maximal margin, which depends critically on a small subset of points and is therefore often called a *hard margin*. We will refer to the bound in terms of the 2-norm of the margin slack vector as the 2-norm soft margin bound, and similarly for the 1-norm soft margin.

4.4 Other Bounds on Generalisation and Luckiness

The previous section considered bounds on generalisation performance in terms of measures of the margin distribution. We argued that the bounds we must use in high dimensional spaces must be able to take advantage of favourable input distributions that are in some sense aligned with the target function. The bounds must avoid dependence on the dimension of the input space in favour of dependence on quantities measured as a result of the training algorithm, quantities that effectively assess how favourable the input distribution is. We described three results showing dependences on three different measures of the margin distribution. We will briefly argue in this section that bounds of this type do not necessarily have to depend on margin value. In particular the size of a sample compression scheme can be used to bound the generalisation by a relatively straightforward argument due to Littlestone and Warmuth. A sample compression scheme is defined by a fixed rule

$$\rho : S \mapsto \rho(S)$$

for constructing a classifier from a given set of labelled data. Given a large training set, it is compressed by finding the smallest subset (the compression set) $\hat{S} \subseteq S$ for which the reconstructed classifier $\rho(\hat{S})$ correctly classifies the whole set S. Fix a number $d < \ell$. Suppose that for a particular training set we obtain a compressed set of size d. This can only occur in $\binom{\ell}{d}$ ways. For each such choice the probability that the resulting hypothesis has error more than ε, and yet correctly classifies the remaining $\ell - d$ randomly generated training points, is bounded by

$$(1-\varepsilon)^{\ell-d} \leq \exp(-\varepsilon(\ell-d)).$$

Hence, the probability that a compression set of size d has error greater than ε_d can be bounded by

$$\binom{\ell}{d} \exp(-\varepsilon_d(\ell-d)). \tag{4.6}$$

For

$$\varepsilon_d = \frac{1}{\ell-d}\left(d \ln \frac{e\ell}{d} + \ln \frac{\ell}{\delta}\right),$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

this will be less than δ/ℓ. It follows that the probability of ε_d failing to bound the generalisation error for a compression set of size d is less than δ/ℓ and so the probability that the ε corresponding to the observed size of the compression set is greater than the generalisation error is at most δ. Hence, we have shown the following theorem.

Theorem 4.25 *Consider a compression scheme ρ. For any probability distribution $\mathcal{D}$ on $X \times \{-1, 1\}$, with probability $1-\delta$ over ℓ random examples S, the hypothesis defined by a compression set of size d has error no more than*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \frac{1}{\ell - d}\left(d \log \frac{e\ell}{d} + \log \frac{\ell}{\delta}\right).$$

We will see in Chapter 6 that the support vectors of a Support Vector Machine form a compression scheme that can reconstruct the maximal margin hyperplane. Theorem 4.25 therefore shows that the number of support vectors, a quantity that does not directly involve the margin, can also be used to measure how favourably the distribution generating the data is aligned with the target function, so giving another data dependent bound. This observation has led to the introduction of a general framework which uses a so-called *luckiness function* to assess how well the distribution is aligned with the target function. The size of the margin is just one such measure. Choosing a luckiness function corresponds to asserting a prior belief about the type of relations that may occur between distributions and target functions. If that belief holds true we profit by a correspondingly improved generalisation, though of course there may be a small cost involved when the assumption fails.

4.5 Generalisation for Regression

The problem of regression is that of finding a function which approximates mapping from an input domain to the real numbers based on a training sample. The fact that the output values are no longer binary means that the mismatch between the hypothesis output and its training value will no longer be discrete. We refer to the difference between the two values as the *residual* of the output, an indication of the accuracy of the fit at this point. We must decide how to measure the importance of this accuracy, as small residuals may be inevitable while we wish to avoid large ones. The *loss function* determines this measure. Each choice of loss function will result in a different overall strategy for performing regression. For example least squares regression uses the sum of the squares of the residuals.

Although several different approaches are possible (see Section 4.8), we will provide an analysis for the generalisation performance of a regression function by using the bounds obtained for the classification case, as these will motivate the algorithms described in Chapter 6. Hence, we will consider a threshold test accuracy θ, beyond which we consider a mistake to have been made. We therefore aim to provide a bound on the probability that a randomly drawn test

point will have accuracy less than θ. If we assess the training set performance using the same θ, we are effectively using the real-valued regressors as classifiers and the worst case lower bounds apply. What we must do in order to make use of the dimension free bounds is to allow a margin in the regression accuracy that corresponds to the margin of a classifier. We will use the same symbol γ to denote this margin, which measures the amount by which the training and test accuracy can differ. It should be emphasised that we are therefore using a different loss function during training and testing, where γ measures the discrepancy between the two losses, implying that a training point counts as a mistake if its accuracy is less than $\theta - \gamma$. One way of visualising this method of assessing performance is to consider a band of size $\pm(\theta - \gamma)$ around the hypothesis function. Any training points lying outside this band are considered to be training mistakes. Test points count as mistakes only if they lie outside the wider band of $\pm\theta$. Using this correspondence Theorem 4.18 has the following interpretation for regression estimation.

Theorem 4.26 *Consider performing regression estimation with linear functions $\mathcal{L}$ with unit weight vectors on an inner product space X and fix $\gamma \leq \theta \in \mathbb{R}^+$. For any probability distribution $\mathcal{D}$ on $X \times \mathbb{R}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over ℓ random examples S, any hypothesis $f \in \mathcal{L}$, whose output is within $\theta - \gamma$ of the training value for all of the training set S, has residual greater than θ on a randomly drawn test point with probability at most*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \varepsilon(\ell, \mathcal{L}, \delta, \gamma) = \frac{2}{\ell}\left(\frac{64R^2}{\gamma^2}\log\frac{e\ell\gamma}{4R}\log\frac{128\ell R^2}{\gamma^2} + \log\frac{4}{\delta}\right),$$

provided $\ell > 2/\varepsilon$ and $64R^2/\gamma^2 < \ell$.

The theorem shows how we can bound the probability that the output of a unit norm linear function on a random test point will be out by more than θ provided its residuals on the training set are all smaller than $\theta - \gamma$. Note that as in the classification case the dimension of the feature space does not enter into the formula, ensuring that the bound will be applicable even for the high dimensional feature spaces arising from the use of kernel functions.

We next consider the role that the margin slack variables play in the regression case.

Definition 4.27 Consider using a class $\mathcal{F}$ of real-valued functions on an input space X for regression. We define the *margin slack variable* of an example $(\mathbf{x}_i, y_i) \in X \times \mathbb{R}$ with respect to a function $f \in \mathcal{F}$, target accuracy θ and loss margin γ to be the quantity (see Figure 4.1 for a linear example and Figure 4.2 for a non-linear function)

$$\xi((\mathbf{x}_i, y_i), f, \theta, \gamma) = \xi_i = \max\left(0,\ |y_i - f(\mathbf{x}_i)| - (\theta - \gamma)\right).$$

Note that $\xi_i > \gamma$ implies an error on $(\mathbf{x}_i, y_i)$ of more than θ. The *margin slack vector* $\boldsymbol{\xi}(S, f, \theta, \gamma)$ of a training set

$$S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$$

https://doi.org/10.1017/CBO9780511801389.006 Published online by Cambridge University Press

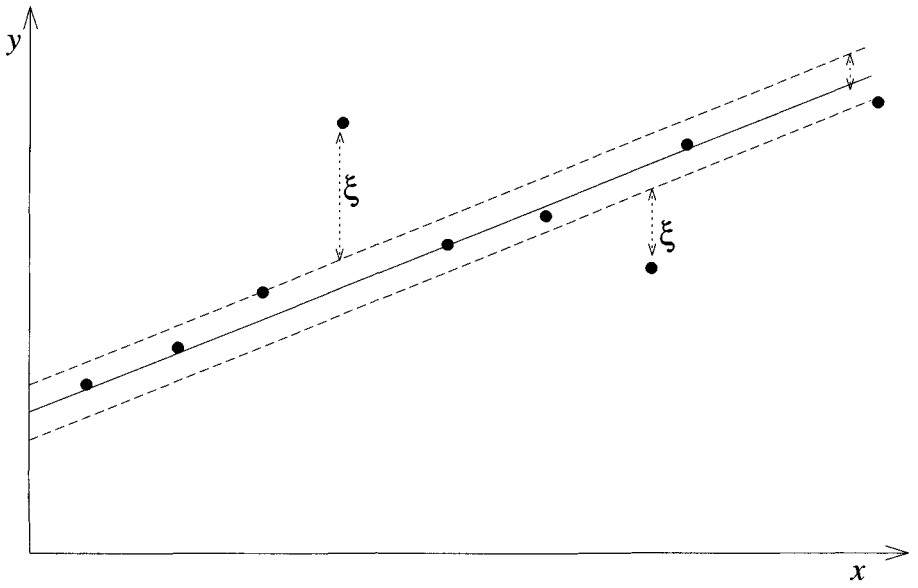

Figure 4.1: The slack variables for a one dimensional linear regression problem

with respect to a function f and target accuracy θ and loss margin γ contains the margin slack variables

$$\xi = \xi(S, f, \gamma) = (\xi_1, \ldots, \xi_\ell),$$

where the dependence on S, f, θ, and γ is dropped when it is clear from the context.

Again the correspondence with the classification case is direct and leads to the following bounds on the generalisation performance of linear regressors in terms of the 2-norm of the slack variables. Note that in the case of regression it no longer makes sense to fix the norm of the weight vector, as in contrast to the classification case rescaling does result in a different functionality. The size of the norm of the weight vector affects the scale at which the covering must be sought for the equivalent unit weight vector of a linear function.

Theorem 4.28 *Consider performing regression with linear functions $\mathcal{L}$ on an inner product space X and fix $\gamma \leq \theta \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $X \times \mathbb{R}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over ℓ random examples S, the probability that a hypothesis $\mathbf{w} \in \mathcal{L}$ has output more than θ away from its true value is bounded by*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{\|\mathbf{w}\|_2^2 R^2 + \|\xi\|_2^2}{\gamma^2}\log^2 \ell + \log\frac{1}{\delta}\right),$$

where $\xi = \xi(\mathbf{w}, S, \theta, \gamma)$ is the margin slack vector with respect to $\mathbf{w}$, θ, and γ.

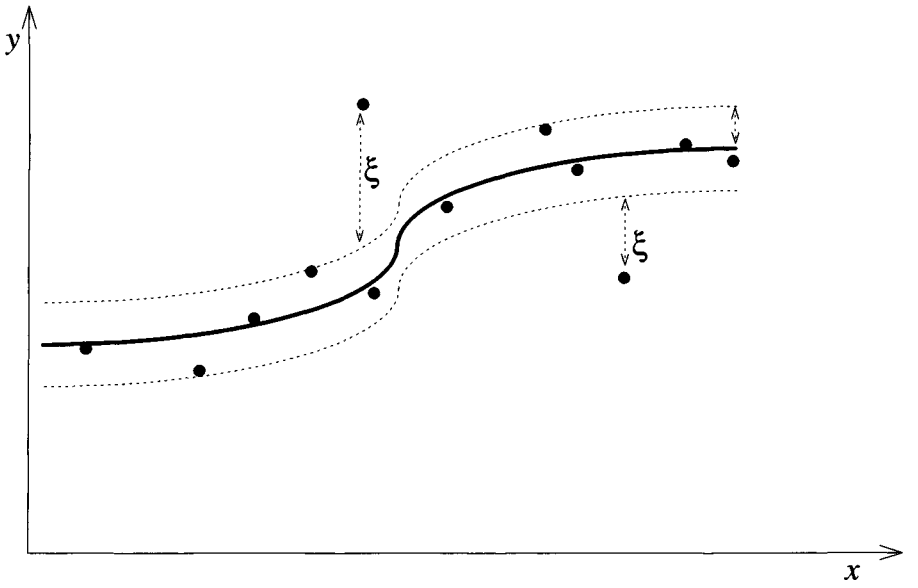

Figure 4.2: The slack variables for a non-linear regression function

The theorem is a significant advance over Theorem 4.26 since it can be applied to all linear functions and can take account of training points whose residuals lie outside the $\theta - \gamma$ tube. The 2-norm of these excess residuals enters into the formula together with the 2-norm of the linear function. If we consider the case when $\gamma = \theta$, the 2-norm of the margin slack vector is the sum of the squares of the residuals on the training sequence, something often referred to as the *sum squared error (SSE)*. We therefore have the following corollary.

Corollary 4.29 Consider performing regression with linear functions $\mathscr{L}$ on an inner product space X and fix $\theta \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathscr{D}$ on $X \times \mathbb{R}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over ℓ random examples S, the probability that a hypothesis $\mathbf{w} \in \mathscr{L}$ has output more than θ away from its true value is bounded by

$$\operatorname*{err}_{\mathscr{D}}(f) \leq \frac{c}{\ell}\left(\frac{\|\mathbf{w}\|_2^2 R^2 + SSE}{\theta^2}\log^2 \ell + \log\frac{1}{\delta}\right),$$

where SSE is the sum squared error of the function $\mathbf{w}$ on the training set S.

The corollary is directly applicable to least squares regression using linear functions, perhaps the most standard form of regression, but here used in a setting where the training sequence has been generated according to an unknown probability distribution. The resulting bound on the probability that a test

point has residual greater than θ is a novel way of assessing performance of such functions. In Chapter 6 we will see that the ridge regression algorithm discussed in Subsection 2.2.2 directly optimises this bound and hence potentially outperforms the standard least squares algorithm.

Finally, we can translate the 1-norm bound of Theorem 4.24 to obtain the following result.

Theorem 4.30 *Consider performing regression with linear functions $\mathcal{L}$ on an inner product space X and fix $\gamma \leq \theta \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $X \times \mathbb{R}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over ℓ random examples S, the probability that a hypothesis $\mathbf{w} \in \mathcal{L}$ has output more than θ away from its true value is bounded by*

$$\operatorname*{err}_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{\|\mathbf{w}\|_2^2 R^2 + \|\boldsymbol{\xi}\|_1^2 \log(1/\gamma)}{\gamma^2} \log^2 \ell + \log \frac{1}{\delta}\right),$$

where $\boldsymbol{\xi} = \boldsymbol{\xi}(\mathbf{w}, S, \theta, \gamma)$ is the margin slack vector with respect to $\mathbf{w}$, θ, and γ.

The bound in terms of the 1-norm of the slacks and 2-norm of the weight vector may seem an unnatural mix of two different norms. However, the use of 2-norm for the linear function is dictated by our prior over the function class, while the norm of the slack variables should be chosen to model the type of noise that has corrupted the training examples. For example, if we optimise the 1-norm bound the resulting regressor takes less account of points that have large residuals and so can handle outliers better than by optimising the 2-norm bound.

4.6 Bayesian Analysis of Learning

In this section we will briefly review the Bayesian approach to learning. Its motivation does not come from bounds on the generalisation and so the section may seem out of place in this chapter. Though Bayesian analysis can be used to estimate generalisation, in this section we only cover that part of the theory needed to motivate the learning strategy. The pac style of analysis we have considered in the earlier sections of this chapter has focused on finding an error bound that will hold with high probability. That approach can be seen as conservative in that it attempts to find a bound on the error probability that will hold with high confidence. The Bayesian approach in contrast attempts to choose output values that are most likely based on the observed training values. This can result in a function not actually in the initial set of hypotheses. Alternatively if we restrict ourselves to choosing a function from the set, it can motivate an optimal choice. It therefore is attempting to make the best possible choice based on the available data. In order to make such a desirable calculation tractable a number of assumptions need to be made including the existence of a prior distribution over the set of hypotheses and a (Gaussian) noise model. These assumptions can render the choice of function less reliable when

compared with the pac analysis, which depends only on the assumed existence of an underlying probability distribution that generates the training and test data independently. Nonetheless we shall see that despite its very different starting point the resulting computation and function are very closely related to the support vector approach.

As described in Chapter 3 the prior distribution can be described by a Gaussian process and choosing its covariance function defines the prior in a similar way to that in which choosing the kernel for a Support Vector Machine defines the feature space. The aim of Bayesian analysis is to update that distribution based on the observation of the particular data. The more a particular data point disagrees with a function the more the posterior probability of that function is reduced. Bayes' formula,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

is used to calculate this altered probability under the assumption that the errors between the true function and the observed outputs are generated by a Gaussian distribution with mean 0 and variance σ^2. Once the updated or a posteriori distribution based on all the observations has been calculated, the learner can either pick the function with highest likelihood or take an average of the outputs over the different functions weighted according to the posterior distribution. The function with highest likelihood is often referred to as the maximum a posteriori or *MAP* hypothesis. The second more powerful approach corresponds to choosing the most likely output given a test input but with no restrictions placed on what function is used. In fact the function will always be in the convex closure of the original set of hypotheses, since it is an average of those functions weighted according the a posteriori distribution. In the case of Gaussian processes the two approaches coincide since a linear class is equal to its own convex closure and the a posteriori distribution is Gaussian.

We restrict consideration to regression since the assumption of a Gaussian distribution of errors is only meaningful in this case. In order to be consistent with our previous notation we will use y_i, $i = 1, \ldots, \ell$, to denote the output values from the training set that are assumed to be corrupted by noise. The underlying true target output values will be denoted by t_i, $i = 1, \ldots, \ell$. Unfortunately, this notation is the reverse of that adopted in some of the literature. The relation between the vectors $\mathbf{y}$ and $\mathbf{t}$ is assumed to be Gaussian,

$$P(\mathbf{y}|\mathbf{t}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{t})'\Omega^{-1}(\mathbf{y}-\mathbf{t})\right],$$

where $\Omega = \sigma^2\mathbf{I}$. As mentioned above, the aim of the Bayesian analysis is to calculate the probability distribution for the true output t given a novel input $\mathbf{x}$, and a training set S, which we split into the matrix $\mathbf{X}$ of input vectors and vector $\mathbf{y}$ of corresponding outputs. Hence, we wish to estimate $P(t|\mathbf{x}, S)$ and in

particular find where it achieves its maximum. First we use Bayes' formula to express

$$P(t, \mathbf{t}|\mathbf{x}, S) = P(t, \mathbf{t}|\mathbf{y}, \mathbf{x}, \mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{t}, \mathbf{x}, \mathbf{X})P(t, \mathbf{t}|\mathbf{x}, \mathbf{X})}{P(\mathbf{y}|\mathbf{x}, \mathbf{X})}$$
$$= \frac{P(\mathbf{y}|\mathbf{t})P(t, \mathbf{t}|\mathbf{x}, \mathbf{X})}{P(\mathbf{y}|\mathbf{x}, \mathbf{X})} \propto P(\mathbf{y}|\mathbf{t})P(t, \mathbf{t}|\mathbf{x}, \mathbf{X}),$$

where we have treated the denominator as a constant since it does not depend on our choice of hypothesis and hence does not vary once the training set and test point are known. The second factor in the final expression is the prior distribution for the true outputs given a set of inputs with no knowledge about the output values contained in the training set. This will be determined by the choice of Gaussian process through a covariance function. The first factor is the weighting given to a particular hypothesis identified by its output values on the training set inputs. The weighting is based entirely on the discrepancy between the hypothesis outputs and the values given in the training set. The task remaining for the Bayesian learner is to 'marginalise' over $\mathbf{t}$, by which is meant summing the probability of a particular value for t over all possible values that the parameters $\mathbf{t}$ might take. The advantage of using Gaussian distributions is that the resulting distribution is also Gaussian and that its mean and variance can be calculated analytically, hence giving the maximum value for t together with what can be regarded as an error bar on the accuracy of the prediction. Chapter 6 will describe this calculation and compare the resulting decision rule with that obtained using Support Vector Machines.

4.7 Exercises

1. Prove Proposition 4.5.
2. Describe how Theorem 4.9 could be made to apply for all values of γ, indicating what weakening of the bound would result.
3. In Section 4.4 consider choosing ε_d so that formula (4.6) is less than δ_d for some values δ_d satisfying

$$\sum_{i=1}^{\ell} \delta_i = 1.$$

Show that a generalisation of Theorem 4.25 results. Given that we know that the probability of obtaining d support vectors is p_d, what choice of δ_d will give the best expected bound in the generalised theorem?

4.8 Further Reading and Advanced Topics

The analysis of generalisation based on the VC dimension was developed by Vapnik and Chervonenkis starting from the mid 1960s [162]. It has been

successfully applied in different fields, motivating for example the development in statistics of a uniform law of large numbers [117]. Most of its basic theoretical results were already present in Vapnik's book [157]. The development of the pac model of machine learning, on the other hand, can be traced to the seminal paper by Valiant in 1984 [155], which laid the foundations of what became known as computational learning theory: a large set of models describing, among others, on-line learning, query learning, unsupervised and supervised learning, and recently also applied to reinforcement learning. The introduction of the VC results within this theory, together with the lower bounds, is contained in the landmark paper by Blumer et al. [18], and has greatly influenced the entire field of machine learning. VC theory has since been used to analyse the performance of learning systems as diverse as decision trees, neural networks, and others; many learning heuristics and principles used in practical applications of machine learning have been explained in terms of VC theory.

Many introductory books to computational learning theory exist, for example [6], [71], although they are often mainly focused on the pac model, somewhat neglecting the rich fields of on-line, query, and unsupervised learning. A good introduction to the statistical analysis of pattern recognition is given by Devroye et al. [33]. VC theory has recently also come to be known as statistical learning theory, and is extensively described in the recent book of Vapnik [159], and in other books that preceded it [157] [158], as well as earlier papers by Vapnik and Chervonenkis [162][164][165]. An easy introduction to the theory is provided by [169], and a complete account, including also very recent results, can be found in [5]. An international conference is held every year on computational learning theory, known as the COLT conference, where new results are presented. Websites such as **www.neurocolt.org** offer large repositories of recent papers.

The question of how the margin might affect generalisation was raised by many researchers including Duda and Hart [35], Vapnik and Chervonenkis [166], and Mangasarian. Vapnik and Chervonenkis [163] obtained bounds for the case when the margin is measured on the combined training and test sets using a quantity analogous to the fat-shattering dimension. The fat-shattering dimension itself (sometimes also called the scale-sensitive dimension, or V_γ dimension) has appeared implicitly in a number of earlier references but was introduced into computational learning theory by [72] and shown by Alon et al. [2] to characterise the so-called Glivenko Cantelli classes. The fat-shattering dimension of linear classifiers was obtained by different authors ([54] and [138]), while the proof presented in this chapter appears in [9].

The first papers to prove the large margin results were [138], [10] with the second reference containing the percentile result of Subsection 4.3.2. The soft margin bounds involving the margin slack vector for both classification and regression are due to [141], [142], [139], [140] and use techniques similar to those discussed in Chapter 2 for obtaining mistake bounds in the non-separable case (see Section 2.5 for further references). Those results are summarised in [9] and [149]. A quantity related to the margin slack vector is the so-called 'hinge loss', used to obtain mistake bounds in the on-line learning framework [48]. Margin analysis has since been applied to describe systems like Adaboost [127], Bayesian

classifiers [32], decision trees [12], neural networks [10] , and is now a standard tool in machine learning. The margin analysis has also been extended to take account of the margin of the test example [137].

Anthony and Bartlett used the fat-shattering dimension to obtain results for regression similar Theorem 4.26. Different analyses of generalisation are possible for regression, such as in [159]. The book [5] provides an excellent introduction to the analysis of regression.

The reason why margin analysis requires different tools from VC theory is that the quantity used to characterise the richness of a hypothesis class, the margin, depends on the data. Only after training the learning machine can one know what is the complexity of the resulting hypothesis. This style of analysis, which provides a way of exploiting benign collusions between the target function and input distribution, is often called data dependent analysis, or data dependent structural risk minimisation. The first data dependent result was Theorem 4.25 on the generalisation power of compression schemes and which is due to Littlestone and Warmuth [79][42], while the paper [138] introduced the general luckiness framework mentioned in Section 4.4. Other data dependent results include micro-choice algorithms, and pac-Bayesian bounds [93][94]. More recent bounds include [133] and [37], who like [138] bring out the connection between classification and regression.

Bayesian analysis is a traditional field within statistics, and has been applied to pattern recognition for a long time [35]. In recent years, a new surge of interest in Bayesian analysis has come from the neural networks community, mainly thanks to the work of MacKay [82]. An introduction to such analysis is provided by the books of Bishop [16] and Neal [102]. More recently, attention has been directed at Gaussian processes, a standard tool in statistics, described in [120] and [180]. We will return to the subject of Gaussian processes in Chapter 6. A Bayesian analysis of generalisation of Gaussian processes has been performed by Sollich [150] and Opper and Vivarelli [106]. Other analyses of generalisation are possible, based on statistical mechanics (see for example [105]), or on the theory of on-line algorithms [75].

These references are also given on the website **www.support-vector.net**, which will be kept up to date with new work, pointers to software and papers that are available on-line.