$$R[f] = \int \frac{1}{2}|f(x) - y| \, d\mathrm{P}(x, y). \tag{1.18}$$

The risk can be defined for any loss function, provided the integral exists. For the present zero-one loss function, the risk equals the probability of misclassification.[7]

Capacity

Statistical learning theory (Chapter 5, [570, 559, 561, 136, 562, 14]), or VC (Vapnik-Chervonenkis) theory, shows that it is imperative to restrict the set of functions from which f is chosen to one that has a *capacity* suitable for the amount of available training data. VC theory provides *bounds* on the test error. The minimization of these bounds, which depend on both the empirical risk and the capacity of the function class, leads to the principle of *structural risk minimization* [559].

VC dimension

Shattering

The best-known capacity concept of VC theory is the *VC dimension*, defined as follows: each function of the class separates the patterns in a certain way and thus induces a certain labelling of the patterns. Since the labels are in $\{\pm 1\}$, there are at most 2^m different labellings for m patterns. A very rich function class might be able to realize all 2^m separations, in which case it is said to *shatter* the m points. However, a given class of functions might not be sufficiently rich to shatter the m points. The VC dimension is defined as the largest m such that there exists a set of m points which the class can shatter, and ∞ if no such m exists. It can be thought of as a one-number summary of a learning machine's capacity (for an example, see Figure 1.4). As such, it is necessarily somewhat crude. More accurate capacity measures are the *annealed VC entropy* or the *growth function*. These are usually considered to be harder to evaluate, but they play a fundamental role in the conceptual part of VC theory. Another interesting capacity measure, which can be thought of as a scale-sensitive version of the VC dimension, is the *fat shattering dimension* [286, 6]. For further details, cf. Chapters 5 and 12.

VC Bound

Whilst it will be difficult for the non-expert to appreciate the results of VC theory in this chapter, we will nevertheless briefly describe an example of a VC bound:

7. The risk-based approach to machine learning has its roots in statistical decision theory [582, 166, 43]. In that context, $f(x)$ is thought of as an *action*, and the loss function measures the loss incurred by taking action $f(x)$ upon observing x when the true output (state of nature) is y.

Like many fields of statistics, decision theory comes in two flavors. The present approach is a *frequentist* one. It considers the risk as a function of the distribution P and the decision function f. The *Bayesian* approach considers parametrized families P_Θ to model the distribution. Given a prior over Θ (which need not in general be a finite-dimensional vector), the *Bayes risk* of a decision function f is the *expected* frequentist risk, where the expectation is taken over the prior. Minimizing the Bayes risk (over decision functions) then leads to a *Bayes decision function*. Bayesians thus act as if the parameter Θ were actually a random variable whose distribution is known. Frequentists, who do not make this (somewhat bold) assumption, have to resort to other strategies for picking a decision function. Examples thereof are considerations like *invariance* and *unbiasedness*, both used to restrict the class of decision rules, and the *minimax* principle. A decision function is said to be minimax if it minimizes (over all decision functions) the maximal (over all distributions) risk. For a discussion of the relationship of these issues to VC theory, see Problem 5.9.