Chapter 1

An Overview of Statistical Learning Theory

Vladimir Vapnik[1]

Abstract. Statistical learning theory was introduced in the late 1960's. Until the 1990's it was a purely theoretical analysis of the problem of function estimation from a given collection of data. In the middle of the 1990's new types of learning algorithms (called support vector machines) based on the developed theory were proposed. This made statistical learning theory not only a tool for the theoretical analysis but also a tool for creating practical algorithms for estimating multidimensional functions. This article presents a very general overview of statistical learning theory including both theoretical and algorithmic aspects of the theory. The goal of this overview is to demonstrate how the abstract learning theory established conditions for generalization which are more general than those discussed in classical statistical paradigms and how the understanding of these conditions inspired new algorithmic approaches to function estimation problems. A more detailed overview of the theory (without proofs) can be found in Vapnik (1995). In Vapnik (1998) one can find a detailed description of the theory (including proofs).

[1]This chapter is reprinted with permission from Vladimir Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks, Vol.10, No.5, pp.988-1000, 1999 (Copyright © 1999 IEEE). The author wants to thank Filip Mulier for discussions and help making the published article more clear and readable.

1.1 Setting of the Learning Problem

In this section we consider a model of learning and show that analysis of this model can be conducted in the general statistical framework of minimizing expected loss using observed data. We show that practical problems such as pattern recognition, regression estimation, and density estimation are particular case of this general model.

1.1.1 Function estimation model

The model of learning from examples can be described using three components:

1. A generator of random vectors x, drawn independently from a fixed but unknown distribution $P(x)$.

2. A supervisor that returns an output vector y for every input vector x, according to a conditional distribution function[2] $P(y|x)$, also fixed but unknown.

3. A learning machine capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$.

The problem of learning is that of choosing from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one which predicts the supervisor's response in the best possible way. The selection is based on a training set of ℓ random independently identically distributed (i.i.d.) observations

$$(x_1, y_1), ..., (x_\ell, y_\ell) \tag{1.1}$$

drawn according to $P(x, y) = P(x)P(y|x)$.

1.1.2 Problem of risk minimization

In order to choose the best available approximation to the supervisor's response, one measures the *loss* or discrepancy $L(y, f(x, \alpha))$ between the response y of the supervisor to a given input x and the response $f(x, \alpha)$ provided by the learning machine. Consider the expected value of the loss, given by the *risk functional*

$$R(\alpha) = \int L(y, f(x, \alpha)) dP(x, y). \tag{1.2}$$

The goal is to find the function $f(x, \alpha_0)$ which minimizes the risk functional $R(\alpha)$ over the class of functions $f(x, \alpha)$, $\alpha \in \Lambda$ in the situation where the joint probability distribution $P(x, y)$ is unknown and the only available information is contained in the training set (1.1).

1.1.3 Three main learning problems

This formulation of the learning problem is rather general. It encompasses many specific problems. Below we consider the main ones: the problems of pattern recognition, regression estimation, and density estimation.

[2] This is the general case which includes a case where the supervisor uses a function $y = f(x)$.

The Problem of Pattern Recognition. Let the supervisor's output y take only values $y \in \{0, 1\}$ and let $f(x, \alpha)$, $\alpha \in \Lambda$, be a set of *indicator* functions (functions which take on either value zero or one). Consider the following loss-function

$$L(y, f(x, \alpha)) = \begin{cases} 0 & \text{if } y = f(x, \alpha) \\ 1 & \text{if } y \neq f(x, \alpha). \end{cases} \tag{1.3}$$

For this loss function, the functional (1.2) provides the probability of classification error (i.e. when the answers y given by the supervisor and the answers given by the indicator function $f(x, \alpha)$ differ). Therefore, the problem is to find the function which minimizes the probability of classification errors when the probability measure $P(x, y)$ is unknown, but the data (1.1) are given.

The Problem of Regression Estimation. Let the supervisor's answer y be a real value, and let $f(x, \alpha)$, $\alpha \in \Lambda$ be a set of real functions which contains the *regression function*

$$f(x, \alpha_0) = \int y dP(y|x).$$

It is known that if $f(x, \alpha) \in L_2$ then the regression function is the one which minimizes the functional (1.2) with the the following loss-function

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2. \tag{1.4}$$

Thus the problem of regression estimation is the problem of minimizing the risk functional (1.2) with the loss function (1.4) in the situation where the probability measure $P(x, y)$ is unknown but the data (1.1) are given.

The Problem of Density Estimation. Finally, consider the problem of density estimation from the set of densities $p(x, \alpha)$, $\alpha \in \Lambda$. For this problem we consider the following loss-function

$$L(p(x, \alpha)) = -\log p(x, \alpha). \tag{1.5}$$

It is known that the desired density minimizes the risk functional (1.2) with the loss-function (1.5). Thus, again, in order to estimate the density from the data one has to minimize the risk-functional under the condition that the corresponding probability measure $P(x)$ is unknown but i.i.d. data

$$x_1, ..., x_n$$

are given.

The General Setting of the Learning Problem. The general setting of the learning problem can be described as follows. Let the probability measure $P(z)$ be defined on the space Z. Consider the set of functions $Q(z, \alpha), \alpha \in \Lambda$. The goal is: minimize the risk functional

$$R(\alpha) = \int Q(z, \alpha) dP(z), \quad \alpha \in \Lambda \tag{1.6}$$

for the probability measure $P(z)$ unknown but given an i.i.d. sample

$$z_1, ..., z_\ell. \tag{1.7}$$

The learning problems considered above are particular cases of this general problem of *minimizing the risk functional (1.6) on the basis of empirical data (1.7)*, where z denotes a pair (x, y) and $Q(z, \alpha)$ is the specific loss function (for example, either (1.3), (1.4) or (1.5)). In the sequel we will describe results obtained for the general statement of the problem. When applying this to specific problems, one has to substitute the corresponding loss-functions in the obtained formulas.

1.1.4 Empirical risk minimization induction principle

In order to minimize the risk functional (1.6) for an unknown probability measure $P(z)$ the following induction principle is usually employed.

The expected risk functional $R(\alpha)$ is replaced by the *empirical risk* functional

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \tag{1.8}$$

constructed on the basis of the training set (1.7). The principle is to approximate the function $Q(z, \alpha_0)$ which minimizes the risk (1.6) by the function $Q(z, \alpha_\ell)$ which minimizes the empirical risk (1.8). This principle is called the Empirical Risk Minimization induction principle (ERM principle).

1.1.5 Empirical risk minimization principle and the classical methods

The ERM principle is quite general. The classical methods for solving a specific learning problem, such as the least squares method in the problem of regression estimation or the maximum likelihood method in the problem of density estimation are realizations of the ERM principle for the specific loss functions considered above.

Indeed, in order to specify the regression problem one introduces an $n + 1$ dimensional variable $z = (x, y) = (x^1, ..., x^n, y)$ and uses loss function (1.4). Using this loss function in the functional (1.8) yields the functional

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2$$

which one needs to minimize for finding the regression estimate (least square method).

In order to estimate a density function from a given set of functions $p(x, \alpha)$ one uses the loss function (1.5). Putting this loss function into (1.8) one obtains the maximum likelihood method with functional

$$R_{emp}(\alpha) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \ln p(x_i, \alpha).$$

which one needs to minimize in order to find the approximation to the density.

Since the ERM principle is a general formulation of these classical estimation problems, any theory concerning the ERM principle applies to the classical methods as well.

1.1.6 Four parts of learning theory

Learning theory has to address the following four questions:

1. *What are the conditions for consistency of the ERM principle?*

   To answer this question one has to specify the *necessary and sufficient* conditions for convergence in probability[3]of the following sequences of the random values:

   - The values of risks $R(\alpha_\ell)$ converging to the minimal possible value of the risk $R(\alpha_0)$ (where $R(\alpha_\ell)$, $\ell = 1, 2, ...$ are the expected risks for functions $Q(z, \alpha_\ell)$ each minimizing the empirical risk $R_{emp}(\alpha_\ell)$)

   $$R(\alpha_\ell) \longrightarrow^{P}_{\ell\to\infty} R(\alpha_0), \tag{1.9}$$

   - The values of obtained empirical risks $R_{emp}(\alpha_\ell)$, $i = 1, 2, ...$ converging to the minimal possible value of the risk $R(\alpha_0)$

   $$R_{emp}(\alpha_\ell) \longrightarrow^{P}_{\ell\to\infty} R(\alpha_0). \tag{1.10}$$

   Equation (1.9) shows that solutions found using ERM converge to the best possible one. Equation (1.10) shows that empirical risk values converge to the value of the smallest risk.

2. *How fast does the sequence of smallest empirical risk values converge to the smallest actual risk?* In other words, what is the rate of generalization of a learning machine that implements the empirical risk minimization principle?

3. *How can one control the rate of convergence (the rate of generalization) of the learning machine?*

4. *How can one construct algorithms that can control the rate of generalization?*

The answers to these questions form the four parts of learning theory:

1. The theory of consistency of learning processes.

2. The non-asymptotic theory of the rate of convergence of learning processes.

3. The theory of controlling the generalization of learning processes.

4. The theory of constructing learning algorithms.

[3]Convergence in probability of values $R(\alpha_\ell)$ means that for any $\varepsilon > 0$ and for any $\eta > 0$ there exists a number $\ell_0 = \ell_0(\varepsilon, \eta)$ such that for any $\ell > \ell_0$ with probability at least $1 - \eta$ the inequality $R(\alpha_\ell) - R(\alpha_0) < \varepsilon$ holds true.

1.2 The Theory of Consistency of Learning Processes

The theory of consistency is an asymptotic theory. It describes *the necessary and sufficient conditions* for convergence of the solutions, obtained using the proposed method, to the best possible as the number of observations is increased. The following question arises:

> *Why do we need a theory of consistency if our goal is to construct algorithms for a small (finite) sample size?*

The answer is:

> *We need a theory of consistency because it provides not only sufficient but also necessary conditions for convergence of the empirical risk minimization inductive principle. Therefore, any theory of the empirical risk minimization principle must satisfy these necessary and sufficient conditions.*

In this section we introduce the main capacity concept (the so-called VC entropy) which defines the generalization ability of the ERM principle. In the next sections we show that the non-asymptotic theory of learning is based on different types of bounds that evaluate this concept for a fixed amount of observations.

1.2.1 The key theorem of the learning theory

The key theorem of the theory concerning the ERM based learning processes is the following [27]:

Theorem 1 (The Key Theorem) *Let $Q(z,\alpha)$, $\alpha \in \Lambda$ be a set of functions that has a bounded loss for probability measure $P(z)$*

$$A \leq \int Q(z,\alpha)P(z) \leq B \quad \forall \alpha \in \Lambda.$$

Then for the ERM principle to be consistent it is necessary and sufficient that the empirical risk $R_{emp}(\alpha)$ converges uniformly to the actual risk $R(\alpha)$ over the set $Q(z,\alpha)$, $\alpha \in \Lambda$ as

$$\lim_{\ell \to \infty} Prob\{\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \varepsilon\} = 0, \qquad \forall \varepsilon. \tag{1.11}$$

This type of convergence is called uniform one-sided convergence. In other words, according to the Key Theorem, the conditions for consistency of the ERM principle are equivalent to the conditions for existence of uniform one-sided convergence (1.11). This theorem is called the Key Theorem because it asserts that any analysis of the convergence properties of the ERM principle must be the *worst case analysis*. The necessary condition for consistency (not only the sufficient condition) depends on whether or not the deviation for the worst function over the given set of of functions

$$\Delta(\alpha_{worst}) = \sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha))$$

converges to zero in probability. From this theorem it follows that the analysis of the ERM principle requires an analysis of the properties of uniform convergence of the expectations to their probabilities over the given set of functions.

1.2.2 The necessary and sufficient conditions for uniform convergence

To describe the necessary and sufficient condition for uniform convergence (1.11), we introduce a concept called *the entropy of the set of functions* $Q(z,\alpha)$, $\alpha \in \Lambda$, *on the sample of size* ℓ. We introduce this concept in two steps: first for sets of indicator functions and then for sets of real valued functions.

Entropy of the Set of Indicator Functions. Let $Q(z,\alpha)$, $\alpha \in \Lambda$ be a set of indicator functions (i.e. functions which take only the values zero or one). Consider a sample

$$z_1,\ldots,z_\ell. \tag{1.12}$$

Let us characterize the diversity of this set of functions $Q(z,\alpha), \alpha \in \Lambda$ on the given sample by a quantity $N^\Lambda(z_1,\ldots,z_\ell)$ that represents the number of different separations of this sample that can be obtained using functions from the given set of indicator functions. Let us write this in another form. Consider the set of ℓ-dimensional binary vectors

$$q(\alpha) = (Q(z_1,\alpha),\ldots,Q(z_\ell,\alpha)), \quad \alpha \in \Lambda$$

that one obtains when α takes various values from Λ. Then geometrically speaking $N^\Lambda(z_1,\ldots,z_\ell)$ is the number of different vertices of the ℓ-dimensional cube that can be obtained on the basis of the sample $z_1,\ldots,z_\ell$ and the set of functions $Q(z,\alpha), \alpha \in \Lambda$.Let us call the value

$$H^\Lambda(z_1,\ldots,z_\ell) = \ln N^\Lambda(z_1,\ldots,z_\ell)$$

the *random entropy*. The random entropy describes the diversity of the set of functions on the given data. $H^\Lambda(z_1,\ldots,z_\ell)$ is a random variable since it was constructed using random i.i.d. data. Now we consider the expectation of the random entropy over the joint distribution function $F(z_1,\ldots,z_\ell)$:

$$H^\Lambda(\ell) = E\ln N^\Lambda(z_1,\ldots,z_\ell).$$

We call this quantity the entropy of the set of indicator functions $Q(z,\alpha)$, $\alpha \in \Lambda$ on samples of size ℓ. It depends on the set of functions $Q(z,\alpha)$, $\alpha \in \Lambda$, the probability measure $P(z)$, and the number of observations ℓ. The entropy describes the expected diversity of the given set of indicator functions on the sample of size ℓ.

The main result of the theory of consistency for the pattern recognition problem (the consistency for indicator loss function) is the following theorem [24]:

Theorem 2 *For uniform two-sided convergence of the frequencies to their probabili-*

ties[4] *it is necessary and sufficient that the equality*

$$\lim_{\ell\to\infty}\frac{H^{\Lambda}(\ell)}{\ell}=0,\quad \forall\varepsilon>0 \tag{1.14}$$

holds.

By slightly modifying the condition (1.14) one can obtain the necessary and sufficient condition for one-sided uniform convergence (1.11).

Entropy of the Set of Real Functions. Now we generalize the concept of entropy to sets of real valued functions. Let $A\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$, be a set of bounded loss functions. Using this set of functions and the training set (1.12) one can construct the following set of ℓ-dimensional real-valued vectors

$$q(\alpha)=(Q(z_1,\alpha),...,Q(z_\ell,\alpha)),\quad \alpha\in\Lambda. \tag{1.15}$$

This set of vectors belongs to the ℓ-dimensional cube with the edge $B-A$ and has a finite ε-net[5] in the metric C. Let $N=N^{\Lambda}(\varepsilon;z_1,...,z_\ell)$ be the number of elements of the minimal ε-net of the set of vectors $q(\alpha),\ \alpha\in\Lambda$. The logarithm of the (random) value $N^{\Lambda}(\varepsilon;z_1,...,z_\ell)$

$$H^{\Lambda}(\varepsilon;z_1,...,z_\ell)=\ln N^{\Lambda}(\varepsilon;z_1,...,z_\ell)$$

is called the *random VC-entropy*[6] of the set of functions $A\le Q(z,\alpha)\le B$ on the sample $z_1,...,z_\ell$. The expectation of the random VC-entropy

$$H^{\Lambda}(\varepsilon;\ell)=EH^{\Lambda}(\varepsilon;z_1,...,z_\ell)$$

is called the *VC-entropy* of the set of functions $A\le Q(z,\alpha)\le B,\ \alpha\in\Lambda$ on the sample of the size ℓ. Here expectation is taken with respect to product-measure $P(z_1,...,z_\ell)=P(z_1)\cdot...\cdot P(z_\ell)$.

The main results of the theory of uniform convergence of the empirical risk to the actual risk for bounded loss functions include the following theorem [24]:

Theorem 3 *For uniform two-sided convergence of the empirical risks to the actual risks*

$$\lim_{\ell\to\infty}Prob\{\sup_{\alpha\in\Lambda}(|R(\alpha)-R_{emp}(\alpha)|>\varepsilon\}=0,\qquad \forall\varepsilon. \tag{1.16}$$

[4]For sets of indicator functions $R(\alpha)$ defines probability and $R_{emp}(\alpha)$ defines frequency.

$$\lim_{\ell\to\infty}\mathrm{Prob}\{\sup_{\alpha\in\Lambda}|R(\alpha)-R_{emp}(\alpha)|>\varepsilon\}=0,\qquad \forall\varepsilon. \tag{1.13}$$

[5]The set of vectors $q(\alpha),\ \alpha\in\Lambda$ has minimal ε-net $q(\alpha_1),...,q(\alpha_N)$ if:

1. There exist $N=N^{\Lambda}(\varepsilon;z_1,...,z_\ell)$ vectors $q(\alpha_1),...,q(\alpha_N)$, such that for any vector $q(\alpha^*),\ \alpha^*\in\Lambda$ one can find among these N vectors one $q(\alpha_r)$ which is ε-close to this vector (in a given metric). For a C metric that means $\rho(q(\alpha^*),q(\alpha_r))=\max_{1\le i\le\ell}|Q(z_i\alpha^*)-Q(z_i,\alpha_r)|\le\varepsilon$.
2. N is minimal number of vectors which possess this property.

[6]Note that VC-entropy is different from classical metrical ε-entropy $H_{cl}^{\Lambda}(\varepsilon)=\ln N^{\Lambda}(\varepsilon)$ where $N^{\Lambda}(\varepsilon)$ is cardinality of the minimal ε-net of the set of functions $Q(z,\alpha),\alpha\in\Lambda$.

it is necessary and sufficient that the equality

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}(\varepsilon, \ell)}{\ell} = 0, \quad \forall \varepsilon > 0 \tag{1.17}$$

be valid.

By slightly modifying the condition (1.16) one can obtain the necessary and sufficient condition for one-sided uniform convergence (1.11). According to the key assertion this implies the necessary and sufficient conditions for consistency of the ERM principle.

1.2.3 Three milestones in learning theory

In this section, we consider for simplicity reasons a set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$ (i.e. we consider the problem of pattern recognition). The results obtained for sets of indicator functions can be generalized to sets of real-valued functions. In the previous section we introduced the entropy for sets of indicator functions

$$H^{\Lambda}(\ell) = E \ln N^{\Lambda}(z_1, \ldots, z_{\ell}).$$

Now, we consider two new functions that are constructed on the basis of the values $N^{\Lambda}(z_1, ..., z_{\ell})$: the *Annealed VC-entropy*

$$H^{\Lambda}_{ann}(\ell) = \ln E N^{\Lambda}(z_1, ..., z_{\ell})$$

and the *Growth function*

$$G^{\Lambda}(\ell) = \ln \sup_{z_1, ..., z_{\ell}} N^{\Lambda}(z_1, ..., z_{\ell}).$$

These functions are determined in such a way that for any ℓ the inequalities

$$H^{\Lambda}(\ell) \leq H^{\Lambda}_{ann}(\ell) \leq G^{\Lambda}(\ell)$$

are valid. On the basis of these functions the three main milestones in Statistical Learning Theory are constructed. In the previous section we introduced the equation

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}(\ell)}{\ell} = 0$$

describing the *necessary and sufficient condition* for consistency of the ERM principle. This equation is the first milestone in learning theory: any machine that is minimizing empirical risk should satisfy it.

On the other hand, this equation says nothing about the rate of convergence of obtained risks $R(\alpha_{\ell})$ to the minimal one $R(\alpha_0)$. It is possible that the ERM principle is consistent but has an arbitrary slow asymptotic rate of convergence. The question is then: *Under which conditions does one have a fast asymptotic rate of convergence?*

We say that the asymptotic rate of convergence is fast if for any $\ell > \ell_0$ the exponential bound

$$P\{R(\alpha_\ell) - R(\alpha_0) > \varepsilon\} < e^{-c\varepsilon^2\ell}$$

holds true, where $c > 0$ is some constant. The equation

$$\lim_{\ell\to\infty} \frac{H^\Lambda_{ann}(\ell)}{\ell} = 0$$

describes the *sufficient* condition for fast convergence[7]. This constitutes the second milestone in Statistical Learning Theory: guaranteeing a fast asymptotic rate of convergence. Note that both the equation describing the necessary and sufficient condition for consistency and the one that describes the sufficient condition for fast convergence of the ERM method are valid for a *given* probability measure $P(z)$ (both VC-entropy $H^\Lambda(\ell)$ and VC-annealed entropy $H^\Lambda_{ann}(\ell)$ are constructed using this measure).

However, our goal is to construct a learning machine for solving many different problems (i.e. for many different probability measures). The next question is then: *Under what conditions is the ERM principle consistent and rapidly converging independently of the probability measure?* The following equation describes the necessary and sufficient conditions for consistency of ERM for any probability measure:

$$\lim_{\ell\to\infty} \frac{G^\Lambda(\ell)}{\ell} = 0.$$

This condition is also sufficient for fast convergence. This equation is the third milestone in Statistical Learning Theory. It describes the conditions under which the learning machine implementing the ERM principle has an asymptotic high rate of convergence, independent of the problem to be solved.

These milestones form a foundation for constructing both distribution independent bounds and rigorous distribution dependent bounds for the rate of convergence of learning machines.

1.3 Bounds on the Rate of Convergence of the Learning Processes

In order to estimate the quality of the ERM method for a given sample size it is necessary to obtain non-asymptotic bounds on the rate of uniform convergence.

A non-asymptotic bound of the rate of convergence can be obtained using a new capacity concept, called the VC dimension (abbreviation for Vapnik-Chervonenkis dimension), which allows us to obtain a constructive bound for the growth function. The concept of VC-dimension is based on a remarkable property of the Growth-function $G^\Lambda(\ell)$.

[7] The necessity of this condition for fast convergence is an open question.

1.3.1 The structure of the growth function

Theorem 4 [23-24] *Any growth function either satisfies the equality*

$$G^{\Lambda}(\ell) = \ell \ln 2$$

or is bounded by the inequality

$$G^{\Lambda}(\ell) < h(\ln\frac{\ell}{h} + 1),$$

where h is an integer for which

$$G^{\Lambda}(h) = h\ln 2$$

$$G^{\Lambda}(h+1) \neq (h+1)\ln 2.$$

In other words the Growth function will be either a linear function or will be bounded by a logarithmic function. (e.g. it cannot be of the form $G^{\Lambda}(\ell) = c\sqrt{\ell}$). We say that the VC dimension of the set of indicator functions $Q(z,\alpha), \alpha \in \Lambda$ is infinite if the Growth function for this set of functions is linear. We say that the VC dimension of the set of indicator functions $Q(z,\alpha), \alpha \in \Lambda$ is finite and equals h if the Growth function is bounded by a logarithmic function with coefficient h.

The finiteness of the VC-dimension of the set of indicator functions implemented by the learning machine forms the necessary and sufficient condition for consistency of the ERM method independent of the probability measure. Finiteness of the VC-dimension also implies fast convergence.

1.3.2 Equivalent definition of the VC dimension

In this section we give an equivalent definition of the VC dimension of sets of indicator functions and we generalize this definition to sets of real functions.

The VC dimension of a set of indicator functions. The *VC-dimension of a set of indicator functions* $Q(z,\alpha)$, $\alpha \in \Lambda$, is the maximum number h of vectors $z_1, ..., z_h$ which can be separated in all 2^h possible ways using functions of this set[8] (*shattered* by this set of functions). If for any n there exists a set of n vectors which can be shattered by the set $Q(z,\alpha)$, $\alpha \in \Lambda$, then the VC-dimension is equal to infinity.

The VC dimension of a set of real valued functions. Let $a \leq Q(z,\alpha) \leq A$, $\alpha \in \Lambda$, be a set of real valued functions bounded by constants a and A (a can approach $-\infty$ and A can be equal to ∞). Let us consider along with the set of real valued functions $Q(z,\alpha)$, $\alpha \in \Lambda$, the set of indicator functions

$$I(z,\alpha,\beta) = \theta\{Q(z,\alpha) - \beta\}, \ \alpha \in \Lambda \tag{1.18}$$

[8] Any indicator function separates a set of vectors into two subsets: the subset of vectors for which this function takes value 0 and the subset of vectors for which it takes value 1.

where $a < \beta < A$ is some constant, $\theta(u)$ is a step function:

$$\theta(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geq 0. \end{cases}$$

The *VC dimension of the set of real valued functions* $Q(z, \alpha)$, $\alpha \in \Lambda$, is defined to be the VC-dimension of the set of indicator functions (1.18).

1.3.3 Two important examples

Example 1

1. The VC-dimension of the set of *linear indicator functions*

$$Q(z, \alpha) = \theta\left\{\sum_{p=1}^{n} \alpha_p z_p + \alpha_0\right\}$$

in n-dimensional coordinate space $Z = (z_1, ..., z_n)$ is equal to $h = n + 1$, since using functions of this set one can shatter at most $n+1$ vectors. Here $\theta\{\cdot\}$ is the step function, which takes value 1 if the expression between brackets is positive and takes value 0 otherwise.

2. The VC-dimension of the set of *linear functions*

$$Q(z, \alpha) = \sum_{p=1}^{n} \alpha_p z_p + \alpha_0, \quad \alpha_0, ..., \alpha_n \in (-\infty, \infty)$$

in n-dimensional coordinate space $Z = (z_1, ..., z_n)$ is also equal to $h = n + 1$ because the VC-dimension of the corresponding linear indicator functions is equal to $n+1$ (using $\alpha_0 - \beta$ instead of α_0 does not change the set of indicator functions).

Example 2

We call a hyperplane

$$(w^* \cdot x) - b = 0, \quad |w^*| = 1$$

the Δ-margin separating hyperplane if it classifies vectors x as follows

$$y = \begin{cases} 1 & \text{if } (w^* \cdot x) - b \geq \Delta \\ -1 & \text{if } (w^* \cdot x) - b \leq -\Delta. \end{cases}$$

Classifications of vectors x that fall within the margin $(-\Delta, \Delta)$ are undefined.

Theorem 5 [25, 20–22] *Let vectors $x \in X$ belong to a sphere of radius R, then the set of Δ-margin separating hyperplanes has VC dimension h bounded by the inequality*

$$h \leq \min\left(\left[\frac{R^2}{\Delta^2}\right], n\right) + 1.$$

These examples show that in general the VC dimension of the set of hyperplanes equals $n + 1$, where n is the dimensionality of the input space. However, the VC dimension of the set of Δ-margin separating hyperplanes (with a large value of margin Δ) can be less than $n + 1$. This fact will play an important role towards constructing new function estimation methods.

1.3.4 Distribution independent bounds for the rate of convergence of learning processes

Consider sets of functions which possess a finite VC-dimension h. We distinguish then between the following two cases:

1. The case where the set of loss functions $Q(z,\alpha),\ \alpha \in \Lambda$ is a set of *totally bounded functions*

2. The case where the set of loss functions $Q(z,\alpha),\ \alpha \in \Lambda$ is *not necessarily a set of totally bounded functions.*

Case 1 [The set of totally bounded functions] Without loss of generality, we assume that

$$0 \leq Q(z,\alpha) \leq B,\ \alpha \in \Lambda. \tag{1.19}$$

The main result in the theory of bounds for sets of totally bounded functions is the following [20–22]:

Theorem 6 *With probability at least* $1-\eta$*, the inequality*

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\varepsilon}{2}\left(1+\sqrt{1+\frac{4R_{emp}(\alpha)}{B\varepsilon}}\right), \tag{1.20}$$

holds true simultaneously for all functions of the set (1.19), where

$$\varepsilon = 4\frac{h(\ln\frac{2\ell}{h}+1)-\ln\eta}{\ell}. \tag{1.21}$$

For the set of indicator functions: $B=1$

This Theorem provides bounds for the risks of all functions of the set (1.18) (including the function $Q(z,\alpha_\ell)$ which minimizes the empirical risk (1.8)). The bounds follow from the bound on uniform convergence (1.13) for sets of totally bounded functions that have finite VC dimension.

Case 2 [The set of unbounded functions] Consider the set of (nonnegative) unbounded functions $0 \leq Q(z,\alpha),\ \alpha \in \Lambda$. It is easy to show (by constructing an example) that, without additional information about the set of unbounded functions and/or probability measures, it is impossible to obtain an inequality of type (1.20). Below, we use the following information

$$\sup_{\alpha\in\Lambda}\frac{\left(\int Q^p(z,\alpha)dP(z)\right)^{\frac{1}{p}}}{\int Q(z,\alpha)dP(z)} \leq \tau < \infty \tag{1.22}$$

where $p > 1$ is some fixed constant[9].

The main result for the case of unbounded sets of loss functions is the following [20-22]:

Theorem 7 *With probability at least* $1 - \eta$ *the inequality*

$$R(\alpha) \leq \frac{R_{emp}(\alpha)}{(1 - a(p)\tau\sqrt{\varepsilon})_+}, \quad a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}} \tag{1.23}$$

holds true simultaneously for all functions of the set, where ε *is determined by (1.21),* $(a)_+ = max(a, 0)$.

The Theorem bounds the risks for all functions of the set (1.15) (including the function $Q(z, \alpha_\ell)$).

1.3.5 Problem of constructing rigorous (distribution dependent) bounds

To construct rigorous bounds for the rate of convergence one has to take into account information about the probability measure. Let $\mathcal{P}_0$ be a set of all probability measures and let $\mathcal{P} \subset \mathcal{P}_0$ be a subset of the set $\mathcal{P}_0$. We say that one has prior information about an unknown probability measure $P(z)$ if one knows the set of measures $\mathcal{P}$ that contains $P(z)$. Consider the following generalization of the Growth function

$$\mathcal{G}^{\Lambda}_{\mathcal{P}}(\varepsilon, \ell) = \lg \sup_{P \in \mathcal{P}} E_P N^{\Lambda}(\varepsilon; z_1, ..., z_\ell).$$

For indicator functions $Q(z, \alpha), \alpha \in \Lambda$ and for the extreme case where $\mathcal{P} = \mathcal{P}_0$ the Generalized Growth function $\mathcal{G}^{\Lambda}_{\mathcal{P}}(\varepsilon, \ell)$ coincides with the Growth function $G^{\Lambda}(\ell)$. For another extreme case where $\mathcal{P}$ contains only one function $P(z)$ the Generalized growth function coincides with the annealed VC-entropy.

The following assertion is true [20, 26]:

Theorem 8 *Suppose that a set of loss-functions is bounded*

$$-\inf < A \leq Q(z, \alpha) \leq B < \infty, \qquad \alpha \in \Lambda.$$

Then for sufficiently large ℓ *the following inequality*

$$P\left\{\sup_{\alpha \in \Lambda}\left|\int Q(z, \alpha) dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i, \alpha)\right| > \varepsilon\right\} \leq$$

[9] This inequality describes some general properties of distribution functions of the random variables $\xi_\alpha = Q(z, \alpha)$, generated by the $P(z)$. It describes the "tails of distributions" (the probability of large values for the random variables ξ_α). If the inequality (1.22) with $p > 2$ holds, then the distributions have so-called "light tails" (large values do not occurs very often). In this case rapid convergence is possible. If, however, the inequality (1.22) holds only for $p < 2$ (large values of the random variables ξ_α occur rather often) then the rate of convergence will be small (it will be arbitrarily small if p is sufficiently close to one).

$$12\exp\left\{\left(\frac{G_p^{\Lambda}\Lambda_{ann}(\varepsilon/6(B-A),2\ell)}{\ell}-\frac{\varepsilon^2}{B-A}+\frac{\ln\ell}{\ell}\right)\ell\right\}.$$

holds true.

From this bound it follows that for sufficiently large ℓ with probability $1-\eta$ simultaneously for all $\alpha \in \Lambda$ (including the one that minimizes the empirical risk) the following inequality is valid:

$$\int Q(z,\alpha)dF(z) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha) + \sqrt{\frac{\mathcal{G}_p^{\Lambda}(\varepsilon/6(B_A),2\ell)-\ln\eta/12}{\ell}}.$$

However, this bound is non-constructive because the theory does not specify a method for evaluating the Generalized Growth function. In order to make this bound constructive and rigorous one has to estimate the Generalized Growth function for a given set of loss-functions and a given set of probability measures. This is one of the main subjects of the current learning theory research.

1.4 Theory for Controlling the Generalization of Learning Machines

The theory for controlling the generalization of a learning machine is devoted to constructing an induction principle for minimizing the risk functional which takes into account the *size of the training set* (an induction principle for a *"small" sample size*[10]). The goal is to specify methods which are appropriate for a given sample size.

1.4.1 Structural risk minimization induction principle

The ERM principle is intended for dealing with a large sample size. Indeed, the ERM principle can be justified by considering the inequalities (1.20). When ℓ/h is large, the second summand on the right hand side of inequality (1.20) becomes small. The actual risk is then close to the value of the empirical risk. In this case, a small value of the empirical risk provides a small value of (expected) risk. However, if ℓ/h is small, then even a small $R_{emp}(\alpha_\ell)$ does not guarantee a small value of risk. In this case the minimization for $R(\alpha)$ requires a new principle, based on the simultaneous minimization of the two terms in inequality (1.20), one of which depends on the value of the empirical risk while the second depends on the VC-dimension of the set of functions. To minimize the risk in this case it is necessary to find a method which, along with minimizing the value of empirical risk, controls the VC-dimension of the learning machine.

The following principle, which is called the principle of Structural Risk Minimization (SRM), is intended for minimizing the risk functional with respect to both empirical risk and VC-dimension of the set of functions. Let, the set S of functions $Q(z,\alpha)$, $\alpha \in \Lambda$,

[10]The sample size ℓ is considered to be small if ℓ/h is small, say $\ell/h < 20$.

be provided with a *structure*: so that S is composed of the nested subsets of functions $S_k = \{Q(z, \alpha),\ \alpha \in \Lambda_k\}$, such that

$$S_1 \subset S_2 \subset ... \subset S_n ... \tag{1.24}$$

and $S^* = \bigcup_k S_k$. An *admissible structure* is one that satisfies the following three properties:

1. The set S^* is everywhere dense in S.
2. The VC-dimension h_k of each set S_k of functions is finite.
3. Any element S_k of the structure contains totally bounded functions $0 \leq Q(z, \alpha) \leq B_k,\ \alpha \in \Lambda_k$.

The SRM principle suggests to do the following: for a given set of observations $z_1, ..., z_\ell$ choose the element of structure S_n with $n = n(\ell)$ and the particular function from S_n such that the guaranteed risk (1.20) is minimal. The SRM principle actually suggests a *trade-off between the quality of the approximation and the complexity of the approximating function.* As n increases, empirical risk minima decrease, but on the other hand the term responsible for the confidence interval (summand in (1.20)) increases. The SRM principle takes both factors into account.

The main results of the theory of SRM are the following [9, 22]:

Theorem 9 *For any distribution function the SRM method provides convergence to the best possible solution with probability one.*

In other words SRM method is universally strongly consistent.

Theorem 10 [22] *For admissible structures the method of structural risk minimization provides approximations $Q(z, \alpha_\ell^{n(\ell)})$ for which the sequence of risks $R(\alpha_\ell^{n(\ell)})$ converge to the best one $R(\alpha_0)$ with asymptotic rate of convergence*[11]

$$V(\ell) = r_{n(\ell)} + B_{n(\ell)} \sqrt{\frac{h_{n(\ell)} \ln \ell}{\ell}} \tag{1.25}$$

if the law $n = n(\ell)$ is such that

$$\lim_{\ell \to \infty} \frac{B_{n(\ell)}^2 h_{n(\ell)} \ln \ell}{\ell} = 0. \tag{1.26}$$

In equation (1.25) B_n is the bound for functions from S_n and $r_n(\ell)$ is the rate of approximation

$$r_n = \inf_{\alpha \in \Lambda_n} \int Q(z, \alpha) dP(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dP(z).$$

[11] We say that the random variables ξ_ℓ, $\ell = 1, 2, ...$ converge to the value ξ_0 with asymptotic rate $V(\ell)$ if there exists constant C such that $V^{-1}(\ell) |\xi_\ell - \xi_0| \longrightarrow_{\ell \to \infty}^{P} C$.

1.5 Theory of Constructing Learning Algorithms

In order to implement the SRM induction principle in learning algorithms one has to control two factors in the to be minimized bound (1.20):

1. The value of empirical risk
2. The capacity factor (to choose the element S_n with the appropriate value of VC dimension).

We confine ourselves now to the pattern recognition case and consider two type of learning machines:

1. Neural Networks (NN) that were inspired on the biological analogy with the brain
2. The support vector machines that were inspired on statistical learning theory.

We discuss how each corresponding machine can control these factors.

1.5.1 Methods of separating hyperplanes and their generalization

Consider first the problem of minimizing empirical risk on the set of *linear indicator functions*

$$f(x, w) = \theta\left\{\sum_{i=0}^{n} w_i x^i\right\}, \ w \in W. \tag{1.27}$$

Let

$$(x_1, y_1), ..., (x_\ell, y_\ell)$$

be a training set where $x_j = (x_j^1, ..., x_j^n)$ is a vector, $y_j \in \{0, 1\}$, $j = 1, ..., \ell$.

For minimizing the empirical risk one has to find the parameters $w = (w_1, ..., w_n)$ (weights) which minimize the empirical risk functional

$$R_{emp}(w) = \frac{1}{\ell}\sum_{j=1}^{\ell}(y_j - f(x_j, w))^2. \tag{1.28}$$

There are several methods for minimizing this functional. In the case when the minimum of the empirical risk is zero one can find the exact solution while when the minimum of this functional is nonzero one can find an approximate solution. Therefore, by constructing a separating hyperplane one can control the value of empirical risk. Unfortunately, the set of separating hyperplanes is not flexible enough to provide low empirical risk for many real life problems [13].

Two opportunities were considered to increase the flexibility of the sets of functions:

1. to use a richer set of indicator functions which are superpositions of linear indicator functions
2. to map the input vectors into a high dimensional feature space and construct in this space a Δ-margin separating hyperplane.

The first idea corresponds to the neural network. The second idea leads to support vector machines.

1.5.2 Sigmoid approximation of indicator functions and neural nets

For describing the idea behind the NN let us consider the method of minimizing the functional (1.28). It is impossible to use regular *gradient-based* methods of optimization for minimizing this functional. The gradient of the indicator function $R_{emp}(w)$ is either equal to zero or is undefined. The solution is to approximate the set of indicator functions (1.27) by so-called *sigmoid functions*

$$\overline{f}(x,w) = S\left\{\sum_{i=0}^{n} w_i x^i\right\} \tag{1.29}$$

where $S(u)$ is a smooth monotonic function such that $S(-\infty) = 0$, $S(+\infty) = 1$. For example, the functions

$$S_1(u) = \frac{1}{1+\exp^{-u}}, \qquad S_2(u) = \frac{2\text{arctan(u)} + \pi}{2\pi}.$$

are sigmoid functions.

For the set of sigmoid function, the empirical risk functional

$$R_{emp}(w) = \frac{1}{\ell}\sum_{j=1}^{\ell}(y_j - \overline{f}(x_j,w))^2 \tag{1.30}$$

is smooth in w. It has a gradient $\text{grad}R_{emp}(w)$ and therefore can be minimized using gradient-based methods. For example, the *gradient descent method* uses the following update rule

$$w_{new} = w_{old} - \gamma(\cdot)\text{grad } R_{emp}(w_{old})$$

where the data $\gamma(\cdot) = \gamma(n) \geq 0$ depend on the iteration number n. For convergence of the gradient descent method to a local minimum, it is enough that $\gamma(n)$ satisfy the conditions

$$\sum_{n=1}^{\infty}\gamma(n) = \infty, \qquad \sum_{n=1}^{\infty}\gamma^2(n) < \infty.$$

Thus, the idea is to use the sigmoid approximation at the stage of estimating the coefficients, and use the indicator functions with these coefficients at the stage of recognition.

The generalization of this idea leads to feedforward neural nets. In order to increase the flexibility of the set of decision rules of the learning machine one considers a set of functions which are the superposition of several linear indicator functions (networks of neurons) [13] instead of the set of linear indicator functions (single neuron). All indicator functions in this superposition are replaced by sigmoid functions.

A method for calculating the gradient of the empirical risk for the sigmoid approximation of neural nets, called the *back-propagation method*, was found [15],[12]. Using this gradient descent method, one can determine the corresponding coefficient values (weights) of all elements of the neural net. In the 1990s it was proven that the VC dimension of neural networks depends on the type of sigmoid functions and the number of weights in the neural net. Under some general conditions the VC dimension of the

neural net is bounded (although it is sufficiently large). Suppose that the VC dimension does not change during the neural network training procedure, then the generalization ability of neural net depends on how well the neural net minimizes the empirical risk using a sufficiently large number of training data.

The three main problems encountered when minimizing the empirical risk using the back-propagation method are:

1. The empirical risk functional has many local minima. Optimization procedures guarantee convergence to some local minimum. In general the function which is found using the gradient-based procedure can be far from the best one. The quality of the obtained approximation depends on many factors, in particular on the initial parameter values of the algorithm.

2. Convergence to a local minimum can be rather slow (due to the high dimensionality of the weight-space).

3. The sigmoid function has a scaling factor which affects the quality of the approximation. To choose the scaling factor one has to make a trade-off between quality of approximation and the rate of convergence.

Therefore, a good minimization of the empirical risk depends in many respects on the art of the researcher in this case.

1.5.3 The optimal separating hyperplanes

For introducing the method that serves as an alternative to the neural network, let us consider optimal separating hyperplanes [25]. Suppose the training data

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), \quad x \in R^n, \quad y \in \{+1, \ -1\}$$

can be separated by a hyperplane:

$$(w \cdot x) - b = 0. \tag{1.31}$$

We say that this set of vectors is separated by the *Optimal hyperplane (or the Maximal Margin hyperplane)* if it is separated without error and the distance between the closest vector and the hyperplane is maximal. To describe the separating hyperplane let us use the following form:

$$(w \cdot x_i) - b \geq 1 \qquad \text{if } y_i = 1\ ,$$
$$(w \cdot x_i) - b \leq -1 \qquad \text{if } y_i = -1.$$

In the following we use a compact notation for these inequalities:

$$y_i[(w \cdot x_i) - b] \geq 1, \quad i = 1, \ldots, \ell. \tag{1.32}$$

It is easy to check that the Optimal hyperplane is the one that satisfies the conditions (1.32) and minimizes functional

$$\Phi(w) = \frac{1}{2}||w||^2 = \frac{1}{2}(w, w). \tag{1.33}$$

The minimization is taken with respect to both vector w and scalar b.

The solution to this optimization problem is given by the saddle point of the Lagrange functional (Lagrangian):

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{\ell} \alpha_i \{[(x_i \cdot w) - b]y_i - 1\}, \tag{1.34}$$

where the α_i are the Lagrange multipliers. The Lagrangian has to be minimized with respect to w, b and maximized with respect to $\alpha_i \geq 0$.

In the saddle point, the solutions w_0, b_0, and α^0 should satisfy the conditions

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial b} = 0, \quad \frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = 0.$$

Rewriting these equations in explicit form one obtains the following properties of the Optimal hyperplane:

(i) The coefficients α_i^0 for the Optimal hyperplane should satisfy the constraints

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i = 0, \qquad \alpha_i^0 \geq 0, \qquad i = 1, \ldots, \ell \tag{1.35}$$

(ii) The parameters of the Optimal hyperplane (vector w_0) are a linear combination of the vectors of the training set with

$$w_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i, \qquad \alpha_i^0 \geq 0, \qquad i = 1, \ldots, \ell \tag{1.36}$$

(iii) The solution must satisfy the following Kuhn–Tucker conditions,

$$\alpha_i^0 \{[(x_i \cdot w_0) - b_0]y_i - 1\} = 0, \quad i = 1, \ldots, \ell. \tag{1.37}$$

From these conditions it follows that only some training vectors in expansion (1.36) (called the *support vectors*) can have nonzero coefficients α_i^0 in the expansion of w_0. The support vectors are the vectors for which, in inequality (1.36), the equality is achieved. Therefore we obtain

$$w_0 = \sum_{\text{support vectors}} y_i \alpha_i^0 x_i, \qquad \alpha_i^0 \geq 0. \tag{1.38}$$

Substituting the expression for w_0 back into the Lagrangian and taking into account the Kuhn–Tucker conditions, one obtains the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j}^{\ell} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j). \tag{1.39}$$

It remains to maximize this functional in the non-negative quadrant

$$\alpha_i \geq 0, \quad i = 1, ..., \ell$$

under the constraint

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0. \tag{1.40}$$

Putting the expression for w_0 in (1.31), we obtain the hyperplane as an expansion on support vectors

$$\sum_{i=1}^{\ell} \alpha_i^0 (x, x_i) + b_0 = 0. \tag{1.41}$$

To construct the Optimal hyperplane in the case when the data are linearly non-separable, we introduce non-negative variables $\xi_i \geq 0$ and the functional

$$\Phi(\xi) = (w, w) + C \sum_{i=1}^{\ell} \xi_i$$

which we minimize subject to the constraints

$$y_i((w \cdot x_i) - b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, \ell.$$

Using the same formalism with Lagrange multipliers one can show that the optimal hyperplane also has an expansion (1.41) on support vectors. The coefficients α_i can be found by maximizing the same quadratic form as in the separable case (1.39) under slightly different constraints

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, \ell, \text{ and } \sum_{i=1}^{\ell} \alpha_i y_i = 0. \tag{1.42}$$

1.5.4 The support vector network

The support vector network implements the following idea [21]: Map the input vectors into a very high dimensional feature space Z through some non-linear mapping chosen *a priori*. Then construct an optimal separating hyperplane in this space. The goal is to create the situation as described previously in example 2, where for Δ-margin separating hyperplanes the VC dimension is defined by the ratio R^2/Δ^2. In order to generalize well, we control (decrease) the VC dimension by constructing an optimal separating hyperplane (that maximizes the margin). To increase the margin we use very high dimensional spaces.

Example. Consider a mapping that allows us to construct decision polynomials in the input space. To construct a polynomial of degree two, one can create a feature space, Z, which has $N = \frac{n(n+3)}{2}$ coordinates of the form:

$$z_1 = x_1, \ldots, z_n = x_n \ , \qquad n \text{ coordinates} \ ,$$

$$z_{n+1}=x_1^2,\ldots,z_{2n}=x_n^2\ , \qquad n \text{ coordinates}\ ,$$

$$z_{2n+1}=x_1x_2,\ldots,z_N=x_nx_{n-1}\ , \qquad \tfrac{n(n-1)}{2} \text{ coordinates}\ ,$$

where $x=(x_1,\ldots,x_n)$. The separating hyperplane constructed in this space is a separating second degree polynomial in the input space. To construct a polynomial of degree k in an n dimensional input space one has to construct a $O(n^k)$ dimensional feature space, where one then constructs the optimal hyperplane. The problem then arises of how to computationally deal with such high-dimensional spaces: for constructing a polynomial of degree 4 or 5 in a 200 dimensional space it is necessary to construct hyperplanes in a billion dimensional feature space.

In 1992 it was noted [5] that both for describing the optimal separating hyperplane in the feature space (1.41) and estimating the corresponding coefficients of expansion of the separating hyperplane (1.39) one uses the inner product between two vectors $z(x_1)$ and $z(x_2)$, which are images in the feature space of the input vectors x_1 and x_2. Therefore, if one can estimate the inner product of the two vectors in the feature space $z(x_1)$ and $z(x_2)$ as a function of two variables in the input space

$$(z_i \cdot z) = K(x, x_i),$$

then it will be possible to construct the solutions which are equivalent to the optimal hyperplane in the feature space. To get this solution one only needs to replace the inner product (x_i, x_j) in equations (1.39) and (1.41) by the function $K(x_i, x_j)$. In other words, one constructs nonlinear decision functions in the input space

$$I(x) = \text{sign}\left(\sum_{\text{support vectors}} \alpha_i K(x_i, x) + b_0\right), \tag{1.43}$$

that are equivalent to the linear decision functions (1.33) in the feature space. The coefficients α_i in (1.43) are defined by solving the equation

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2}\sum_{i,j}^{\ell} \alpha_i\alpha_j y_i y_j K(x_i, x_j) \tag{1.44}$$

under constraints (1.42). In 1909 Mercer proved a theorem which defines the general form of inner products in Hilbert spaces.

Theorem 11 *The general form of the inner product in Hilbert space is defined by the symmetric positive definite function $K(x,y)$ that satisfies the condition*

$$\int K(x,y)z(x)z(y)dxdy \geq 0$$

for all functions $z(x)$, $z(y)$ satisfying the inequality

$$\int z^2(x)dx \leq \infty.$$

Therefore any function $K(x, y)$ satisfying Mercer's condition can be used for constructing rule (1.42) which is equivalent to constructing an optimal separating hyperplane in some feature space. The learning machines which construct decision functions of the type (1.43) are called *Support Vectors Networks or Support Vector Machines*[12].

Using different expressions for inner products $K(x, x_i)$ one can construct different learning machines with arbitrary types of (nonlinear in the input space) decision surfaces. For example to specify polynomials of any fixed order d one can use the following functions for the inner product in the corresponding feature space

$$K(x, x_i) = ((x \cdot x_i) + 1)^d.$$

Radial Basis Function machines with decision functions of the form

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i \exp\left\{-\frac{|x - x_i|^2}{\sigma^2}\right\}\right)$$

can be implemented by using a function of the type

$$K(x, x_i) = \exp\left\{-\frac{|x - x_i|^2}{\sigma^2}\right\}.$$

In this case the SVM machine will find both the centers x_i and the corresponding weights α_i.

The SVM possesses some useful properties:

- The optimization problem for constructing an SVM has a unique solution.
- The learning process for constructing an SVM is rather fast.
- Simultaneously with constructing the decision rule, one obtains the set of support vectors.
- Implementation of a new set of decision functions can be done by changing only one function (kernel $K(x_i, x)$), which defines the dot product in Z-space.

1.5.5 Why can neural networks and support vectors networks generalize?

The generalization ability of both Neural Networks and Support Vectors Networks is based on the factors described in the theory for controlling the generalization of the learning processes. According to this theory, to guarantee a high rate of generalization of the learning machine one has to construct a structure

$$S_1 \subset S_2 \subset ... \subset S$$

[12]This name stresses that for constructing this type of machine, the idea of expanding the solution on support vectors is crucial. In the SVM the complexity of construction depends on the number of support vectors rather than on the dimensionality of the feature space.

on the set of decision functions $S = \{Q(z, \alpha), \alpha \in \Lambda\}$ and then choose both an appropriate element S_k of the structure and a function $Q(z, \alpha_\ell^k) \in S_k$ within this element that minimizes bound (1.20). The bound (1.16) can be rewritten in the simple form

$$R(\alpha_\ell^k) \leq R_{emp}(\alpha_\ell^k) + \Omega(\frac{\ell}{h_k}) \tag{1.45}$$

where the first term is an estimate of the risk and the second term is the confidence interval for this estimate.

In designing a neural network, one determines a set of admissible functions with some VC-dimension h^*. For a given amount ℓ of training data the value h^* determines the confidence interval $\Omega(\frac{\ell}{h^*})$ for the network. Choosing the appropriate element of a structure is therefore a problem of designing the network for a given training set. During the learning process this network minimizes the first term in the bound (1.45) (the number of errors on the training set). If it happens that at the stage of designing the network one constructs a network that is too complex (for the given amount of training data), the confidence interval $\Omega(\frac{\ell}{h^*})$ will be large. In this case, even if one could minimize the empirical risk as small as zero, the amount of errors on the test set can become big. This case is called *overfitting*. To avoid overfitting (and get a small confidence interval) one has to construct networks with small VC-dimension. Therefore, for generalizing well by using a neural network, one must first suggest an appropriate architecture of the neural network, and second, find in this network the function that minimizes the number of errors on the training data. For neural networks these two problems are solved by using some heuristics (see remarks on the back-propagation method).

In support vector methods one can control both parameters: in the separable case one obtains the unique solution that minimizes the empirical risk (down to zero) using a Δ-margin separating hyperplane with the maximal margin (i.e., subset with the smallest VC dimension). In the general case one obtains the unique solution when one chooses the value of the trade-off parameter C.

1.6 Conclusion

This chapter presented a very general overview of statistical learning theory. It demonstrates how an abstract analysis allows us to discover a general model of generalization.

According to this model, the generalization ability of learning machines depends on capacity concepts which are more sophisticated than merely the dimensionality of the space or the number of free parameters of the loss function (these concepts are the basis for the classical paradigm of generalization).

The new understanding of the mechanisms behind generalization not only changes the theoretical foundation of generalization (for example from this new viewpoint, the Occam razor principle is not always correct), but also changes the algorithmic approaches to function estimation problems. The approach described is rather general. It can be applied for various function estimation problems including regression, density estimation, solving inverse equations and so on.

Statistical Learning Theory started more than 30 years ago. The development of this theory did not involve many researchers. After the success of the SVM in solving

real life problems, the interest in statistical learning theory significantly increased. For the first time, abstract mathematical results in statistical learning theory have a direct impact on algorithmic tools of data analysis. In the last three years a lot of articles have appeared that analyze the theory of inference and the SVM method from different perspectives. These include:

1. Obtaining improved constructive bounds instead of the classical ones described in this chapter (which are more in the spirit of the non-constructive bound based on the Growth function than on bounds based on the VC dimension concept). Success in this direction could lead, in particular, to creating machines that generalize better than the SVM based on the concept of optimal hyperplane.

2. Extending the SVM ideology to many different problems of function and data-analysis.

3. Developing a theory that allows us to create kernels that possess desirable properties (for example that can enforce desirable invariants).

4. Developing a new type of induction inference that is based on direct generalization from the training set to the test set, avoiding the intermediate problem of estimating a function (the transductive type inference).

The hope is that this very fast growing area of research will significantly boost all branches of data analysis.

Bibliography

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, *Journal of the ACM* **44**(4) (1997) 617–631.

[2] P.L. Bartlett, P. Long, and R.C. Williamson, Fat-shattering and the learnability of real-valued functions, *Journal of Computer and System Sciences* **52**(3) (1996) 434–452.

[3] P.L. Bartlett and J. Shawe-Taylor, Generalization performance on support vector machines and other pattern classifiers, In B. Schölkopf, C. Burges, and A. Smola (ed) *Advances in Kernel Methods. Support Vector Learning*, The MIT Press (1999)

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth, Learnability and the Vapnik-Chervonenkis Dimension, *Journal of the ACM* **36**(4) (1989) 929–965.

[5] B. Boser, I. Guyon, and V.N. Vapnik, A training algorithm for optimal margin classifiers, *Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh ACM (1992) 144–152.

[6] C.J.C. Burges, Simplified support vector decision rule, *Proceedings of 13th Intl. Conf. on Machine Learning*, San Mateo, CA (1996) 71–77.

[7] C.J.C. Burges, Geometry and invariance in kernel based methods, In B. Schölkopf, C. Burges, and A. Smola (ed) *Advances in Kernel Methods. Support Vector Learning*, MIT Press (1999).

[8] C. Cortes and V.N. Vapnik, Support Vector Networks, *Machine Learning* **20** (1995) 273–297.

[9] L. Devroye, L. Gyorfi and G. Lugosi, *A Probability Theory of Pattern recognition*, Springer, N.Y. (1996).

[10] F. Girosi, An equivalence between sparse approximation and support vector machines, *Neural Computation* **10**(6) (1998) 1455–1480.

[11] F. Girosi, M. Jones, and T. Poggio, Regularization theory and neural networks architectures, *Neural Computation* **7**(2) (1995) 219–269.

[12] Y. Le Cun, Learning processes in an asymmetric threshold network, In E. Beinenstock, F. Fogelman-Soulie, and G. Weisbuch (ed) *Disordered systems and biological organizations*, Les Houches, France, Springer-Verlag (1986) 233–240.

[13] M.L. Minsky and S.A. Papert, *Perceptrons*, MIT Press (1969).

[14] M. Opper, On the annealed VC entropy for margin classifiers: a statistical mechanics study, In B. Schölkopf, C. Burges, and A. Smola (ed) *Advances in Kernel Methods, support Vector Learning,* MIT Press (1999).

[15] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, Learning internal representations by error propagation, *Parallel distributed processing: Explorations in macrostructure of cognition,* Vol. I, Badford Books, Cambridge, MA (1986) 318–362.

[16] J. Shawe-Taylor, P.L. Bartlett, R. C. Williamson, and M. Anthony, Structural risk minimization over data-dependent hierarchies, *IEEE Transactions on Information Theory* **44**(5) (1998) 1926–1940.

[17] B. Schölkopf, A Smola, and K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* **10** (1998) 1229–1319.

[18] A. Smola, B. Schölkopf and K-R. Müller, The connection between regularization operators and support vector kernels, *Neural Networks* **11** (1998) 637–649.

[19] M. Talagrand, The Glivenko-Cantelli problem, ten years later, *Journal of Theoretical Probability* **9**(2) (1996) 371–384.

[20] V.N. Vapnik, *Estimation of Dependencies Based on Empirical Data,* [in Russian], Nauka, Moscow (1979) (English translation: (1982) Springer-Verlag, New York).

[21] V.N. Vapnik, *The Nature of Statistical Learning Theory,* Springer-Verlag, New-York (1995).

[22] V.N. Vapnik, *Statistical Learning Theory,* John Wiley, New-York (1998).

[23] V.N. Vapnik and A.Ja. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Reports of Academy of Science USSR **181**(4) (1968).

[24] V.N. Vapnik and A.Ja. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl* **16** (1971) 264–280.

[25] V.N. Vapnik and A.Ja. Chervonenkis, *Theory of Pattern Recognition* [in Russian], Nauka, Moscow (1974) (German translation: W. N. Wapnik and A. Ja. Cherwonenkis (1979) *Theorie der Zeichenerkennung,* Akademia-Verlag, Berlin)

[26] V.N. Vapnik and A.Ja. Chervonenkis, Necessary and sufficient conditions for the uniform convergence of the means to their expectations, *Theory Probab. Appl.* **26** (1981) 532–553.

[27] V.N. Vapnik and A.Ja. Chervonenkis, The necessary and sufficient conditions for consistency of the method of empirical risk minimization [in Russian] (1989) *Yearbook of the Academy of Sciences of the USSR* on *Recognition, Classification, and Forecasting* Nauka Moscow **2** (1989) 217–249 (English translation: (1991) *Pattern Recogn. and Image Analysis,* Vol. 1, No. 3, 284-305).

[28] M. Vidyasagar, *A theory of learning and generalization,* Springer, N.Y. (1997).

[29] G. Wahba, *Spline Models for observational data,* SIAM Vol. 59, Philadelphia (1990).

[30] R.C. Williamson, A. Smola, and B. Schölkopf, Entropy numbers, operators, and support vector kernels, In B. Schölkopf, C. Burges, and A. Smola (ed) *Advances in Kernel Methods. Support Vector Learning.* The MIT Press (1999).