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

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

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

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

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,

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

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

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

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).

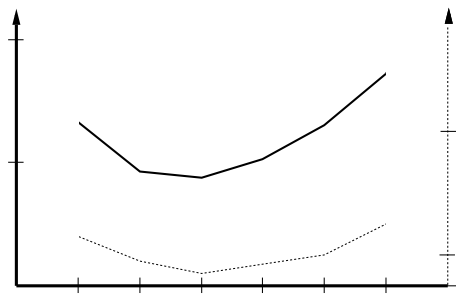

Figure 5.5 Average VC dimension (solid), and total number of test errors, of ten two-class-classifiers (dotted) with polynomial degrees 2 through 7, trained on the USPS set of handwritten digits. The baseline 174 on the error scale, corresponds to the total number of test errors of the ten *best* binary 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]).

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

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:

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

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:*