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].