Chapter 2
Machine Learning

Abstract We present an extensive review on the subject of machine learning by studying existing literature. We focus primarily on the main approaches that have been proposed in order to address the problem of machine learning and how they may be categorized according to type and amount of inference. Specifically, the categorization of the various machine learning paradigms according to the type of inference, involves the following two approaches:

- Model Identification or Parametric Inference; and
- Model Prediction or General Inference.

The general framework of the parametric model, in particular, introduces the principles of Empirical Risk Minimization (ERM) and Structural Risk Minimization. On the other hand, the Transductive Inference Model is defined as an extension to the original paradigm of General Inference. The categorization of machine learning models according to the amount of inference includes the following approaches:

- Rote Learning;
- Learning from Instruction; and
- Learning from Examples.

Specifically, Learning from Examples provides the framework to analyze the problem of minimizing a risk functional on a given set of empirical data which is the fundamental problem within the field of pattern recognition. In essence, the particular form of the risk functional defines the primary problems of machine learning, namely:

- The Classification Problem;
- The Regression Problem; and
- The Density Estimation Problem which is closely related to the Clustering Problem.

Finally, in this chapter we present a conspectus of the theoretical foundations behind Statistical Learning Theory.

D.N. Sotiropoulos and G.A. Tsihrintzis, *Machine Learning Paradigms*,
Intelligent Systems Reference Library 118, DOI 10.1007/978-3-319-47194-5_2

2.1 Introduction

The ability to learn is one of the most distinctive attributes of intelligent behavior. An informal definition of the learning process in general could be articulated as: "*The learning process includes the acquisition of new declarative knowledge, the development of new skills through interaction or practice, the organization of new knowledge into general, effective representations, and the discovery of new facts and theories through observation and experimentation*". The term *machine learning*, on the other hand, covers a broad range of computer programs. In general, any computer program that improves its performance through experience or training can be called a learning program. Machine learning constitutes an integral part of artificial intelligence since the primary feature of any intelligent system is the ability to learn. Specifically, systems that have the ability to learn need not be implicitly programmed for any possible problematic situation. In other words, the development of machine learning alleviates the system designer from the burden of foreseeing and providing solutions for all possible situations.

The study and modelling of learning processes in their multiple manifestations constitute the topic of machine learning. In particular, machine learning has been developed around the following primary research lines:

- *Task-oriented studies*, which are focused on developing learning systems in order to improve their performance in a predetermined set of tasks.
- *Cognitive simulation*, that is, the investigation and computer simulation of human learning processes.
- *Theoretical analysis*, which stands for the investigation of possible learning methods and algorithms independently of the particular application domain.
- *Derivation of machine learning paradigms and algorithms* by developing metaphors for biological processes that may be interesting within the context of machine learning. A typical example is the field of biologically inspired computing which led to the emergence of Artificial Neural Networks and Artificial Immune Systems.

The following sections provide an overview of the various machine learning approaches that have been proposed over the years according to different viewpoints concerning the *underlying learning strategies*. Specifically, Sect. 2.2 provides a more general categorization of the machine learning methodologies based on the particular type of inference utilized while Sect. 2.3 provides a more specialized analysis according to the amount of inference. Finally, Sect. 2.5 gives a theoretical justification of Statistical Learning Theory.

2.2 Machine Learning Categorization According to the Type of Inference

The fundamental elements of statistical inference have existed for more than 200 years, due to the seminal works of Gauss and Laplace. However, their systematic analysis began in the late 1920s. By that time, descriptive statistics was mostly complete since it was shown that many events of the real world are sufficiently described by different statistical laws. Specifically, statisticians have developed powerful mathematical tools, namely distribution functions, that have the ability to capture interesting aspects of reality. However, a crucial question that was yet to be answered concerned the determination of a reliable method for performing statistical inference. A more formal definition of the related problem could be the following: *Given a collection of empirical data originating from some functional dependency, infer this dependency*. Therefore, the analysis of methods of statistical inference signaled the beginning of a new era for statistics which was significantly influenced by two bright events:

1. Fisher introduced the main models of statistical inference in the unified framework of parametric statistics. His work indicated that the various problems related to the estimation of functions from given data (the problems of discriminant analysis, regression analysis, and density estimation) are particular instances of the more general problem dealing with the parameter estimation of a specific parametric model. In particular, he suggested the Maximum Likelihood method as a for the estimation of the unknown parameters in all these models.
2. Glivenko, Cantelli and Kolmogorov, on the other hand, started a general analysis of statistical inference. One of the major findings of this quest was the Glivenko–Cantelli theorem stating that the empirical distribution function always converges to the actual distribution function. Another equally important finding came from Kolmogorov who found the asymptotically exact rate of this convergence. Specifically, he proved that the rate turns out to be exponentially fast and independent of the unknown distribution function.

Notwithstanding, these two events determined the two main approaches that were adopted within the general context of machine learning:

1. *Model Identification* or *particular (parametric) inference* which aims at creating simple statistical methods of inference that can be used for solving real-life problems, and
2. *Model Prediction* or *general inference*, which aims at finding one induction method for any problem of statistical inference.

The philosophy that led to the conception of the model identification approach is based upon the belief that the investigator knows the problem to be analyzed relatively well. Specifically, he/she is aware of the physical law that generates the stochastic properties of the data and the function to be found up to a finite number of parameters. According to the model identification approach, the very essence of the

statistical inference problem is the estimation of these parameters by utilizing the available data. Therefore, the natural solution in finding these parameters is obtained by utilizing information concerning the statistical law and the target function is the adaptation of the maximum likelihood method. The primary purpose of this theory is to justify the corresponding approach by discovering and describing its favorable properties.

On the contrary, the philosophy that led to the conception of the model prediction approach is focused on the fact that there is no reliable a priori information concerning the statistical law underlying the problem or the desirable function to be approximated. Therefore, it is necessary to find a method in order to infer the approximation function from the given examples in each situation. The corresponding theory of model prediction must be able to describe the conditions under which it is possible find the best approximation to an unknown function in a given set of functions with an increasing number of examples.

2.2.1 Model Identification

The model identification approach corresponding to the principle of parametric inference was developed very quickly since its original conception by Fisher. In fact, the main ideas underlying the parametric model were clarified in the 1930s and the main elements of theory of parametric inference were formulated within the next 10 years. Therefore, the time period between the 1930 and 1960 was the "golden age" of parametric inference which dominated statistical inference. At that time, there was only one legitimate approach to statistical inference, namely the theory that served the model identification approach. The classical parametric paradigm falls within the general framework introduced by Fisher according to which any signal Y can be modelled as consisting of a deterministic component and a random counterpart:

$$Y = f(X) + \epsilon \tag{2.1}$$

The deterministic part $f(X)$ is defined by the values of a known family of functions which are determined up to a limited number of parameters. The random part ϵ corresponds to the noise added to the signal, defined by a known density function. Fisher considered the estimation of the parameters of the function $f(X)$ as the goal of statistical analysis. Specifically, in order to find these parameters he introduced the maximum likelihood method. Since the main goal of Fisher's statistical framework was to estimate the model that generated the observed signal, his paradigm is identified by the term "*Model Identification*". In particular, Fisher's approach reflects the traditional idea of Science concerning the process of inductive inference, which can be roughly summarized by the following steps:

1. Observe a phenomenon.
2. Construct a model of that phenomenon (*inductive step*).

3. Make predictions using this model (*deductive step*).

The philosophy of this classical paradigm is based upon on the following beliefs:

1. *In order to find a dependency from the data, the statistician is able to define a set of functions, linear in their parameters, that contain a good approximation to the desired function. The number of parameters describing this set is small.*
   This belief was specifically supported by referring to the Weierstrass theorem, according to which any continuous function with a finite number of discontinuities can be approximated on a finite interval by polynomials (functions linear in their parameters) with any degree of accuracy. The main idea was that this set of functions could be replaced by an alternative set of functions, not necessarily polynomials, but linear with respect to a small number of parameters. Therefore, one could obtain a good approximation to the desired function.
2. *The statistical law underlying the stochastic component of most real-life problems is the normal law.*
   This belief was supported by referring to the Central Limit Theorem, which states that under wide conditions the sum of a large number of random variables is approximated by the normal law. The main idea was that if randomness in a particular problem is the result of interaction among a large number of random components, then the stochastic element of the problem will be described by the normal law.
3. *The maximum likelihood estimate may serve as a good induction engine in this paradigm.*
   This belief was supported by many theorems concerning the conditional optimality of the maximum likelihood method in a restricted set of methods or asymptotically. Moreover, there was hope that this methodology would offer a good tool even for small sample sizes.

Finally, these three beliefs are supported by the following more general philosophy:
If there exists a mathematical proof that some method provides an asymptotically optimal solution, then in real life this method will provide a reasonable solution for a small number of data samples.

The classical paradigm deals with the identification of stochastic objects which particulary relate to the problems concerning the estimation of densities and conditional densities.

Density Estimation Problem

The first problem to be considered is the density estimation problem. Letting $\boldsymbol{\xi}$ be a random vector then the probability of the random event $F(\mathbf{x}) = P(\boldsymbol{\xi} < \mathbf{x})$ is called a *probability distribution function of the random vector* $\boldsymbol{\xi}$ where the inequality is interpreted coordinatewise. Specifically, the random vector $\boldsymbol{\xi}$ has a density function if there exists a nonnegative function $p(\mathbf{u})$ such that for all $\mathbf{x}$ the equality

$$F(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} p(\mathbf{u}) d\mathbf{u} \tag{2.2}$$

is valid. The function $p(\mathbf{x})$ is called a *probability density* of the random vector. Therefore, by definition, the problem of estimating a probability density from the data requires a solution of the integral equation:

$$\int_{-\infty}^{\mathbf{x}} p(\mathbf{u}, \alpha) d\mathbf{u} = F(\mathbf{x}) \tag{2.3}$$

on a given set of densities $p(\mathbf{x}, \alpha)$ where $\alpha \in \Lambda$. It is important to note that while the true distribution function $F(\mathbf{x})$ is unknown, one is given a random independent sample

$$\mathbf{x}_1, \ldots, \mathbf{x}_l \tag{2.4}$$

which is obtained in accordance with $F(\mathbf{x})$. Then it is possible to construct a series of approximations to the distribution function $F(\mathbf{x})$ by utilizing the given data set (2.4) in order to form the so-called *empirical distribution function* which is defined by the following equation:

$$F_l(\mathbf{x}) = \frac{1}{l} \sum_{i=1}^{l} \theta(\mathbf{x} - \mathbf{x}_i), \tag{2.5}$$

where $\theta(\mathbf{u})$ corresponds to the step function defined as:

$$\theta(\mathbf{u}) = \begin{cases} \mathbf{1}, & \text{when all the coordinates of vector } \mathbf{u} \text{ are positive,} \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{2.6}$$

Thus, the problem of density estimation consists of finding an approximation to the solution of the integral equation (2.3). Even if the probability density function is unknown, an approximation to this function can be obtained.

Conditional Probability Estimation Problem

Consider pairs $(\omega, \mathbf{x})$ where $\mathbf{x}$ is a vector and ω is scalar which takes on only k values from the set $\{0, 1, \ldots, k-1\}$. According to the definition, the conditional probability $P(\omega, \mathbf{x})$ is a solution of the integral equation:

$$\int_{-\infty}^{\mathbf{x}} P(\omega|\mathbf{t}) dF(\mathbf{t}) = F(\omega, \mathbf{x}), \tag{2.7}$$

where $F(\mathbf{x})$ is the distribution function of the random vector $\mathbf{x}$ and $F(\omega, \mathbf{x})$ is the joint distribution function of pairs $(\omega, \mathbf{x})$. Therefore, the problem of estimating the conditional probability in the set of functions $P_\alpha(\omega|\mathbf{x})$, where $\alpha \in \Lambda$, is to obtain an approximation to the integral equation (2.7) when both distribution functions $F(\mathbf{x})$ and $F(\omega, \mathbf{x})$ are unknown, but the following set of samples are available:

$$(\omega_1, \mathbf{x}_1), \ldots, (\omega_l, \mathbf{x}_l). \tag{2.8}$$

As in the case of the density estimation problem, the unknown distribution functions $F(\mathbf{x})$ and $F(\omega, \mathbf{x})$ can be approximated by the empirical distribution functions (2.5) and function:

$$F_l(\omega, \mathbf{x}) = \frac{1}{l} \sum_{i=1}^{l} \theta(\mathbf{x} - \mathbf{x}_i)\delta(\omega, \mathbf{x}_i), \tag{2.9}$$

where the function $\delta(\omega, \mathbf{x})$ is defined as:

$$\delta(\omega, \mathbf{x}) = \begin{cases} 1, & \text{if the vector } \mathbf{x} \text{ belongs to class } \omega, \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

Thus, the problem of conditional probability estimation may be resolved by obtaining an approximation to the solution of the integral equation (2.7) in the set of functions $P_\alpha(\omega|\mathbf{x})$ where $\alpha \in \Lambda$. This solution, however, is difficult to get since the probability density functions $F(\mathbf{x})$ and $F(\omega, \mathbf{x})$ are unknown and they can only be approximated by the empirical functions $F_l(\mathbf{x})$ and $F_l(\omega, \mathbf{x})$.

Conditional Density Estimation Problem

The last problem to be considered is the one related to the conditional density estimation. By definition, this problem consists in solving the following integral equation:

$$\int_{-\infty}^{y} \int_{-\infty}^{\mathbf{x}} p(t|\mathbf{u})dF(\mathbf{u})dt = F(y, \mathbf{x}), \tag{2.11}$$

where the variables y are scalars and the variables $\mathbf{x}$ are vectors. Moreover, $F(\mathbf{x})$ is a probability distribution function which has a density, $p(y, \mathbf{x})$ is the conditional density of y given $\mathbf{x}$, and $F(y, \mathbf{x})$ is the joint probability distribution function defined on the pairs $(y, \mathbf{x})$. The desirable conditional density function $p(y|\mathbf{x})$ can be obtained by considering a series of approximation functions which satisfy the integral equation (2.11) on the given set of functions and the i.i.d pairs of the given data:

$$(y_1, \mathbf{x}_1), \ldots, (y_l, \mathbf{x}_l) \tag{2.12}$$

when both distributions $F(\mathbf{x})$ and $F(y, \mathbf{x})$ are unknown. Once again, it is possible to approximate the empirical distribution function $F_l(\mathbf{x})$ and the empirical joint distribution function:

$$F_l(y, \mathbf{x}) = \frac{1}{l} \sum_{i=1}^{l} \theta(y - y_i)\theta(\mathbf{x} - \mathbf{x}_i). \tag{2.13}$$

Therefore, the problem is to get an approximation to the solution of the integral equation (2.11) in the set of functions $p_a(y, \mathbf{x})$ where $\alpha \in \Lambda$, when the probability

distribution functions are unknown but can be approximated by $F_l(\mathbf{x})$ and $F_l(y, \mathbf{x})$ using the data (2.12).

2.2.2 Shortcoming of the Model Identification Approach

All three problems of stochastic dependency estimation that were thoroughly discussed previously can be described in the following general way. Specifically, they are reduced to solving the following linear continuous operator equation

$$Af = F, f \in \mathcal{F} \tag{2.14}$$

given the constraint that some functions that form the equation are unknown. The unknown functions, however, can be approximated by utilizing a given set of sample data. In this way it possible to obtain approximations to the distribution functions $F_l(\mathbf{x})$ and $F_l(y, \mathbf{x})$. This formulation can reveal a main difference between the problem of density estimation and the problems of conditional probability and conditional density estimation. Particularly, in the problem of density estimation, instead of an accurate right-hand side of the equation only an approximation is available. Therefore, the problem involves getting an approximation to the solution of Eq. (2.14) from the relationship

$$Af \approx F_l, f \in \mathcal{F}. \tag{2.15}$$

On the other hand, in the problems dealing with the conditional probability and conditional density estimation not only the right-hand side of Eq. (2.14) is known only approximately, but the operator A is known only approximately as well. This being true, the true distribution functions appearing in Eqs. (2.7) and (2.11) are replaced by their approximations. Therefore, the problem consists in getting an approximation to the solution of Eq. (2.14) from the relationship

$$A_l f \approx F_l, f \in \mathcal{F}. \tag{2.16}$$

The Glivenko–Cantelli theorem ensures that the utilized approximation functions converge to the true distribution functions as the number of observations goes to infinity. Specifically, the Glivenko–Cantelli theorem states that the convergence

$$\sup_{\mathbf{x}} |F(\mathbf{x}) - F_l(\mathbf{x})| \xrightarrow[l\to\infty]{P} 0 \tag{2.17}$$

takes place. A fundamental disadvantage of this approach is that solving the general operator Eq. (2.14) results in an ill-posed problem. Ill-posed problems are extremely difficult to solve since they violate the well-posedness conditions introduced by Hadamard involving the existence of a solution, the uniqueness of that solution and

the continuous dependence of the solution on the empirical data. That is, the solutions of the corresponding integral equations are unstable.

Moreover, the wide application of computers, in the 1960s, for solving scientific and applied problems revealed additional shortcomings of the model identification approach. It was the fist time that researchers utilized computers in an attempt to analyze sophisticated models that had many factors or in order to obtain more precise approximations.

In particular, the computer analysis of large scale multivariate problems revealed the phenomenon that R. Bellman called "*the curse of dimensionality*". It was observed that increasing the number of factors that have to be taken into consideration requires an exponentially increasing amount of computational resources. Thus, in real-life multidimensional problems where there might be hundreds of variables, the belief that it is possible to define a reasonably small set of functions that contains a good approximation to the desired one is not realistic.

Approximately at the same time, Tukey demonstrated that the statistical components of real-life problems cannot be described by only classical distribution functions. By analyzing real-life data, Tukey discovered that the corresponding true distributions are in fact different. This entails that it is crucial to take this difference into serious consideration in order to construct effective algorithms.

Finally, James and Stein showed that even for simple density estimation problems, such as determining the location parameters of a $n > 2$ dimensional normal distribution with a unit covariance matrix, the maximum likelihood method is not the best one.

Therefore, all three beliefs upon which the classical parametric paradigm relied turned out to be inappropriate for many real-life problems. This had an enormous consequence for statistical science since it looked like the idea of constructing statistical inductive inference models for real-life problems had failed.

2.2.3 Model Prediction

The return to the general problem of statistical inference occurred so imperceptibly that it was not recognized for more than 20 years since Fisher's original formulation of the parametric model. Of course, the results from Glivenko, Cantelli and Kolmogorov were known but they were considered to be inner technical achievements that are necessary for the foundation of statistical inference. In other words, these results could not be interpreted as an indication that there could be a different type of inference which is more general and more powerful than the classical parametric paradigm.

This question was not addressed until after the late 1960s when Vapnik and Chervonenkis started a new paradigm called *Model Prediction* (or predictive inference). The goal of model prediction is to predict events well, but not necessarily through the identification of the model of events. The rationale behind the model prediction paradigm is that the problem of estimating a model is hard (ill-posed) while the

problem of finding a rule for good prediction is much easier (better-posed). Specifically, it could happen that there are many rules that predict the events well and are very different from the true model. Nonetheless, these rules can still be very useful predictive tools.

The model prediction paradigm was initially boosted when in 1958 F. Rosenblatt, a physiologist, suggested a learning machine (computer program) called the Perceptron for solving the simplest learning problem, namely the pattern classification/recognition problem. The construction of this machine incorporated several existing neurophysiological models of learning mechanisms. In particular, F. Rosenblatt demonstrated that even with the simplest examples the Perceptron was able to generalize without constructing a precise model of the data generation process. Moreover, after the introduction of the Perceptron, many learning machines were suggested that had no neurobiological analogy but they did not generalize worse than Perceptron. Therefore, a natural question arose:

Does there exist something common in these machines? Does there exist a general principle of inductive inference that they implement?

Immediately, a candidate was found as a general induction principle, the so-called *empirical risk minimization* (ERM) principle. The ERM principle suggests the utilization of a decision rule (an indicator function) which minimizes the number of training errors (empirical risk) in order to achieve good generalization on future (test) examples. The problem, however, was to construct a theory for that principle.

At the end of 1960s, the theory of ERM for the pattern recognition problem was constructed. This theory includes the general *qualitative theory* of generalization that described the necessary and sufficient conditions of consistency of the ERM induction principle. Specifically, the consistency of the ERM induction principle suggests that it is valid for any set of indicator functions, that is {0, 1}-valued functions on which the machine minimizes the empirical risk. Additionally, the new theory includes the general *quantitative theory* describing the bounds on the probability of the (future) test error for the function minimizing the empirical risk.

The application of the ERM principle, however, does not necessarily guarantee consistency, that is convergence to the best possible solution with an increasing number of observations. Therefore, the primary issues that drove the development of the ERM theory were the following:

1. Describing situations under which the method is consistent, that is, to find the necessary and sufficient conditions for which the ERM method defines functions that converge to the best possible solution with an increasing number of observations. The resulting theorems thereby describe the qualitative model of ERM principle.
2. Estimating the quality of the solution obtained on the basis of the given sample size. This entails, primarily, to estimate the probability of error for the function that minimizes the empirical risk on the given set of training examples and secondly to estimate how close this probability is to the smallest possible for the given set of functions. The resulting theorems characterize the generalization ability of the ERM principle.

In order to address both issues for the pattern recognition problem it is necessary to construct a theory that could be considered as a generalization of the Glivenko–Cantelli–Kolmogorov results. This is equivalent to the following statements:

1. For any given set of events, to determine whether the uniform law of large numbers holds, that is whether uniform convergence takes place.
2. If uniform convergence holds, to find the bounds for the *non-asymptotic* rate of uniform convergence.

This was the theory constructed by Vapnik and Chervonenkis which was based on a collection of new concepts, the so-called capacity concepts for a set of indicator functions. The most important new concept was the so-called VC dimension of the set of indicator functions which characterizes the variability of the set of indicator functions. Specifically, it was found that both the necessary and sufficient conditions of consistency and the rate of convergence of the ERM principle depend on the capacity of the set of functions that are implemented by the learning machine. The most preeminent results of the new theory that particularly relate to the VC dimension are the following:

1. For distribution-independent consistency of the ERM principle, the set of functions implemented by the learning machine must have a finite VC dimension.
2. Distribution-free bounds on the rate of uniform convergence depend on the VC dimension, the number of errors, and the number of observations.

The bounds for the rate of uniform convergence not only provide the main theoretical basis for the ERM inference, but also motivate a new method of inductive inference. For any level of confidence, an equivalent form of the bounds define bounds on the probability of the test error *simultaneously for all functions of the learning machine* as a function of the training errors, of the VC dimension of the set of functions implemented by the learning machine, and of the number of observations. This form of the bounds led to a new idea for controlling the generalization ability of learning machines:

In order to achieve the smallest bound on the test error by minimizing the number of training errors, the machine (set of functions) with the smallest VC dimension should be used.

These two requirements define a pair of contradictory goals involving the simultaneous minimization of the number of training errors and the utilization of a learning machine (set of functions) with a small VC dimension. In order to minimize the number of training errors, one needs to choose a function from a wide set of functions, rather than from a narrow set, with small VC dimension. Therefore, to find the best guaranteed solution, one has to compromise between the accuracy of approximation of the training data and the capacity (VC dimension) of the machine that is used for the minimization of errors. The idea of minimizing the test error by controlling two contradictory factors was formalized within the context of a new induction principle, the so-called Structural Risk Minimization (SRM) principle.

The fundamental philosophy behind the SRM principle is the so-called Occam's razor which was originally proposed by William of Occam in the fourteenth century,

stating that *entities should not be multiplied beyond necessity*. In particular, the most common interpretation of Occam's razor is that *the simplest explanation is the best*. The assertion coming from SRM theory, however, is different and suggests that one should choose the explanation provided by the machine with the smallest capacity (VC dimension).

The SRM principle constitutes an integral part of the model prediction paradigm which was established by the pioneering work of Vapnik and Chervonenkis. Specifically, one of the most important achievements of the new theory concerns the discovery that the generalization ability of a learning machine depends on the capacity of the set of functions which are implemented by the learning machine which is different from the number of free parameters. Moreover, the notion of capacity determines the necessary and sufficient conditions ensuring the consistency of the learning process and the rate of convergence. In other words, it reflects intrinsic properties of inductive inference.

In order to extend the model prediction paradigm, Vapnik introduced the *Transductive Inference* paradigm in the 1980s. The goal of transductive inference is to estimate the values of an unknown predictive function at a given point of interest, but not in the whole domain of its definition. The rationale behind this approach is that it is possible to achieve more accurate solutions by solving less demanding problems. The more general philosophical underpinning behind the transductive paradigm can be summarized by the following imperative:

If you possess a restricted amount of information for solving some general problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

In many real-life problems, the goal is to find the values of an unknown function only at points of interest, namely the testing data points. In order to solve this problem the model prediction approach uses a two-stage procedure which is particularly illustrated in Fig. 2.1.

At the first stage (inductive step) a function is estimated from a given set of functions, while at the second stage (deductive step) this function is used in order to evaluate the values of the unknown function at the points of interest. It is obvious that at the first stage of this two-stage scheme one addresses a problem that is more general than the one that needs to be solved. This is true since estimating an unknown

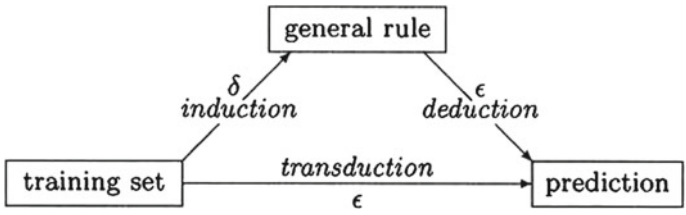

Fig. 2.1 Inference models

function involves estimating its values at all points in the function domain when only a few are of practical importance. In situations when there is only a restricted amount of information, it is possible to be able to estimate the values of the unknown function reasonably well at the given points of interest but cannot estimate the values of the function well at any point within the function domain. The direct estimation of function values only at points of interest using a given set of functions forms the transductive type of inference. As clearly depicted in Fig. 2.1, the transductive solution derives results in one step, directly from particular to particular (transductive step).

2.3 Machine Learning Categorization According to the Amount of Inference

Although machine learning paradigms can be categorized according to the type of inference that is performed by the corresponding machines, a common choice is to classify learning systems based on the amount of inference. Specifically, this categorization concerns the amount of inference that is performed by the learner which is one of the two primary entities in machine learning, the other being the supervisor (teacher). The supervisor is the entity that has the required knowledge to perform a given task, while the learner is the entity that has to learn the knowledge in order to perform a given task. In this context, the various learning strategies can be distinguished by the amount of inference the learner performs on the information given by the supervisor.

Actually, there are two extreme cases of inference, namely performing no inference and performing a remarkable amount of inference. If a computer system (the learner) is programmed directly, its knowledge increases but it performs no inference since all cognitive efforts are developed by the programmer (the supervisor). On the other hand, if a systems independently discovers new theories or invents new concepts, it must perform a very substantial amount of inference since it is deriving organized knowledge from experiments and observations. An intermediate case could be a student determining how to solve a math problem by analogy to problem solutions contained in a textbook. This process requires inference but much less than discovering a new theorem in mathematics.

Increasing the amount of inference that the learner is capable of performing, the burden on the supervisor decreases. The following taxonomy of machine learning paradigms captures the notion of trade-off in the amount of effort that is required of the learner and of the supervisor. Therefore, there are four different learning types that can be identified, namely *rote learning*, *learning from instruction*, *learning by analogy* and *learning from examples*.

2.3.1 Rote Learning

Rote learning consists in the direct implanting if knowledge into a learning system. Therefore, there is no inference or other transformation of the knowledge involved on the part of the learner. There are, of course, several variations of this method such as:

- Learning by being programmed or modified by an external entity. This variation requires no effort in the part of the learner. A typical paradigm is the usual style of computer programming.
- Learning by memorization of given facts and data with no inference drawn from incoming information. For instance, the primitive database systems.

2.3.2 Learning from Instruction

Learning from instruction (or *learning by being told*) consists in acquiring knowledge from a supervisor or other organized source, such as a textbook, requiring that the learner transforms the knowledge from the input language to an internal representation. The new information is integrated with the prior knowledge for effective use. The learner is required to perform some inference, but a large fraction of the cognitive burden remains with the supervisor, who must present and organize knowledge in a way that incrementally increases the learner's actual knowledge. In other words, learning from instruction mimics education methods. In this context, the machine learning task involves building a system that can accept and store instructions in order to efficiently cope with a future situation.

2.3.3 Learning by Analogy

Learning by analogy consists in acquiring new facts or skills by transforming and increasing existing knowledge that bears strong similarity to the desired new concept or skill into a form effectively useful in the new situation. A learning-by-analogy system could be applied in order to convert an existing computer program into one that performs a closely related function for which it was not originally designed. Learning by analogy requires more inference on the part of the learner that does rote learning or learning from instruction. A fact or skill analogous in relevant parameters must be retrieved from memory which will be subsequently transformed in order to be applied to the new situation.

2.4 Learning from Examples

Learning from examples is a model addressing the problem of functional dependency estimation within the general setting of machine learning. The fundamental components of this model, as they are illustrated in Fig. 2.2, are the following:

1. The generator of the data G.
2. The target operator or *supervisor's operator* S.
3. The learning machine LM.

The generator G serves as the main environmental factor generating the *independently and identically distributed* (i.i.d) random vectors $\mathbf{x} \in \mathbf{X}$ according to some unknown (but fixed) probability distribution function $F(\mathbf{x})$. In other words, the generator G determines the common framework in which the supervisor and the learning machine act. The random vectors $\mathbf{x} \in \mathbf{X}$ are subsequently fed as inputs to the target operator (supervisor S) which finally returns the output values y. It is important to note that although there is no information concerning the transformation of input vector to output values, it is known that the corresponding target operator exists and does not change. The learning machine observes l pairs

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \tag{2.18}$$

(the training set) which contains input vectors $\mathbf{x}$ and the supervisor's response y. During this period the learning machine constructs some operator which will be used for prediction of the supervisor's answer y_i on an particular observation vector $\mathbf{x}$ generated by the generator G. Therefore, the goal of the learning machine is to construct an appropriate approximation. In order to be a mathematical statement, this general scheme of learning from examples needs some clarification. First of all, it is important to describe the kind of functions that are utilized by the supervisor. In this monograph, it is assumed that the supervisor returns the output value y on the input vector $\mathbf{x}$ according to a conditional distribution function $F(y|\mathbf{x})$ including the case when the supervisor uses a function of the form $y = f(\mathbf{x})$. Thus, the learning machine observes the training set, which is drawn randomly and independently according to a joint distribution function $F(\mathbf{x}, y) = F(\mathbf{x})F(y|\mathbf{x})$ and by utilizing this training set it constructs an approximation to the unknown operator. From a formal point

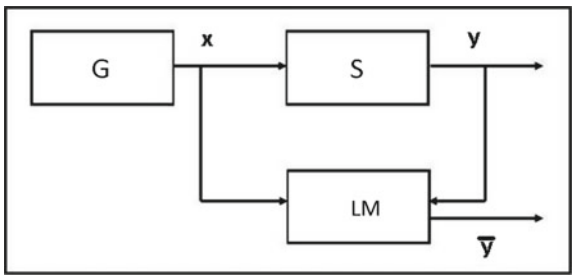

Fig. 2.2 Learning from examples

of view, the process of constructing an operator consists of developing a learning machine having the ability to implement some fixed set of functions given by the construction of the machine. Therefore, *the learning process is a process of choosing an appropriate function from a given set of functions*.

2.4.1 The Problem of Minimizing the Risk Functional from Empirical Data

Each time the problem of selecting a function with desired qualities arises, one should look in the set of possible functions for the one that satisfies the given quality criterion in the best possible way. Formally, this means that on a subset Z of the vector space $\mathbb{R}^n$, a set of admissible functions $\{g(\mathbf{z})\}$, $\mathbf{z} \in Z$, is given, and a functional

$$R = R(g(\mathbf{z})) \tag{2.19}$$

is defined as the criterion of quality for the evaluation of any given function. It is then required to find the function $g'(\mathbf{z})$ minimizing the functional (2.19) assuming that the minimum of the functional corresponds to the best quality and that the minimum of (2.19) exists in $\{g(\mathbf{z})\}$. In the case when the set of functions $\{g(\mathbf{z})\}$ and the functional $R(g(\mathbf{z}))$ were explicitly given, finding the function $g'(\mathbf{z})$ which minimizes (2.19) would a problem of the calculus of variations. In real-life problems, however, this is merely the case since the most common situation is that the risk functional is defined on the basis of a given probability distribution $F(\mathbf{z})$ defined on Z. Formally, the risk functional is defined as the mathematical expectation given by the following equation

$$R(g(\mathbf{z})) = \int L(\mathbf{z}, g(\mathbf{z}))dF(\mathbf{z}) \tag{2.20}$$

where the function $L(\mathbf{z}, g(\mathbf{z}))$ is integrable for any $g(\mathbf{z}) \in \{g(\mathbf{z})\}$. Therefore, the problem is to minimize the risk functional (2.20) in the case when the probability distribution $F(\mathbf{z})$ is unknown but the sample

$$\mathbf{z}_1, \ldots, \mathbf{z}_l \tag{2.21}$$

of observations drawn randomly and independently according to $F(\mathbf{z})$ is available.

It is important to note that there is a substantial difference between problems arising when the optimization process involves the direct minimization of the functional (2.19) and those encountered when the functional (2.20) is minimized on the basis of the empirical data (2.21). In the case of minimizing (2.19) the problem reduces to organizing the search for the function $g'(\mathbf{z})$ from the set $\{g(\mathbf{z})\}$ which minimizes (2.19). On the other hand, when the functional (2.20) is to be minimized on the basis of the empirical data (2.21), the problem reduces to formulating a constructive criterion that will be utilized in order to choose the optimal function rather than

organizing the search of the functions in $\{g(\mathbf{z})\}$. Therefore, the question in the first case is: *How to obtain the minimum of the functional in the given set of functions?* On the other hand, in the second case the question is: *What should be minimized in order to select from the set* $\{g(\mathbf{z})\}$ *a function which will guarantee that the functional (2.20) is small?*

Strictly speaking, the direct minimization of the risk functional (2.20) based on the empirical data (2.21) is impossible based on the utilization of methods that are developed in optimization theory. This problem, however, lies within the core of mathematical statistics.

When formulating the minimization problem for the functional (2.20), the set of functions $g(\mathbf{z})$ will be given in a parametric form $\{g(\mathbf{z}, \alpha), \alpha \in \Lambda\}$. Here α is parameter from the set Λ such that the value $\alpha = \alpha^*$ defines the specific function $g(z, \alpha^*)$ in the set $g(z, \alpha)$. Therefore, identifying the required function is equivalent to determining the corresponding parameter $\alpha \in \Lambda$. The exclusive utilization of parametric sets of functions does not imply a restriction on the problem, since the set Λ, to which the parameter α belongs, is arbitrary. In other words, Λ can be a set of scalar quantities, a set of vectors, or a set of abstract elements. Thus, in the context of the new notation the functional (2.20) can be rewritten as

$$R(a) = \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}), \alpha \in \Lambda, \tag{2.22}$$

where

$$Q(\mathbf{z}, \alpha) = L(\mathbf{z}, g(\mathbf{z}, \alpha)). \tag{2.23}$$

The function $Q(\mathbf{z}, \alpha)$ represents a *loss function* depending on the variables $\mathbf{z}$ and α.

The problem of minimizing the functional (2.22) may be interpreted in the following simple way: It is assumed that each function $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ (e.g. each function of $\mathbf{z}$ for a fixed $\alpha = \alpha^*$), determines the amount of loss resulting from the realization of the vector $\mathbf{z}$. Thus, the *expected loss* (with respect to $\mathbf{z}$) for the function $Q(\mathbf{z}, \alpha^*)$ will be determined by the integral

$$R(\alpha^*) = \int Q(\mathbf{z}, \alpha^*) dF(\mathbf{z}). \tag{2.24}$$

This functional is the so-called *risk functional* or *risk*. The problem, then, is to choose in the set $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, a function $Q(\mathbf{z}, \alpha_0$ which minimizes the risk when the probability distribution function is unknown but independent random observations $\mathbf{z}_1, \ldots, \mathbf{z}_l$ are given.

Let $\mathcal{P}_0$ be the set of all possible probability distribution functions on Z and $\mathcal{P}$ some subset of probability distribution functions from $\mathcal{P}_0$. In this context, the term "unknown probability distribution function", means that the only available information concerning $F(\mathbf{z})$ is the trivial statement that $F(\mathbf{z}) \in \mathcal{P}_0$.

2.4.2 Induction Principles for Minimizing the Risk Functional on Empirical Data

The natural problem that arises at this point concerns the minimization of the risk functional defined in Eq. (2.24) which is impossible to perform directly on the basis of an unknown probability distribution function $F(\mathbf{x})$ (which defines the risk). In order to address this problem Vapnik and Chervonenkis introduced a new induction principle, namely the principle of *Empirical Risk Minimization*. The principle of empirical risk minimization suggests that instead of minimizing the risk functional (2.22) one could alternatively minimize the functional

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(\mathbf{z}_i, \alpha), \tag{2.25}$$

which is called the *empirical risk functional*. The empirical risk functional is constructed on the basis of the data $\mathbf{z}_1, \ldots, \mathbf{z}_l$ which are obtained according to the distribution $F(\mathbf{z})$. This functional is defined in explicit form and may be subject to direct minimization. Letting the minimum of the risk functional (2.22) be attained at $Q(\mathbf{z}, \alpha_0)$ and the minimum of the empirical risk functional (2.25) be attained at $Q(\mathbf{z}, \alpha_l)$, then the latter may be considered as an approximation to the function $Q(\mathbf{z}, \alpha_0)$. This principle of solving the empirical risk minimization problem is called the empirical risk minimization (induction) principle.

2.4.3 Supervised Learning

In supervised learning (or learning *with a teacher*), the available data are given in the form of input-output pairs. In particular, each data sample consists of a particular input vector and the related output value. The primary purpose of this learning paradigm is to obtain a concise description of the data by finding a function which yields the correct output value for a given input pattern. The term supervised learning stems from the fact that the objects under consideration are already associated with target values which can be either integer class identifiers or real values. Specifically, the type of the output values distinguishes the two branches of the supervised learning paradigm corresponding to the learning problems of *classification* and *regression*.

The Problem of Pattern Recognition

The problem of pattern recognition was formulated in the late 1950s. In essence, it can be formulated as follows: A supervisor observes occurring situations and determines to which of k classes each one of them belongs. The main requirement of the problem is to construct a machine which, after observing the supervisor's classification, realizes an approximate classification in the same manner as the supervisor. A formal definition of the pattern recognition learning problem could be obtained

by considering the following statement: In a certain environment characterized by a probability distribution function $F(\mathbf{x})$, situation $\mathbf{x}$ appears randomly and independently. The supervisor classifies each one of the occurred situations into one of k classes. It is assumed that the supervisor carries out this classification by utilizing the conditional probability distribution function $F(\omega|\mathbf{x})$, where $\omega \in \{0, 1, \ldots, k-1\}$. Therefore, $\omega = p$ indicates that the supervisor assigns situation $\mathbf{x}$ the class p. The fundamental assumptions concerning the learning problem of pattern recognition is that neither the environment $F(\mathbf{x})$ nor the decision rule of the supervisor $F(\omega|\mathbf{x})$ are known. However, it is known that both functions exist meaning yielding the existence of the joint distribution $F(\omega, x) = F(\mathbf{x})F(\omega|\mathbf{x})$.

Let $\phi(\mathbf{x}, \alpha),\ \alpha \in \Lambda$ be a set of functions which can take only k discrete values contained within the $\{0, 1, \ldots, k-1\}$ set. In this setting, by considering the simplest loss function

$$L(\omega, \phi) = \begin{cases} 0, & \text{if } \omega = \phi; \\ 1, & \text{if } \omega \neq \phi. \end{cases} \tag{2.26}$$

the problem of pattern recognition may be formulated as the minimization of the risk functional

$$R(\alpha) = \int L(\omega, \phi(\mathbf{x}, \alpha))dF(\omega, \mathbf{x}) \tag{2.27}$$

on the set of functions $\phi(\mathbf{x}, \alpha),\ \alpha \in \Lambda$. The unknown distribution function $F(\omega, \mathbf{x})$ is implicitly described through a random independent sample of pairs

$$(\omega_1, \mathbf{x}_1), \ldots, (\omega_1, \mathbf{x}_l) \tag{2.28}$$

For the loss function (2.26), the functional defined in Eq. (2.27) determines the average probability of misclassification for any given decision rule $\phi(\mathbf{x}, \alpha)$. Therefore, the problem of pattern recognition reduces to the minimization of the average probability of misclassification when the probability distribution function $F(\omega, \mathbf{x})$ is unknown but the sample data (2.28) are given.

In this way, the problem of pattern recognition is reduced to the problem of minimizing the risk functional on the basis of empirical data. Specifically, the empirical risk functional for the pattern recognition problem has the following form:

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(\omega_i, \phi(\mathbf{x}_i, \alpha)),\ \alpha \in \Lambda. \tag{2.29}$$

The special feature of this problem, however, is that the set of loss functions $Q(\mathbf{z}, \alpha),\ \alpha \in \Lambda$ is not as arbitrary as in the general case defined by Eq. (2.23). The following restrictions are imposed:

- The vector $\mathbf{z}$ consists of $n+1$ coordinates: coordinate ω, which takes on only a finite number of values and n coordinates $(x^1, \ldots, x^n)$ which form the vector $\mathbf{x}$.

- The set of functions $Q(\mathbf{z}, \alpha),\ \alpha \in \Lambda$, is given by $Q(\mathbf{z}, \alpha) = L(\omega, \phi(\mathbf{x}, \alpha)),\ \alpha \in \Lambda$ taking only a finite number of values.

This specific feature of the risk minimization problem characterizes the pattern recognition problem. In particular, the pattern recognition problem forms the simplest learning problem because it deals with the simplest loss function. The loss function in the pattern recognition problem describes a set of *indicator* functions, that is functions that take only binary values.

The Problem of Regression Estimation

The problem of regression estimation involves two sets of elements X and Y which are connected by a functional dependence. In other words, for each element $\mathbf{x} \in X$ there is a unique corresponding element $y \in Y$. This relationship constitutes a function when X is a set of vectors and Y is a set of scalars. However, there exist relationships (stochastic dependencies) where each vector $\mathbf{x}$ can be mapped to a number of different y's which are obtained as a result of random trials. This is mathematically described by considering the conditional distribution function $F(y|\mathbf{x})$, defined on Y, according to which the selection of the y values is realized. Thus, the function of the conditional probability expresses the stochastic relationship between y and $\mathbf{x}$.

Let the vectors $\mathbf{x}$ appear randomly and independently in accordance with a distribution function $F(\mathbf{x})$. Then, it is reasonable to consider that the y values are likewise randomly sampled from the *conditional distribution function* $F(y|\mathbf{x})$. In this case, the sample data points may be considered to be generated according to a *joint probability distribution function* $F(\mathbf{x}, y)$. The most intriguing aspect of the regression estimation problem is that the distribution functions $F(\mathbf{x})$ and $F(y|\mathbf{x})$ defining the joint distribution function $F(y, \mathbf{x}) = F(\mathbf{x})F(y|\mathbf{x})$ are *unknown*. Once again, the problem of regression estimation reduces to the approximation of the true joint distribution function $F(y|\mathbf{x})$ through a series of randomly and independently sampled data points of the following form

$$(y_1, \mathbf{x}_1), \ldots, (y_l, \mathbf{x}_l). \tag{2.30}$$

However, the knowledge of the function $F(y, \mathbf{x})$ is often not required as in many cases it is sufficient to determine one of its characteristics, for example the function of the conditional mathematical expectation:

$$r(\mathbf{x}) = \int yF(y|\mathbf{x}) \tag{2.31}$$

This function is called the *regression* and the problem of its estimation in the set of functions $f(\mathbf{x}, \alpha),\ \alpha \in \Lambda$, is referred to as the problem of regression estimation. Specifically, it was proved that the problem of regression estimation can be reduced to the model of minimizing risk based on empirical data under the following conditions:

$$\int y^2 dF(y, \mathbf{x}) < \infty \text{ and } \int r^2(\mathbf{x}) dF(y, \mathbf{x}) < \infty \tag{2.32}$$

Indeed, on the set $f(\mathbf{x}, \alpha)$ the minimum of the functional

$$R(\alpha) = \int (y - f(\mathbf{x}, \alpha))^2 dF(y, \mathbf{x}) \tag{2.33}$$

(provided that it exists) is attained at the regression function if the regression $r(\mathbf{x})$ belongs to $f(\mathbf{x}, \alpha),\ \alpha \in \Lambda$. On the other hand, the minimum of this functional is attained at the function $f(\mathbf{x}, a^*)$, which is the closest to the regression $r(\mathbf{x})$ in the metric $L_2(P)$, defined as

$$L_2(f_1, f_2) = \sqrt{\int (f_1(\mathbf{x}) - f_2(\mathbf{x}))^2 dF(\mathbf{x})} \tag{2.34}$$

if the regression $r(\mathbf{x})$ does not belong to the set $f(\mathbf{x}, \alpha),\ \alpha \in \Lambda$.

Thus, the problem of estimating the regression may be also reduced to the scheme of minimizing a risk functional on the basis of a given set of sample data. Specifically, the empirical risk functional for the regression estimation problem has the following form:

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i, \alpha))^2,\ \alpha \in \Lambda \tag{2.35}$$

The specific feature of this problem is that the set of functions $Q(\mathbf{z}, \alpha),\ \alpha \in \Lambda$, is subject to the following restrictions:

- The vector $\mathbf{z}$ consists of $n + 1$ coordinates: the coordinate y and the n coordinates $(x^1, \ldots, x^n)$ which form the vector $\mathbf{x}$. However, in contrast to the pattern recognition problem, the coordinate y as well as the function $f(\mathbf{x}, a)$ may take any value in the interval $(-\infty, \infty)$
- The set of loss functions $Q(\mathbf{z}, \alpha),\ \alpha \in \Lambda$, is of the form $Q(\mathbf{z}, a) = (y - f(\mathbf{x}, \alpha))^2$.

2.4.4 Unsupervised Learning

If the data is only a sample of objects without associated target values, the problem is known as *unsupervised learning*. In unsupervised learning, there is no teacher. Hence, a concise description of the data can be a set of clusters or a probability density stating how likely it is to observe a certain object in the future. The primary objective of unsupervised learning is to extract some structure from a given sample of training objects.

The Problem of Density Estimation

Let $p(\mathbf{x}, \alpha),\ \alpha \in \Lambda$, be a set of probability densities containing the required density

$$p(\mathbf{x}, \alpha_0) = \frac{dF(\mathbf{x})}{d\mathbf{x}}. \tag{2.36}$$

It was shown that the minimum of the risk functional

$$R(\alpha) = \int \ln p(\mathbf{x}, \alpha) dF(\mathbf{x}) \tag{2.37}$$

(if it exists) is attained at the functions $p(\mathbf{x}, \alpha^*)$ which may differ from $p(\mathbf{x}, \alpha_0)$ only on a set of zero measure. Specifically, Bretagnolle and Huber [1] proved the following inequality

$$\int |p(\mathbf{x}, \alpha)) - p(\mathbf{x}, \alpha_0))| d\mathbf{x} \leq 2\sqrt{R(\alpha) - R(\alpha_0)} \tag{2.38}$$

according to which the problem of estimating the density in L_1 is reduced to the minimization of the functional (2.37) on the basis of empirical data. The general form of the L_p metric in a k-dimensional metric space is given by the following equation

$$\|x\|_p = \left(\sum_{i=1}^{k} |x_i|^p \right)^{\frac{1}{p}} \tag{2.39}$$

for $1 \leq p < \infty$ and $x \in \mathbb{R}^k$. In particular, the corresponding empirical risk functional has the following form

$$R_{emp}(\alpha) = -\sum_{i=1}^{l} \ln p(\mathbf{x}_i, \alpha) \tag{2.40}$$

The special feature of the density estimation problem is that the set of functions $Q(\mathbf{z}, \alpha)$ is subject to the following restrictions:

- The vector $\mathbf{z}$ coincides with the vector $\mathbf{x}$,
- The set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, is of the form $Q(\mathbf{z}, \alpha) = -\log p(\mathbf{x}, \alpha)$, where $p(\mathbf{x}, \alpha)$ is a set of density functions. The loss function $Q(\mathbf{z}, \alpha)$ takes on arbitrary values on the interval $(-\infty, \infty)$.

Clustering

A general way to represent data is to specify a similarity between any pair of objects. If two objects share much structure, is should be possible to reproduce the data from the same *prototype*. This is the primary idea underlying *clustering methods* which form a rich subclass of the unsupervised learning paradigm. Clustering is one of the most primitive mental activities of humans, which is used in order to handle the huge amount of information they receive every day. Processing every piece of information as a single entity would be impossible. Thus, humans tend to categorize entities (i.e. objects, persons, events) into clusters. Each cluster is then characterized by the common attributes of the entities it contains.

The definition of clustering leads directly to the definition of a single "cluster". Many definitions have been proposed over the years, but most of them are based on loosely defined terms such as "similar" and "alike" or are oriented to a specific kind of clusters. Therefore, the majority of the proposed definitions for clustering are of vague or of circular nature. This fact reveals that it is not possible to provide a universally accepted formal definition of clustering. Instead, one can only provide an intuitive definition stating that given a fixed number of clusters, the clustering procedure aims at finding a grouping of objects (*clustering*) such that similar objects will be assigned to same group (*cluster*). Specifically, if there exists a partitioning of the original data set such that the similarities of the objects in one cluster are much greater than the similarities among objects from different clusters, then it is possible to extract structure from the given data. Thus, it is possible to represent a whole cluster by one representative data point. More formally, by letting

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_l\} \tag{2.41}$$

be the original set of available data the m - *clustering* $\mathcal{R}$ of X may be defined as the partitioning of X into m sets (*clusters*) $C_1, \ldots, C_m$ such that the following three conditions are met:

- $C_i \neq \emptyset,\ i = 1, \ldots, m$
- $\cup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset,\ i \neq j,\ i, j = 1, \ldots, m$

It must be noted that the data points contained in a cluster C_i are more "similar" to each other and less similar to the data points of the other clusters. The quantification, however, of the terms "similar" and "dissimilar" is highly dependent on the type of the clusters involved. The type of the clusters is determinately affected by the shape of the clusters which in tern depends on the particular measure of *dissimilarity* or *proximity* between clusters.

2.4.5 Reinforcement Learning

Reinforcement learning is learning how to map situations to actions in order to maximize a numerical reward signal. The learner is not explicitly told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. In most interesting and challenging cases, actions may affect not only the immediate reward, but also the next situation and, through that, all subsequent rewards. These two characteristics (trial and error, and delayed reward) are the most important distinguishing features of reinforcement learning. Reinforcement learning does not define a subclass of learning algorithms, but rather a category of learning problems which focuses on designing learning agents which cope with real-life problems. The primary features of such problems involves the interaction of the learning agents with their environments in order to achieve a

particular goal. Clearly, this kind of agents must have the ability to sense the state of the environment to some extent and must be able to take actions affecting that state. The agent must also have a goal or goals relating to the state of the environment.

Reinforcement learning is different from the classical supervised learning paradigm where the learner is explicitly instructed by a knowledgable external supervisor through a series of examples that indicate the desired behavior. This is an important kind of learning, but it is not adequate on its own to address the problem of learning from interaction. In interactive problems, it is often impractical to obtain correct and representative examples of all the possible situations in which the agent has to act. In uncharted territory, where one would expect learning to be more beneficial, an agent must be able to learn from its own experience.

One of the challenges that arises in reinforcement learning and not in other kinds of learning is the tradeoff between exploration and exploitation. To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found effective in producing reward. However, the discovery of such actions requires that the agent has to try actions that he has not selected before. In other words, the agent has to *exploit* what is already known in order to obtain reward, but it is also important to *explore* new situations in order to make better action selections in the future. The dilemma is that neither exploitation nor exploration can be pursued exclusively without failing at the task. The agent must try a variety of actions and progressively favor those that appear to be best. Moreover, when the learning problems involves a stochastic task, each action must be tried many times in order to reliably estimate the expected reward.

Another key feature of reinforcement learning is that it explicitly considers the *whole* problem of a goal-directed agent interacting with an uncertain environment. This is in contrast with many learning approaches that address subproblems without investigating how they fit into a larger picture. Reinforcement learning, on the other hand, starts with a complete, interactive goal-seeking agent. All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments. Moreover, it is usually assumed from the beginning that the agent has to operate despite significant uncertainty about the environment it faces. For learning research to make progress, important subproblems have to be isolated and studied, but they should be incorporated in the larger picture as subproblems that are motivated by clear roles in complete, interactive, goal-seeking agents, even if all the details of the complete agent cannot yet be filled in.

2.5 Theoretical Justifications of Statistical Learning Theory

Statistical learning theory provides the theoretical basis for many of today's machine learning algorithms and is arguably one of the most beautifully developed branches of artificial intelligence in general. Providing the basis of new learning algorithms, however, was not the only motivation for the development of statistical learning theory. It was just as much a philosophical one, attempting to identify the fundamental

element which underpins the process of drawing valid conclusions from empirical data.

The best-studied problem in machine learning is the problem of classification. Therefore, the theoretical justifications concerning Statistical Learning Theory will be analyzed within the general context of supervised learning and specifically pattern classification. The pattern recognition problem, in general, deals with two kind of spaces: the input space $\mathbf{X}$, which is also called the space of *instances*, and the output space $\mathbf{Y}$, which is also called the *label* space. For example, if the learning task is to classify certain objects into a given, finite set of categories, then $\mathbf{X}$ consists of the space of all possible objects (instances) in a certain, fixed representation, while $\mathbf{Y}$ corresponds to the discrete space of all available categories such that $\mathbf{Y} = \{0, \ldots, k-1\}$. This discussion, however, will be limited to the case of binary classification for simplicity reasons which yields that the set of available categories will be restricted to $\mathbf{Y} = \{-1, +1\}$. Therefore, the problem of classification may be formalized as the procedure of estimating a functional dependence of the form $\phi : \mathbf{X} \rightarrow \mathbf{Y}$, that is a relationship between input and output spaces $\mathbf{X}$ and $\mathbf{Y}$ respectively. Moreover, this procedure is realized on the basis of a given set of *training examples* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$, that is pairs of objects with the associated category label. The primary goal when addressing the pattern classification problem is to find such a mapping that yields the smallest possible number of classification errors. In other words, the problem of pattern recognition is to find that mapping for which the number of objects in $\mathbf{X}$ that are assigned to the wrong category is as small as possible. Such a mapping is referred to as a *classifier*. The procedure for determining such a mapping on the basis of a given set of training examples is referred to as a *classification algorithm* or *classification rule*. A very important issue concerning the definition of the pattern recognition problem is that no particular assumptions are made on the spaces $\mathbf{X}$ and $\mathbf{Y}$. Specifically, it is assumed that there exists a *joint distribution function* F on $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ and that the training examples $(\mathbf{x}_i, y_i)$ are sampled independently from this distribution F. This type of sampling is often denoted as *iid* (independently and identically distributed) sampling.

It must be noted that any particular discrimination function $\phi(\mathbf{x})$ is parameterized by a unique parameter $\alpha \in \Lambda_{all}$ which can be anything from a single parameter value to a multidimensional vector. In other words, Λ_{all} denotes the set of all measurable functions from $\mathbf{X}$ to $\mathbf{Y}$ corresponding to the set of all possible classifiers for a given pattern recognition problem. Of particular importance is the so-called Bayes Classifier $\phi_{Bayes}(\mathbf{x})$, identified by the parameter α_{Bayes}, whose discrimination function has the following form

$$\phi_{Bayes}(\mathbf{x}, \alpha_{Bayes}) = \arg\min_{\omega \in \mathbf{Y}} P(Y = \omega | X = \mathbf{x}). \tag{2.42}$$

The Bayes classifier operates by assigning any given pattern to the class with the maximum a posteriori probability. The direct computation of the Bayes classifier, however, is impossible in practice since the underlying probability distribution is completely unknown to the learner. Therefore, the problem of pattern recognition may be

formulated as the procedure of constructing a function $\phi(\mathbf{x}, \alpha) : \mathbf{X} \rightarrow \mathbf{Y}$, uniquely determined by the parameter α, through a series of training points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$ which has risk $R(\alpha)$ as close as possible to the risk $R(\alpha_{Bayes})$ of the Bayes classifier.

2.5.1 Generalization and Consistency

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$ be a sequence of training patterns and α_l be the function parameter corresponding to the classifier obtained by the utilization of some learning algorithm on the given training set. Even though it is impossible to compute the true underlying risk $R(\alpha_l)$ for this classifier according to Eq. (2.27), it is possible to estimate the empirical risk $R_{emp}(\alpha_l)$ according to Eq. (2.29) accounting for the number of errors on the training points.

Usually, for a classifier α_n trained on a particular training set, the empirical risk $R_{emp}(\alpha_l)$ is relatively small since otherwise the learning algorithm will not even seem to be able to explain the training data. A natural question arising at this point is whether a function α_l which makes a restricted number of errors on the training set will perform likewise on the rest of the $\mathbf{X}$ space. This question is intimately related to the notion of *generalization*. Specifically, a classifier α_l is said to generalize well if the difference $|R(\alpha_l) - R_{emp}(\alpha_l)|$ is small. This definition, however, does not imply that the classifier α_l will have a small overall error R_{emp}, but it just means that the empirical error $R_{emp}(\alpha_l)$ is a good estimate of the true error $R(\alpha_l)$. Particularly bad in practice is the situation where $R_{emp}(\alpha_l)$ is much smaller than $R(\alpha_l)$ misleading to the assumption of being overly optimistic concerning the quality of the classifier.

The problem concerning the generalization ability of a given machine learning algorithm may be better understood by considering the following regression example. One is given a set of observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in \mathbf{X} \times \mathbf{Y}$, where for simplicity it is assumed that $\mathbf{X} = \mathbf{Y} = \mathbb{R}$. Figure 2.3 shows a plot of such a dataset, indicated by the round points, along with two possible functional dependencies that could underly the data.

The dashed line α_{dashed} represents a fairly complex model that fits the data perfectly resulting into a zero training error. The straight line, on the other hand, does not completely explain the training data, in the sense that there are some residual errors, leading to a small training error. The problem regarding this example concerns the inability to compute the true underlying risks $R(\alpha_{dashed})$ and $R(\alpha_{straight})$ since the two possible functional dependencies have very different behavior. For example, if the straight line classifier $\alpha_{straight}$ was the true underlying risk, then the dashed line classifier α_{dashed} would have a high true risk, as the L_2 distance between the true and the estimated function is very large. The same also holds when the true functional dependence between the spaces $\mathbf{X}$ and $\mathbf{Y}$ is represented by the dashed line while the straight line corresponds to the estimated functional dependency. In both cases, the true risk would be much higher than the empirical risk.

This example emphasizes the need to make the correct choice between a relatively complex function model, leading to a very small training error, and a simpler function

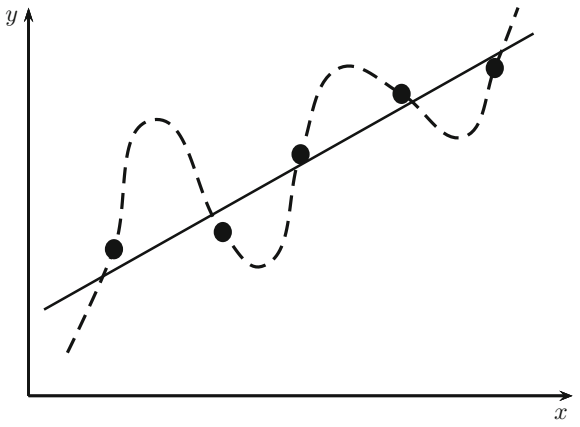

Fig. 2.3 Regression example

model at the cost of a slightly higher training error. In one form or another, this issue was extensively studied within the context of classical statistics as the *bias-variance* dilemma. The bias-variance dilemma involves the following dichotomy. If a linear fit is computed for any given data set, then every functional dependence discovered would be linear but as a consequence of the *bias* imposed from the choice of the linear model which does not necessarily comes from the data. On the hand, if a polynomial model of sufficiently high degree is fit for any given data set, then the approximation ability of the model would fit the data perfectly but it would suffer from a large variance depending on the initial accuracy of the measurements. In other words, within the context of applied machine learning, complex explanations show *overfitting*, while overly simple explanations imposed by the learning machine design lead to *underfitting*. Therefore, the concept of generalization can be utilized in order to determine the amount of increase in the training error in order to tolerate for a fitting a simpler model and quantify the way in which a given model is simpler than another one.

Another concept, closely related to generalization, is the one of consistency. However, as opposed to the notion of generalization discussed above, consistency is not a property of an individual function, but a property of a set of functions. The notion of consistency, as it is described in classical statistics, aims at making a statement about what happens in the limit of infinitely many sample points. Intuitively, it seems reasonable to request that a learning algorithm, when presented with more and more training points, should eventually converge to an optimal solution.

Given any particular classification algorithm and a set of l training points, α_l denotes the parameter identifying the obtained classifier where the exact procedure for its determination is not of particular importance. Note that any classification algorithm chooses its functions from some particular function space identified by the complete parameter space Λ such that $\mathcal{F} = \{\phi(\mathbf{x}, \alpha) : \alpha \in \Lambda\}$. For some algorithms this space is given explicitly, while for others it only exists implicitly via the mechanism of the algorithm. No matter how the parameter space Λ is defined,

the learning algorithm attempts to choose the parameter $\alpha_l \in \Lambda$ which it considers as the best classifier in Λ, based on the given set of training points. On the other hand, in theory the best classifier in Λ is the one that has the smallest risk which is uniquely determined by the following equation:

$$\alpha_\Lambda = \arg\min_{\alpha \in \Lambda} R(\alpha). \tag{2.43}$$

The third classifier of particular importance is the Bayes classifier α_{Bayes} introduced in Eq. (2.42). Bayes classifier, while being the best existing classifier, it may be not be included within the parameter space Λ under consideration, so that $R(\alpha_\Lambda) > R(\alpha_{Bayes})$.

Let $(\mathbf{x}_i, y_i)_{i \in \mathbb{N}}$ be an infinite sequence of training points which have been drawn independently from some probability distribution P and, for each $l \in \mathbb{N}$, let α_l be a classifier constructed by some learning algorithm on the basis of the first l training points. The following types of consistency may be defined:

1. The learning algorithm is called *consistent with respect to Λ and P* if the risk $R(\alpha_l)$ converges in probability to the risk $R(\alpha_\Lambda)$ of the best classifier Λ, that is for all $\epsilon > 0$,
$$P(R(\alpha_l) - R(\alpha_\Lambda) > \epsilon) \to 0 \text{ as } n \to \infty \tag{2.44}$$

2. The learning algorithm is called *Bayes-consistent with respect to P* if the risk $R(R(\alpha_l))$ converges to the risk $R(\alpha_{Bayes})$ of the Bayes classifier, that is for all $\epsilon > 0$,
$$P(R(\alpha_l) - R(\alpha_{Bayes}) > \epsilon) \to 0 \text{ as } n \to \infty \tag{2.45}$$

3. The learning algorithm is called universally consistent with respect to Λ (*resp universally Bayes-consistent*) if it is consistent with respect to Λ (*reps Bayes-consistent*) for all probability distributions P.

It must be noted that none of the above definitions involves the empirical risk $R_{emp}(\alpha_l)$ of a classifier. On the contrary, they exclusively utilize the true risk $R(\alpha_l)$ as a quality measure reflecting the need to obtain a classifier which is as good as possible. The empirical risk constitutes the most important estimator of the true risk of a classifier so that the requirement involving the convergence of the true risk $(R(\alpha_l) \to R(\alpha_{Bayes}))$ should be extended to the convergence of the empirical risk $(R_{emp}(\alpha_l) \to R(\alpha_{Bayes}))$.

2.5.2 Bias-Variance and Estimation-Approximation Trade-Off

The goal of classification is to get a risk as close as possible to the risk of the Bayes classifier. A natural question that arises concerns the possibility of choosing the complete parameter space Λ_{all} as the parameter space Λ utilized by a particular

classifier. This question raises the subject of whether the selection of the overall best classifier, obtained in the sense of the minimum empirical risk,

$$\alpha_l = \arg \min_{\alpha \in \Lambda_{all}} R_{emp}(\alpha) \tag{2.46}$$

implies consistency. The answer for this question is unfortunately negative since the optimization of a classifier over too large parameter (function) spaces, containing all the Bayes classifiers for all probability distributions P, will lead to inconsistency. Therefore, in order to learn successfully it is necessary to work with a smaller parameter (function) space Λ.

Bayes consistency deals with the convergence of the term $R(\alpha_l) - R(\alpha_{Bayes})$ which can be decomposed in the following form:

$$R(\alpha_l) - R(\alpha_{Bayes}) = \underbrace{(R(\alpha_l) - R(\alpha_\Lambda))}_{estimation \quad error} + \underbrace{(R(\alpha_\Lambda) - R(\alpha_{Bayes}))}_{approximation \quad error} \tag{2.47}$$

The first term on the right hand side is called the *estimation error* while the second term is called the *approximation error*. The first term deals with the uncertainty introduced by the random sampling process. That is, given a finite sample, it is necessary to estimate the best parameter (function) in Λ. Of course, in this process there will be a hopefully small number of errors which is identified by the term estimation error. The second term, on the other hand, is not influenced by random qualities. It particularly deals with the error made by looking for the best parameter (function) in a small parameter (function) space Λ, rather than looking for the best parameter (function) in the entire space Λ_{all}. Therefore, the fundamental question in this context is how well parameters (functions) in Λ can be used to approximate parameters (functions) in Λ_{all}.

In statistics, the estimation error is also called the *variance*, and the approximation error is called the *bias* of an estimator. The first term measures the variation of the risk of the function corresponding to the parameter α_l estimated on the sample, while the second one measures the bias introduced in the model by choosing a relatively small function class.

In this context the parameter space Λ may be considered as the means to balance the trade-off between estimation and approximation error. This is particularly illustrated in Fig. 2.4 which demonstrates that the selection of a very large parameter space Λ yields a very small approximation error term since there is high probability that the Bayes classifier will be contained in Λ or at least it can be closely approximated by some element in Λ. The estimation error, however, will be rather large in this case since the space Λ will contain more complex functions which will lead to overfitting. The opposite effect will happen if the function class corresponding to the parameter space Λ is very small.

The trade-off between estimation and approximation error is explicitly depicted in Fig. 2.5. According to the graph, when the parameter space Λ corresponds to a small complexity function space utilized by the classification algorithm, then the

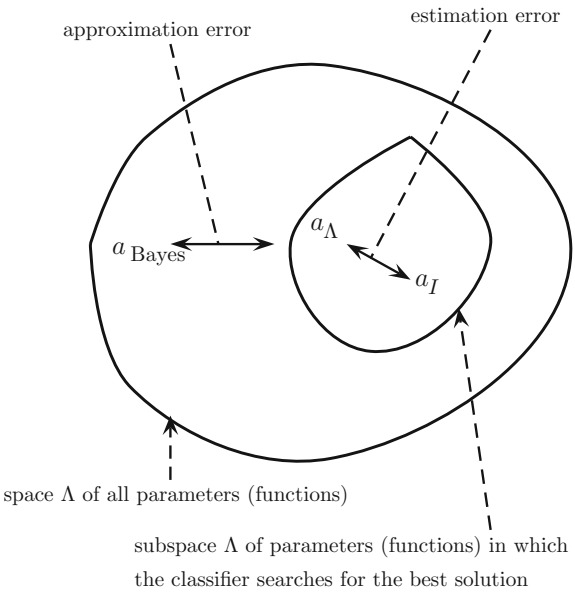

Fig. 2.4 Illustration of estimation and approximation error

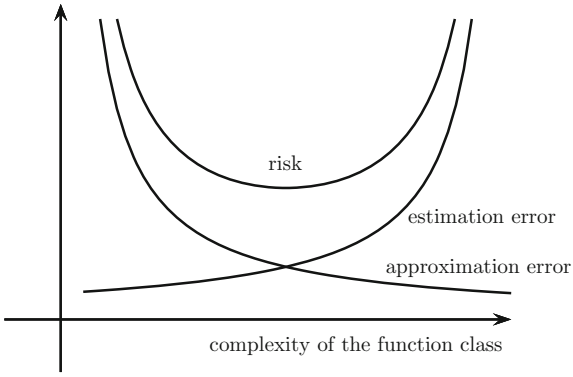

Fig. 2.5 Trade-off between estimation and approximation error

estimation error will be small but the approximation error will be large (underfitting). On the other hand, if the complexity of Λ is large, then the estimation error will also be large, while the approximation error will be small (overfitting). The best overall risk is achieved for "moderate" complexity.

2.5.3 Consistency of Empirical Minimization Process

As discussed in Sect. 2.4.2, the ERM principle provides a more powerful way of classifying data since it is impossible to directly minimize the true risk functional

given by Eq. (2.27). In particular, the ERM principle addresses the problem related with the unknown probability distribution function $F(\omega, \mathbf{x})$ which underlies the data generation process by trying to infer a function $f(\mathbf{x}, \alpha)$ from the set of identically and independently sampled training data points. The process of determining this function is based on the minimization of the so-called empirical risk functional which for the problem of pattern classification is given by Eq. (2.29).

The fundamental underpinning behind the principle of Empirical Risk Minimization is the law of large numbers which constitutes one of the most important theorems in statistics. In its simplest form it states that under mild conditions, the mean of independent, identically-distributed random variables ξ_i, which have been drawn from some probability distribution function P of finite variance, converges to the mean of the underlying distribution itself when the sample size goes to infinity:

$$\frac{1}{l}\sum_{i=1}^{l}\xi_i \rightarrow E(\xi)\, for\ l \rightarrow \infty. \tag{2.48}$$

A very important extension to the law of large numbers was originally provided by the Chernoff inequality (Chernoff 1952) which was subsequently generalized by Hoeffding (Hoeffding 1963). This inequality characterizes how well the empirical mean approximates the expected value. Namely, if ξ_i, are random variables which only take values in the [0, 1] interval, then

$$P\left(\left|\frac{1}{l}\sum_{i=1}^{l}\xi_i - E(\xi)\right| \geq \epsilon\right) \leq \exp(-2l\epsilon^2). \tag{2.49}$$

This theorem can be applied to the case of the empirical and the true risk providing a bound which states how likely it is that for a given function, identified by the parameter α, the empirical risk is close to the actual risk:

$$P(|R_{emp}(\alpha) - R(\alpha)| \geq \epsilon) \leq \exp(-2l\epsilon^2) \tag{2.50}$$

The most important fact concerning the bound provided by Chernoff in Eq. (2.50) is its probabilistic nature. Specifically, it states that the probability of a large deviation between the test error and the training error of a function $f(\mathbf{x}, \alpha)$ is small when the sample size is sufficiently large. However, by not ruling out the presence of cases where the deviation is large, it just says that for a fixed function $f(\mathbf{x}, \alpha)$, this is very unlikely to happen. The reason why this has to be the case is the random process that generates the training samples. Specifically, in the unlucky cases when the training data are not representative of the true underlying phenomenon, it is impossible to infer a good classifier. However, as the sample size gets larger, such unlucky cases become very rare. Therefore, any consistency guarantee can only be of the form "the empirical risk is close to the actual risk, with high probability".

Another issue related to the ERM principle is that the Chernoff bound in (Eq. 2.50) is not enough in order to prove the consistency of the ERM process. This is true since the Chernoff inequality holds only for a fixed function $f(\mathbf{x}, \alpha)$ which does not depend on the training data. While this seems to be a subtle mathematical difference, this is where the ERM principle can go wrong as the classifier α_l does depend on the training data.

2.5.4 Uniform Convergence

It turns out the conditions required to render the ERM principle consistent involve *restricting the set of admissible functions*. The main insight provided by the VC theory is that the consistency of the ERM principle is determined by the *worst case* behavior over all functions $f(\mathbf{x}, \alpha)$, where $\alpha \in \Lambda$, that the learning machine could use. This worst case corresponds to a version of the law of large numbers which is *uniform* over all functions parameterized by Λ.

A simplified description of the uniform law of large numbers which specifically relates to the consistency of the learning process is given in Fig. 2.6. Both the empirical and the actual risk are plotted as functions of the α parameter and the set of all possible functions, parameterized by the set Λ, is represented by a single axis of the plot for simplicity reasons. In this context, the ERM process consists of picking the parameter α that yields the minimal value of R_{emp}. This process is consistent if the minimum of R_{emp} converges to that of R as the sample size increases. One way to ensure the convergence of the minimum of all functions in Λ is uniform convergence over Λ. Uniform convergence over Λ requires that for all functions $f(\mathbf{x}, \alpha)$, where $\alpha \in \Lambda$, the difference between $R(\alpha)$ and $R_{emp}(\alpha)$ should become small *simultaneously*. In other words, it is required that there exists some large l such that for sample size at least n, it is certain that for all functions $f(\mathbf{x}, \alpha)$, where $\alpha \in \Lambda$, the difference

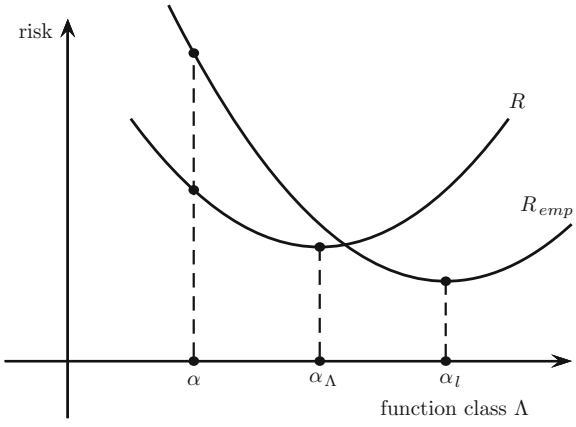

Fig. 2.6 Convergence of the empirical risk to the actual risk

$|R(\alpha) - R_{emp}(\alpha)|$ is smaller than a given ϵ. Mathematically, this statement can be expressed using the following inequality:

$$\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| \leq \epsilon. \tag{2.51}$$

In Fig. 2.6, this means that the two plots of R and R_{emp} become so close that their distance is never larger than ϵ. This, however, does not imply that in the limit of infinite sample sizes, the *minimizer* of the empirical risk, α_l, will lead to a value of the risk that is as good as the risk of the best function, α_Λ, in the function class. The latter is true when uniform convergence is imposed over all functions that are parameterized by Λ. Intuitively it is clear that if it was known that for all functions $f(\mathbf{x}, \alpha)$, where $\alpha \in \Lambda$, the difference $|R(\alpha) - R_{emp}(\alpha)|$ is small, then this holds in particular for any function identified by the parameter α_l that might have been chosen based on the given training data. That is, for any function $f(\mathbf{x}, a)$, where $\alpha \in \Lambda$, it is true that:

$$|R(\alpha) - R_{emp}(\alpha)| \leq \sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)|. \tag{2.52}$$

Inequality (2.52) also holds for any particular function parameter α_l which has been chosen on the basis of a finite sample of training points. Therefore, the following conclusion can be drawn:

$$P(|R(\alpha) - R_{emp}(\alpha)| \geq \epsilon) \leq P(\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| \geq \epsilon), \tag{2.53}$$

where the quantity on the right hand side represents the very essence of the uniform law of large numbers. In particular, the law of large numbers is said to uniformly hold over a function class parameterized by Λ if for all $\epsilon > 0$,

$$P(\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| \geq \epsilon) \to 0 \text{ as } l \to \infty. \tag{2.54}$$

Inequality (2.52) can be utilized in order to show that if the uniform law of large numbers holds for some function class parameterized by Λ, then the ERM is consistent with respect to Λ. Specifically, Inequality (2.52) yields that:

$$|R(\alpha_l) - R(\alpha_\Lambda| \leq 2 \sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)|, \tag{2.55}$$

which finally concludes:

$$P(|R(\alpha_l) - R(\alpha_\Lambda| \geq \epsilon) \leq P(\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| \geq \frac{\epsilon}{2}). \tag{2.56}$$

The right hand side of Inequality (2.56) tends to 0, under the uniform law of large numbers, which then leads to consistency of the ERM process with respect to the underlying function class parameterized by Λ. Vapnik and Chervonenkis [5]

proved that uniform convergence as described by Inequality (2.54) is a necessary and sufficient condition for the consistency of the ERM process with respect to Λ. It must be noted that the condition of uniform convergence crucially depends on the set of functions for which it must hold. Intuitively, it seems clear that the larger the function space parameterized by Λ, the larger the quantity $\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)|$. Thus, the larger Λ, the more difficult it is to satisfy the uniform law of large numbers. That is, for larger function spaces (corresponding to larger parameter spaces Λ) consistency is harder to achieve than for smaller function spaces. This abstract characterization of consistency as a uniform convergence property, whilst theoretically intriguing, is not at all that useful in practice. This is true, since in practice it is very difficult to infer whether the uniform law of large numbers holds for a given function space parameterized by Λ. Therefore, a natural question that arises at this point is whether there are properties of function spaces which ensure uniform convergence of risks.

2.5.5 Capacity Concepts and Generalization Bounds

Uniform convergence was referred to as the fundamental property of a function space determining the consistency of the ERM process. However, a closer look at this convergence is necessary in order to make statements concerning the behavior of a learning system when it is exposed to a limited number of training samples. Therefore, attention should be focused on the probability

$$P(\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)| > \epsilon) \tag{2.57}$$

which will not only provide insight into which properties of function classes determines the consistency of the ERM process, but will also provide bounds on the risk. Along this way, two notions are of primary importance, namely:

1. the *union bound* and
2. the method of *symmetrization* by a ghost sample.

The Union Bound

The union bound is a simple but convenient tool in order to transform the standard law of large numbers of individual functions into a uniform law of large numbers over a set of finitely many functions parameterized by a set $\Lambda = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$. Each of the functions $\{f(\mathbf{x}, \alpha_i) : \alpha_i \in \Lambda\}$, satisfies the standard law of large numbers in the form of a Chernoff bound provided by Inequality (2.50), that is

$$P(|R(\alpha_i) - R_{emp}(\alpha_i)| \geq \epsilon) \leq 2\exp(-2l\epsilon^2). \tag{2.58}$$

In order to transform these statements about the individual functions $\{f(x, \alpha_i) : \alpha_i \in \Lambda\}$ into a uniform law of large numbers the following derivations are necessary

$$P(\sup_{\alpha\in\Lambda}|R(\alpha)-R_{emp}(\alpha)|>\epsilon)=P\Big(\bigcup_{i=1}^{m}|R(\alpha_i)-R_{emp}(\alpha_i)|>\epsilon\Big)$$
$$\leq\sum_{i=1}^{m}P(|R(\alpha_i)-R_{emp}(\alpha_i)|>\epsilon)$$
$$\leq 2m\exp(-2l\epsilon^2) \tag{2.59}$$

It is clear that the difference between the Chernoff bound given by Inequality (2.50) and the right hand side of Eq. (2.59) is just a factor of m. Specifically, if the function space $\mathcal{F}=\{f(\mathbf{x},\alpha_i):\alpha_i\in\Lambda\}$ is fixed, this factor can be regarded as a constant and the term $2m\exp(-2l\epsilon^2)$ still converges to 0 as $l\to\infty$. Hence, the empirical risk converges to 0 uniformly over $\mathcal{F}$ as $l\to\infty$. Therefore, it is proved that an ERM process over a finite set Λ of function parameters is consistent with respect to Λ.

Symmetrization

Symmetrization is an important technical step towards using capacity measures of function classes. Its main purpose is to replace the event $\sup_{\alpha\in\Lambda}|R(\alpha)-R_{emp}(\alpha)|$ by an alternative event which can be solely computed on a given sample size. Assume that a new *ghost sample* $\{(\mathbf{x}_i',y_i')\}_{i=1}^{l}$ is added to the initial training sample $\{(\mathbf{x}_i,y_i)\}_{i=1}^{l}$. The ghost sample is just another sample which is also drawn iid from the same underlying distribution and which is independent of the first sample. The ghost sample, however, is a mathematical tool that is not necessary to be physically sampled in practice. It is just an auxiliary set of training examples where the corresponding empirical risk will be denoted by $R'_{emp}(\alpha)$. In this context, Vapnik and Chervonenkis [6] proved that for $m\epsilon^2\geq 2$

$$P(\sup_{\alpha\in\Lambda}|R(\alpha)-R_{emp}(\alpha)|>\epsilon)\leq 2P\Big(\sup_{\alpha\in\Lambda}|R_{emp}(\alpha)-R'_{emp}(\alpha)|>\frac{\epsilon}{2}\Big). \tag{2.60}$$

Here, the first P refers to the distribution of an iid sample l, while the second one refers to the distribution of two samples of size l, namely the original sample and the ghost one which form an iid sample of size $2l$. In the latter case, R_{emp}, measures the empirical loss on the first half of the sample, and R'_{emp} on the second half. This statement is referred to as the *symmetrization lemma* referring to the fact that the attention is focused on an event which depends on a symmetric way on a sample of size l. Its meaning is that if the empirical risks of two independent l-samples are close to each other, then they should also be close to the true risk. The main purpose of this lemma is to provide a way to bypass the need to directly estimate the quantity $R(\alpha)$ by computing the quantity $R'_{emp}(\alpha)$ on a finite sample size.

In the previous section, the uniform bound was utilized as a means to constraint the probability of uniform convergence in terms of a probability of an event referring to a finite function class. The crucial observation is now that even if Λ parameterizes an infinite function class, the different ways in which it can classify a training set of l sample points is finite. Namely, for any given training point in the training sample, a

function can take only values within the set $\{-1, +1\}$ which entails that on a sample of l points $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$, a function can act in at most 2^l different ways. Thus, even for an infinite function parameter class Λ, there are at most 2^l different ways the corresponding functions can classify the l points of finite sample. This means that when considering the term $\sup_{\alpha \in \Lambda} |R_{emp}(\alpha) - R'_{emp}(\alpha)|$, the supremum effectively runs over a finite set of function parameters. In this context, the supremum over Λ on the right hand side of Inequality (2.60) can be replaced by the supremum over a finite function parameter class with at most 2^{2l} function parameters. This number comes as a direct consequence from the fact that there is a number of $2l$ sample points for both the original and the ghost samples.

The Shattering Coefficient

For the purpose of bounding the probability in Eq. (2.57), the symmetrization lemma implies that the function parameter class Λ is effectively finite since it can be restricted to the $2l$ points appearing on the right hand side of Inequality (2.60). Therefore, the function parameter class contains a maximum number of 2^{2l} elements. This is because only the values of the functions on the sample points and the ghost sample points count. In order to formalize this, let $Z_l = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$ be a given sample of size l and let $|\Lambda_{Z_l}|$ be the cardinality of Λ when restricted to $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$, that is, the number of function parameters from Λ that can be distinguished from their values on $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$. Moreover, let $\mathcal{N}(\Lambda, l)$ be the maximum number of functions that can be distinguished in this way, where the maximum runs over all possible choices of samples, so that

$$\mathcal{N}(\Lambda, l) = \max \{|\Lambda_{Z_l}| : \mathbf{x}_1, \ldots, \mathbf{x}_l \in \mathbf{X}\}. \tag{2.61}$$

The quantity $\mathcal{N}(\Lambda, l)$ is referred to as the *shattering coefficient of the function class parameterized by Λ with respect to the sample size l*. It has a particularly simple interpretation: it is the number of different outputs $\{y_1, \ldots, y_l\}$ that the functions parameterized by Λ can achieve on samples of a given size l. In other words, it measures the *number of ways that the function space can separate the patterns into two classes*. Whenever, $\mathcal{N}(\Lambda, l) = 2^l$, this means that there exists a sample of size l on which all possible separations can be achieved by functions parameterized by Λ. In this case, the corresponding function space is said to *shatter* l points. It must be noted that because of the maximum in the definition of $\mathcal{N}(\Lambda, l)$, shattering means that there *exists* a sample of l patterns which can be shattered in all possible ways. This definition, however, does not imply that all possible samples of size l will be shattered by the function space parameterized by Λ. The shattering coefficient can be considered as a capacity measure for a class of functions in the sense that it measures the "size" of a function class in a particular way. This way involves counting the number of functions in relation to a given sample of finite training points.

Uniform Convergence Bounds

Given an arbitrary, possibly infinite, class of function parameters consider the right hand side of Inequality (2.60) where the sample of $2l$ points will be represented by a

set Z_{2l}. Specifically, the set Z_{2l} may be interpreted as the combination of l points from the original sample and l points from the ghost sample. The main idea is to replace the supremum over Λ by the supremum over $\Lambda_{Z_{2l}}$ where the set Z_{2l} contains at most $\mathcal{N}(\Lambda, l) \leq 2^{2l}$ different functions, then apply the union bound on this finite set and then the Chernoff bound. This leads to a bound as in Inequality (2.59), with $\mathcal{N}(\Lambda, l)$ playing the role of m. Essentially, those steps can be written down as follows:

$$\begin{aligned} P(\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)| > \epsilon) &\leq 2P\Big(\sup_{\alpha\in\Lambda} |R_{emp}(\alpha) - R'_{emp}(\alpha)| > \frac{\epsilon}{2}\Big) \\ &= 2P\Big(\sup_{\alpha\in\Lambda_{Z_{2l}}} |R_{emp}(\alpha) - R'_{emp}(\alpha)| > \frac{\epsilon}{2}\Big) \\ &\leq 2\mathcal{N}(\Lambda, 2l) \exp\Big(\frac{-l\epsilon^2}{4}\Big), \end{aligned} \tag{2.62}$$

yielding the following inequality

$$P(\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)| > \epsilon) \leq 2\mathcal{N}(\Lambda, 2l) \exp\Big(\frac{-l\epsilon^2}{4}\Big). \tag{2.63}$$

The notion of uniform bound may be utilized in order to infer whether the ERM process is consistent for a given class of function parameters Λ. Specifically, the right hand side of Inequality (2.63) guarantees that the ERM process is consistent for a given class of function parameters Λ when it converges to 0 as $l \to \infty$. In this context, the most important factor controlling convergence is the quantity $\mathcal{N}(\Lambda, 2l)$. This is true since the second factor of the product $2\mathcal{N}(\Lambda, 2l) \exp(\frac{-l\epsilon^2}{4})$ is always the same for any given class of function parameters. Therefore, when the shattering coefficient is considerably smaller than 2^{2l}, say $\mathcal{N}(\Lambda, 2l) \leq (2n)^k$, it is easy to derive that the right hand side of the uniform bound takes the form

$$2\mathcal{N}(\Lambda, 2l) \exp\Big(\frac{-l\epsilon^2}{4}\Big) = 2\exp\Big(k\log(2l) - l\frac{\epsilon^2}{4}\Big), \tag{2.64}$$

which converges to 0 as $l \to \infty$. On the other hand, when the class of function parameters coincides with the complete parameter space Λ_{all} then the shattering coefficient takes its maximum value such that $\mathcal{N}(\Lambda, 2l) = 2^{2l}$. This entails that the right hand side of Inequality (2.63) takes the form

$$2\mathcal{N}(\Lambda, 2l) \exp\Big(\frac{-l\epsilon^2}{4}\Big) = 2\exp\Big(l(2\log(2)) - \frac{\epsilon^2}{4}\Big), \tag{2.65}$$

which does not converge to 0 as $l \to \infty$.

The union bound, however, cannot directly guarantee the consistency of the ERM process when utilizing the complete parameter space Λ_{all}. The reason is that Inequality (2.63) gives an upper bound on $P(\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)| > \epsilon)$ which merely provides a sufficient condition for consistency but not a necessary one. According to

(Devroye et al. 1996) a necessary and sufficient condition for the consistency of the ERM process is that

$$\log \frac{\mathcal{N}(\Lambda, 2l)}{l} \to 0. \tag{2.66}$$

2.5.6 Generalization Bounds

Sometimes it is useful to reformulate the uniform convergence bound so that the procedure of initially fixing ϵ and subsequently computing the probability that the empirical risk deviates from the true risk more than ϵ is reversed. In other words, there are occasions when it would be reasonable to initially specify the probability of the desired bound and then get a statement concerning the proximity between the empirical and the true risk. This can be achieved by setting the right hand side of Inequality (2.63) equal to some $\delta > 0$ and then solving for ϵ. The resulting statement declares that with probability at least $1-\delta$, any function in $\{f(\mathbf{x}, a) : \alpha \in \Lambda\}$ satisfies

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{4}{l}(\log(2\mathcal{N}(\Lambda, 2l) - \log(\delta)))}. \tag{2.67}$$

Consistency bounds can also be derived by utilizing Inequality (2.67). In particular, it is obvious that the ERM process is consistent for a given function class parameterized by Λ when the term $\frac{\sqrt{\log(2\mathcal{N}(\Lambda,2l))}}{l}$ converges to 0 as $l \to \infty$. The most important aspect concerning the generalization bound provided by Inequality (2.67) is that it holds for any function in $\{f(\mathbf{x}, a) : \alpha \in \Lambda\}$. This constitutes a highly desired property since the bound holds in particular for the function which minimizes the empirical risk, identified by the function parameter α_l. On the other hand, the bound holds for learning machines that do not truly minimize the empirical risk. This is usually interpreted as a negative property since by taking into account more information about a function, one could hope to obtain more accurate bounds.

Essentially, the generalization bound states that when both $R_{emp}(\alpha)$ and the square root term are small simultaneously then it is highly probable that the error on future points (actual risk) will be small. Despite sounding like a surprising statement this claim does not involve anything impossible. It only says that the utilization of a function class $\{f(\mathbf{x}, a) : \alpha \in \Lambda\}$ with relatively small $\mathcal{N}(\Lambda, l)$, which can nevertheless explain data sampled from the problem at hand, is not likely to be a coincidence. In other words, when a relatively small function class happens to "explain" data sampled from the problem under consideration, then there is a high probability that this function class captures some deeper aspects of the problem. On the other hand, when the problem is too difficult to learn from the given amount of training data, then it is necessary to use a function class so large that can "explain" nearly everything. This results in a small empirical error but at the same time increases the magnitude of the square root term. Therefore, according to the insight provided by the generalization bound, the difficulty of a particular learning problem is entirely determined by the

suitability of the selected function class and by the prior knowledge available for the problem.

The VC Dimension

So far, the various generalization bounds were expressed in terms of the shattering coefficient $\mathcal{N}(\Lambda, l)$. Their primary downside is that they utilize capacity concepts that are usually difficult to evaluate. In order to avoid this situation, Vapnik and Chervonenkis [2] introduced the so-called VC dimension which is one of the most well known capacity concepts. Its primary purpose is to characterize the growth behavior of the shattering coefficient using a single number.

A sample of size l is said to be *shattered by the function parameter class* Λ if this class parameterizes functions that can realize any labelling on the given sample, that is $|\Lambda_{Z_l}| = 2^l$. The VC dimension of Λ, is now defined as the largest number l such that there exists a sample of size l which is shattered by the functions parameterized by Λ. Formally,

$$VC(\Lambda) = \max\{l \in \mathbb{N} : |\Lambda_{Z_l}| = 2^l \text{ for some } Z_l\} \tag{2.68}$$

If the maximum does not exist, the VC dimension is defined to be infinity. For example, the VC dimension of the set of *liner indicator functions*

$$Q(\mathbf{z}, \alpha) = \theta\Big\{\sum_{p=1}^{l} a_p z_p + a_0\Big\} \tag{2.69}$$

in l-dimensional coordinate space $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_l)$ is equal to $l + 1$, since using functions from this set one can shattered at most $l + 1$ vectors. Moreover, the VC dimension of the set of linear functions

$$Q(\mathbf{z}, \alpha) = \sum_{p=1}^{l} a_p z_p + a_0 \tag{2.70}$$

in l-dimensional coordinate space $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_l)$ is equal to $l + 1$, since a linear function can shatter at most $l + 1$ points.

A combinatorial result proved simultaneously by several people [3, 4, 7] characterizes the growth behavior of the shattering coefficient and relates it to the VC dimension. *Let Λ be a function parameter class with finite VC dimension d. Then*

$$\mathcal{N}(\Lambda, l) \leq \sum_{i=0}^{d} \binom{n}{i} \tag{2.71}$$

for all $l \in \mathbb{N}$. In particular, for all $l \geq d$ the following inequality holds

$$\mathcal{N}(\Lambda, l) \leq \left(\frac{en}{d}\right)^d. \tag{2.72}$$

The importance of this statement lies in the last fact. If $l \geq d$, then the shattering coefficient behaves like a polynomial function of the sample size l. According to this result, when the VC dimension of a function parameter class is finite then the corresponding shattering coefficients will grow polynomially with l. Therefore, *the ERM process is consistent with respect to a function parameter space Λ if and only if $VC(\Lambda)$ is finite.*

A fundamental property shared by both the shattering coefficient and the VC dimension is that they do not depend on the underlying probability distribution P, since they only depend on the function parameter class Λ. On the one hand, this is an advantage, as the capacity concepts apply to all possible probability distributions. On the other hand, this can be considered as a disadvantage, as the capacity concepts do not take into account particular properties of the distribution at hand.

A particular class of distribution independent bounds is highly related with the concept of Structural Risk Minimization. Specifically, these bounds concern the subset of totally bounded functions

$$0 \leq Q(\mathbf{z}, \alpha) \leq B, \ \alpha \in \Lambda \tag{2.73}$$

with finite VC dimension such as the set of indicator functions. The main result for this set of functions is the following theorem: *With probability at least $1 - \delta$, the inequality*

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\epsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}\alpha}{B\epsilon}}\right) \tag{2.74}$$

holds true simultaneously for all functions of the set (2.73) *where*

$$\epsilon = 4\frac{d(\ln \frac{2l}{d} + 1) - \ln \delta}{l} \tag{2.75}$$

and $B = 1$.

The Structural Risk Minimization Principle

The ERM process constitutes a fundamental learning principle which efficiently deals with problems involving training samples of large size. This fact is specifically justified by considering Inequality (2.74) which formulates the conditions that guarantee the consistency of the ERM process. In other words, when the ratio l/d is large, the second summand on the right hand side of Inequality (2.74) will be small. The actual risk is then close to the value of the empirical risk. In this case, a small value of the empirical risk ensures a small value of the actual risk. On the other hand, when the ratio l/d is small, then even a small value for the empirical risk will not guarantee a small value for the actual risk. The latter case indicates the necessity for a new learning principle which will focus on acquiring a sufficiently small value for the actual risk $R(\alpha)$ by simultaneously minimizing both terms on the right hand side of Inequality (2.74). This is the basic underpinning behind the principle of Structural Risk Minimization (SRM). In particular, SRM is intended to

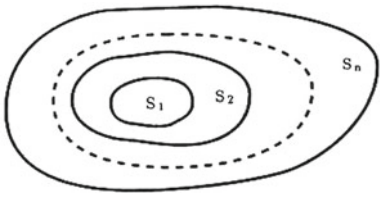

Fig. 2.7 Admissible structure of function sets

minimize the risk functional $R(\alpha)$ with respect to both the empirical risk and the VC dimension of the utilized set of function parameters Λ.

The SRM principle is based on a nested structural organization of the function set $S = Q(\mathbf{z}, \alpha),\ \alpha \in \Lambda$ such that

$$S_1 \subset S_2 \cdots \subset S_n \cdots, \tag{2.76}$$

where $S_k = \{Q(\mathbf{z}, \alpha) : \alpha \in \Lambda_k\}$ are subsets of the original function space such that $S^* = \cup_k S_k$ as it is illustrated in Fig. 2.7

Moreover, the set of utilized functions must form an *admissible structure* which satisfies the following three properties:

1. The VC dimension d_k of each set S_k of functions is finite,
2. Any element S_k of the structure contains totally bounded functions

$$0 \le Q(\mathbf{z}, \alpha) \le B_k,\ \alpha \in \Lambda_k,$$

3. The set S^* is everywhere dense in S in the $L_1(F)$ metric where $F = F(\mathbf{z})$ is the distribution function from which examples are drawn.

Note that in view of (2.76) the following assertions are true:

1. The sequence of values of VC dimensions d_k for the elements S_k of the structure S in nondecreasing with increasing k

$$d_1 \le d_2 \le \cdots \le d_n \le \cdots,$$

2. The sequence of values of the bounds B_k for the elements S_k of the structure S in nondecreasing with increasing k

$$B_1 \le B_2 \le \cdots \le B_n \le \cdots,$$

Denote by $Q(z, \alpha_l^k)$ the function that minimizes the empirical risk in the set of functions S_k. Then with probability $1 - \delta$ one can assert that the actual risk for this function is bounded by the following inequality

$$R(\alpha_l^k) \le R_{emp}(\alpha_l^k) + B_k \epsilon_k(l) \left(1 + \sqrt{1 + \frac{4R_{emp}\alpha_l^k}{B\epsilon_k(l)}}\right), \tag{2.77}$$

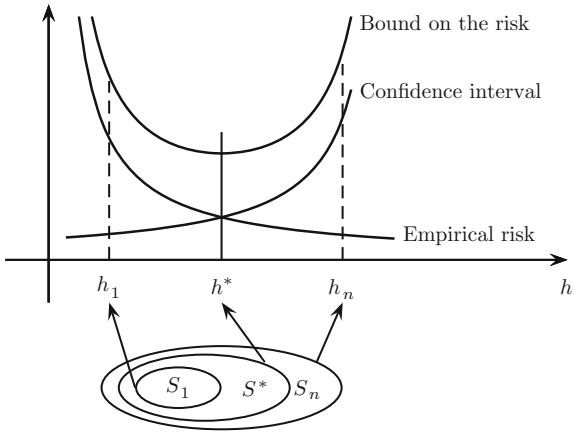

Fig. 2.8 Admissible structure of function sets

where

$$\epsilon_k(l) = 4\frac{d_k(\ln\frac{2l}{d_k}+1) - \ln\frac{\delta}{4}}{l}. \tag{2.78}$$

For a given set of observations $\mathbf{z}_1, \ldots, \mathbf{z}_l$, the SRM method actually suggests that one should choose the element S_k of the structure for which the smallest bound on the risk is achieved. In other words, the SRM principle introduces the notion of a *trade-off between the quality of approximation and the complexity of the approximating function* as it is particularly illustrated in Fig. 2.8.

Therefore, the SRM principle is based upon the following idea: *To provide the given set of functions with an admissible structure and then to find the function that minimizes risk (2.77) over given elements of the structure*. This principle is called the principle of structural risk minimization in order to stress the importance of choosing the element of the structure that possesses an appropriate capacity.

References

1. Bousquet, O., Boucheron, S., Lugosi, G.: Introduction to statistical learning theory. In: Advanced Lectures on Machine Learning, pp. 169–207. Springer, Heidelberg (2004)
2. Pednault, E.P.: Statistical Learning Theory. Citeseer (1997)
3. Sauer, N.: On the density of families of sets. J. Comb. Theory Ser. A **13**(1), 145–147 (1972)
4. Shelah, S.: A combinatorial problem; stability and order for models and theories in infinitary languages. Pac. J. Math. **41**(1), 247–261 (1972)
5. Vapnik, V.: The Nature of Statistical Learning Theory. Springer Science & Business Media, New York (2013)
6. Vapnik, V.N.: An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)
7. Vapnik, V.N., Chervonenkis, A.J.: Theory of Pattern Recognition (1974)

http://www.springer.com/978-3-319-47192-1

Machine Learning Paradigms
Artificial Immune Systems and their Applications in Software Personalization
Sotiropoulos, D.; Tsihrintzis, G.A.
2017, XVI, 327 p. 71 illus., 18 illus. in color., Hardcover
ISBN: 978-3-319-47192-1