

Chapter 2

FALSIFIABILITY AND PARSIMONY: VC DIMENSION AND THE NUMBER OF ENTITIES (1980–2000)

2.1 SIMPLIFICATION OF VC THEORY

For about ten years this book did not attract much attention either in Russia or in the West. It attracted attention later.

In the meantime, in 1984 (five years after the publication of the original version of this book and two years after its English translation) an important event happened. Leslie Valiant published a paper where he described his vision of how learning theory should be built [122].

Valiant proposed the model that later was called the *Probably Approximately Correct* (PAC) learning model. In this model, the goal of learning is to find a rule that reasonably well approximates the best possible rule. One has to construct algorithms which guarantee that such a rule will be found with some probability (not necessarily one). In fact, the PAC model is one of the major statistical models of convergence, called consistency. It has been widely used in statistics since at least Fisher's time.

Nevertheless Valiant's article was a big success. In the mid-1980s the general machine learning community was not very well connected to statistics. Valiant introduced to this community the concept of consistency and demonstrated its usefulness. The theory of consistency of learning processes as well as generalization bounds was the subject of our 1968 and 1971 articles [143, 11], and was described in detail in our 1974 book [12, 173] devoted to pattern recognition, and in a more general setting in *EDBED*. However, at that time these results were not well known in the West.¹

¹In 1989 I met Valiant in Santa Cruz, and he told me that he did not know of our results when he wrote

In the 20th century, and especially in the second half of it, mass culture began to play an important role. For us it is important to discuss the “scientific component” of mass culture.

With the increasing role of science in everyday life, the general public began to discuss scientific discoveries in different areas: physical science, computer science (cybernetics), cognitive science (pattern recognition), biology (genomics), and philosophy. The discussions were held using very simplified scientific models that could be understood by the masses. Also scientists tried to appeal to the general public by promoting their philosophy using simplified models (for example, as has been done by Wiener). There is nothing wrong with this.

However, when science becomes a mass profession, the elements of the scientific mass culture in some cases start to substitute for the real scientific culture: It is much easier to learn the slogans of the scientific mass culture than it is to learn many different concepts from the original scientific sources. Science and “scientific mass culture,” however, are built on very different principles. In *Mathematical Discoveries*, Polya describes the principle of creating scientific mass culture observed by the remarkable mathematician Zermello. Here is the principle:

Gloss over the essentials and attract attention to the obvious.

Something that could remind this principle happened when (after appearance Valiant’s article) the adaptation of ideas described in EDBED started. In the PAC adaptation the VC theory was significantly simplified by removing its essential parts.

In *EDBED* the main idea was the necessary and sufficient aspects of the theory based on three capacity concepts: the VC entropy, the Growth function, and the VC dimension. It stresses that the most accurate bounds can be obtained based on the VC entropy concept. This, however, requires information about the probability measure. One can construct less accurate bounds that are valid for all probability measures. To do this one has to calculate the Growth function which can have a different form for different sets of admissible functions. The Growth function can be upper bounded by the standard function that depends on only one integer parameter (the VC dimension). This also decreases accuracy, but makes the bounds simpler.

These three levels of the theory provide different possibilities for further developments in learning technology. For example, one can try to create theory for the case when the probability measure belongs to some specific sets of measures (say smooth ones), or one can try to find a better upper bound for the Growth function using a standard function that depends on say two (or more) parameters. This can lead to more accurate estimates and therefore to more advanced algorithms. The important component of the theory described in *EDBED* was the structural risk minimization principle. It was considered to be the main driving force behind predictive learning technology.

PAC theory started just from the definition of the VC dimension based on the combinatorial lemma used to estimate the bound for the Growth function (see *EDBED*, Chapter 6, Section A2). The main effort was placed on obtaining VC type bounds for

his article, and that he even visited a conference at Moscow University to explain this to me. Unfortunately we never met in Moscow. After his article was published Valiant tried to find the computer science aspects of machine learning research suggesting analyzing the computational complexity of learning problems. In 1990 he wrote [123]: “If the computational requirements is removed from the definition then we are left with the notion of non-parametric inference in sense of statistics as discussed in particular by Vapnik [*EDBED*].”

different classes of functions (say for neural networks), and on the generalizations of the theory for the set of nonindicator functions. In most cases these generalizations were based on extensions of the VC dimension concept for real-valued functions made in the style described in *EDBED*. The exception was the fat-shattering concept [141] related to VC entropy for real-valued functions described in Chapter 7.

In the early 1990s, some PAC researchers started to attack the VC theory. First, the VC theory was declared a “worst-case theory” since it is based on the uniform convergence concept. In contrast to this “worst-case-theory” the development of “real-case theory” was announced. However, this is impossible (see Section 1.2 of this Afterword) since the (one-sided) uniform convergence forms the necessary and sufficient conditions for consistency of learning (that is also true for PAC learning). Then in the mid-1990s an attempt was made to rename the Vapnik–Chervonenkis lemma (*EDBED*, Chapter 6, Sections 8 and A2) as the Sauer lemma. For the first time we published the formulation of this lemma in 1968 in the *Reports of the Academy of Sciences of USSR* [143]. In 1971, we published the corresponding proofs in the article devoted to the uniform law of large numbers [11]. In 1972, two mathematicians N. Sauer [130] and S. Shelah [131] independently proved this combinatorial lemma.

Researchers, who in the 1980s learned from *EDBED* (or from our articles) both the lemma and its role in statistical learning theory, renamed it in the 1990s.² Why? My speculation is that renaming it was important for creating the following legend:

In 1984 the PAC model was introduced. Early in statistics a concept called the VC dimension was developed. This concept plays an important role in the Sauer lemma, which is a key instrument in PAC theory.

Now, due to new developments in the VC theory and the interest in the advanced topics of statistical learning theory, this legend has died, and as a result interest in PAC theory has significantly decreased.

This is, however, a shame because the computational complexity aspects of learning stressed by Valiant remain relevant.

2.2 CAPACITY CONTROL

2.2.1 BELL LABS

In 1990 Larry Jackel, the head of the Adaptive Systems Research Department at AT&T Bell Labs, invited me to spend half a year with his group. It was a time of wide discussions on the VC dimension concept and its relationship to generalization problems. The obvious interpretation of the VC dimension was the number of free parameters that led to the curse of dimensionality. John Denker, a member of this department, showed,

²N. Sauer did not have in mind statistics proving this lemma. This is the content of the abstract of his article: “P. Erdős (oral communication) transmitted to me in Nice the following question: . . . (*the formulation of the lemma*). . . . In this paper we will answer this question in the affirmative by determining the exact upper bounds.”

however, that the VC dimension is not necessarily the number of free parameters. He came up with the example

$$y = \theta\{\sin ax\}, x \in R^1, a \in (0, \infty)$$

a set of indicator functions that has only one free parameter yet possesses an infinite VC dimension (see Section 2.7.5, footnote 7). In *EDBED* another situation was described: when the VC dimension was smaller than the number of free parameters. These intriguing facts could lead to new developments in learning theory.

Our department had twelve researchers. Six of them, L. Jackel, J. Denker, S. Solla, C. Burges, G. Nohl, and H.P. Graf were physicists, and six, Y. LeCun, L. Bottou, P. Simard, I. Guyon, B. Boser, and Y. Bengio were computer scientists. The main direction of research was to advance the understanding of pattern recognition phenomena. To do this they relied on the principles of research common in physics.

The main principle of research in physics can be thought of as the complete opposite of the Zermello principle for creating scientific mass culture. It can be formulated as follows:

Find the essential in the nonobvious.

The entire story of creating modern technology can be seen as an illustration of this principle. At the time when electricity, electromagnetic waves, annihilation, and other physical fundamentals were discovered they seemed to be insignificant elements of nature. It took a lot of joint efforts of theorists, experimental physicists and engineers to prove that these negligible artifacts are very important parts of nature and make it work.

The examples given by Denker and another one described in *EDBED* (see Chapter 10, Section 5) could be an indication that such a situation in machine learning is quite possible.

The goal of our department was to understand and advance new general principles of learning that are effective for solving real-life problems. As a model problem for on-going experiments, the department focused on developing automatic systems that could read handwritten digits. This task was chosen for a number of reasons. First, it was known to be a difficult problem, with traditional machine vision approaches making only slow progress. Second, lots of data were available for training and testing. And third, accurate solutions to the problem would have significant commercial importance.

Initial success in our research department led to the creation of a development group supervised by Charlie Stenard. This group, which worked closely with us, had as a goal the construction of a machine for banks that could read handwritten checks from all over the world. Such a machine could not make too many errors (the number of errors should be comparable to the number made by humans). However, the machine could refuse to read some percentage of checks.

I spent ten years with this department. During this time check reading machines became an important instrument in the banking industry. About 10% of checks in US banks are read by technology developed at Bell Labs.

During these years the performance of digit recognition was significantly improved. However, it *never* happened that significant improvements in quality of classification were the results of smart engineering heuristics. All jumps in performance were results of advances in understanding fundamentals of the pattern recognition problem.

2.2.2 NEURAL NETWORKS

When I joined the department, the main instrument for pattern recognition was neural networks constructed by Yann LeCun, one of the originators of neural networks. For the digit recognition problem he designed a series of convolutional networks called LeNet. In the early 1990s this was a revolutionary idea. The traditional scheme of applying pattern recognition techniques was the following: a researcher constructs several very carefully crafted features and uses them as inputs for a statistical parametric model. To construct the desired rule they estimated the parameters of this model. Therefore good rules in many respects reflected how smart the researcher was in constructing features.

LeNet uses as input a high-dimensional vector whose coordinates are the raw image pixels. This vector is processed using a multilayer convolutional network with many free parameters. Using the back propagation technique, LeNet tunes the parameters to minimize the training loss.³

For the digit recognition problem, the rules obtained by LeNet were significantly better than any rules obtained by the classical style algorithms. This taught a great lesson: one does not need to go into the details of the decision rule; it is enough to create an “appropriate architecture” and an “appropriate minimization method” to solve the problem.

2.2.3 NEURAL NETWORKS: THE CHALLENGE

The success of neural nets in solving pattern recognition problems was a challenge for theorists. Here is why. When one is trying to understand how the brain is working two different questions arise:

- (1) What happens? What are the principles of generalization that the brain executes?
- (2) How does it happen? How does the brain execute these principles?

Neural networks attempt to answer the second question using an artificial brain model motivated by neurophysiologists.

According to the VC theory, however, this is not very important. VC theory declares that two and only two factors are responsible for generalization. They are the value of empirical loss, and the capacity of the admissible set of functions (the VC entropy, Growth function, or the VC dimension). The SRM principle states that any method that controls these two factors well (minimizing the right-hand side of the VC bounds) is strongly universally consistent.

It was clear that artificial neural networks executed the structural risk minimization principle. However, they seemed to do this rather inefficiently. Indeed, the loss function that artificial neural networks minimize has many local minima. One can guarantee convergence to one of these minima but cannot guarantee good generalization. Neural networks practitioners define some initial conditions that they believe will lead to a

³As computer power increased, LeCun constructed more powerful generations of LeNet.

“good” minimum. Also, the back-propagation method based on the gradient procedure of minimization in high-dimensional spaces requires a very subtle treatment of step values. The choice of these values does not have a good recommendation.

In order to control capacity the designer chooses an appropriate number of elements (neurons) for the networks. Therefore for different training data sizes one has to design different neural networks. All these factors make neural networks more of an art than a science.

Several ideas that tried to overcome the described shortcomings of neural networks were checked during 1991 and 1992 including measuring the VC dimension (capacity) of the learning machine [144, 142] and constructing local learning rules [145]. Now these ideas are developing in a new situation. However, in 1992 they were overshadowed by a new learning concept called Support Vector Machines (SVMs).

2.3 SUPPORT VECTOR MACHINES (SVMs)

The development of SVMs has a 30-year history, from 1965 until 1995. It was completed in three major steps.

2.3.1 STEP ONE: THE OPTIMAL SEPARATING HYPERPLANE

In 1964, Chervonenkis and I came up with an algorithm for constructing an optimal separating hyperplane called the generalized portrait method. Three chapters of our 1974 book *Theory of pattern recognition*, contain the detailed theory of this algorithm [12, 173]. In *EDBED* (Addendum I), a simplified version of this algorithm is given. Here are more details. The problem was: given the training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell), \quad (2.1)$$

construct the hyperplane

$$(w_0, x) + b_0 = 0 \quad (2.2)$$

that separates these data and has the largest margin. In our 1974 book and in *EDBED* we assumed that the data were separable. The generalization of this algorithm for constructing an optimal hyperplane in the nonseparable case was introduced in 1995 [132]. We will discuss it in a later section.

Thus, the goal was to maximize the functional

$$\rho_0 = \min_{\{i: y_i=1\}} \left[\left(\frac{w}{|w|}, x_i \right) + b \right] - \max_{\{j: y_j=-1\}} \left[\left(\frac{w}{|w|}, x_j \right) + b \right]$$

under the constraints

$$y_i((w, x_i) + b) \geq 1, \quad i = 1, \dots, \ell. \quad (2.3)$$

It is easy to see that this problem is equivalent to finding the minimum of the quadratic form

$$R_1(w, b) = (w, w)$$

subject to the linear constraints (2.3). Let this minimum be achieved when $w = w_0$. Then

$$\rho_0 = \frac{2}{\sqrt{(w_0, w_0)}}.$$

To minimize the functional (w, w) subject to constraints (2.3) the standard Lagrange optimization technique was used. The Lagrangian

$$L(\alpha) = \frac{1}{2}(w, w) - \sum_{i=1}^{\ell} \alpha_i ([y_i((w, x_i) + b) - 1]) \quad (2.4)$$

(where $\alpha_i \geq 0$ are the Lagrange multipliers) was constructed and its minimax (minimum over w and b and maximum over the multipliers $\alpha_i \geq 0$) was found. The solution of this quadratic optimization problem has the form

$$w_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i. \quad (2.5)$$

To find these coefficients one has to maximize the functional:

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (2.6)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \dots, \ell.$$

Substituting (2.5) back into (2.2) we obtain the separating hyperplane expressed in terms of the Lagrange multipliers

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 (x, x_i) + b_0 = 0. \quad (2.7)$$

2.3.2 THE VC DIMENSION OF THE SET OF ρ -MARGIN SEPARATING HYPERPLANES

The following fact plays an important role in SVM theory. Let the vectors $x \in R^n$ belong to the sphere of radius $R = 1$. Then the VC dimension h of the set of hyperplanes with margin $\rho_0 = (w_0, w_0)^{-1}$ has the bound

$$h \leq \min\{(w_0, w_0), n\} + 1.$$

That is, the VC dimension is defined by the smallest of the two values: the dimensionality n of the vectors x and the value (w_0, w_0) . In Hilbert (infinite dimensional) space,

the VC dimension of the set of separating hyperplanes with the margin ρ_0 depends just on the value (w_0, w_0) .

In *EDBED* I gave a geometrical proof of the bound (See Chapter 10, Section 5). In 1997, Gurvits found an algebraic proof [124]. Therefore, the optimal separating hyperplane executes the SRM principle: it minimizes (to zero) the empirical loss, using the separating hyperplane that belongs to the set with the smallest VC dimension.

One can therefore introduce the following learning machine that executes the SRM principle:

Map input vectors $x \in X$ into (a rich) Hilbert space $z \in Z$, and construct the maximal margin hyperplane in this space.

According to the VC theory the generalization bounds depend on the VC dimension. Therefore by controlling the margin of the separating hyperplane one controls the generalization ability.

2.3.3 STEP TWO: CAPACITY CONTROL IN HILBERT SPACE

The formal implementation of this idea requires one to specify the operator

$$z = \mathcal{F}x$$

which should be used for mapping. Then similar to (2.7) one constructs the separating hyperplane in image space

$$\sum_{i=1}^{\ell} y_i \alpha_i^0(z, z_i) + b_0 = 0,$$

where the coefficients $\alpha_i \geq 0$ are the ones that maximize the quadratic form

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (z_i, z_j) \quad (2.8)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \dots, \ell. \quad (2.9)$$

In 1992 Boser, Guyon and I found an effective way to construct the separating hyperplane in Hilbert space without explicitly mapping the input vectors x into vectors z of the Hilbert space [125].

This was done using Mercer's theorem.

Let vectors $x \in X$ be mapped into vectors $z \in Z$ of some Hilbert space.

1. *Then there exists in X space a symmetric positive definite function $K(x_i, x_j)$ that defines the corresponding inner product in Z space:*

$$(z_i, z_j) = K(x_i, x_j).$$

2. Also, for any symmetric positive definite function $K(x_i, x_j)$ in X space there exists a mapping from X to Z such that this function defines an inner product in Z space.

Therefore, according to Mercer's theorem, the separating hyperplane in image space has the form

$$\sum_{i=1}^{\ell} y_i \alpha_i^0 K(x, x_i) + b_0 = 0,$$

where the coefficients α_i^0 are defined as the solution of the quadratic optimization problem: maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2.10)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \dots, \ell. \quad (2.11)$$

Choosing specific kernel functions $K(x_i, x_j)$ one makes specific mappings from input vectors x into image vectors z .

The idea of using Mercer's theorem to map into Hilbert space was used in the mid-1960s by Aizerman, Braverman, and Rozonoer [2]. Thirty years later we used this idea in a wider context.

2.3.4 STEP THREE: SUPPORT VECTOR MACHINES

In 1995 Cortes and I generalized the maximal margin idea for constructing (in image space) the hyperplane

$$(w_0, z) + b_0 = 0$$

when the training data are nonseparable [132]. This technology became known as Support Vector Machines (SVMs). To construct such a hyperplane we follow the recommendations of the SRM principle.

Problem 1. Choose among the set hyperplanes with the predefined margin

$$\rho^2 = \frac{4}{(w_0, w_0)} \leq H = \frac{1}{h}$$

the one that separates the images of the training data with the smallest number of errors. That is, we minimize the functional

$$R = \sum_{i=1}^{\ell} \theta(\xi_i) \quad (2.12)$$

subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell \quad (2.13)$$

and the constraint

$$(w, w) \leq h, \quad (2.14)$$

where $\theta(u)$ is the step function:

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

For computational reasons, however, we approximate *Problem 1* with the following one.

Problem 2. Minimize the functional

$$R = \sum_{i=1}^{\ell} \xi_i \quad (2.15)$$

(instead of the functional (2.12)) subject to the constraints (2.13) and (2.14).

Using the Lagrange multiplier technique, one can show that the corresponding hyperplane has an expansion

$$\sum_{i=1}^{\ell} y_i \alpha_i^0(z_i, z) + b_0 = 0. \quad (2.16)$$

To find the multipliers one has to maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - h \sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (z_i, z_j)} \quad (2.17)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \quad (2.18)$$

and the constraints

$$0 \leq \alpha_i \leq 1, \quad i = 1, \dots, \ell.$$

Problem 3. Problem 2 is equivalent to the following (reparametrized) one: Minimize the functional

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} \xi_i \quad (2.19)$$

subject to constraints (2.13). This setting implies the following dual space solution: Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (z_i, z_j) \quad (2.20)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \quad (2.21)$$

and the constraints

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell.$$

One can show that for any h there exists a C such that the solutions of Problem 2 and Problem 3 coincide. From a computational point of view Problem 3 is simpler than Problem 2. However, in Problem 2 the parameter h estimates the VC dimension. Since the VC bound depends on the ratio h/ℓ one can choose the VC dimension to be some fraction of the training data, while in the reparametrized Problem 3 the corresponding parameter C cannot be specified; it can be any value depending on the VC dimension and the particular data.

Taking into account Mercer's theorem,

$$(z_i, z_j) = K(x_i, x_j),$$

we can rewrite the nonlinear separating rule in input space X as

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + b_0 = 0, \quad (2.22)$$

where the coefficients are the solution of the following problems:

Problem 1a. Minimize the functional

$$R = \sum_{i=1}^{\ell} \theta(\xi_i) \quad (2.23)$$

subject to the constraints

$$y_i \sum_{j=1}^{\ell} (y_j \alpha_j K(x_j, x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell \quad (2.24)$$

and the constraint

$$\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \leq h. \quad (2.25)$$

Problem 2a. Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - h \sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j)} \quad (2.26)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \quad (2.27)$$

and the constraints

$$0 \leq \alpha_i \leq 1, \quad i = 1, \dots, \ell. \quad (2.28)$$

Problem 3a. Maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad (2.29)$$

subject to the constraint

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

and the constraints

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell.$$

The solution of Problem 3a became the standard SVM method. In this solution only some of the coefficients α_i^0 are different from zero. The vectors x_i for which $\alpha_i^0 \neq 0$ in (2.22) are called the *support vectors*. Therefore, the separating rule (2.22) is the expansion on the support vectors.

To construct a support vector machine one can use any (conditionally) positive definite function $K(x_i, x_j)$ creating different types of SVMs. One can even use kernels in the situation when input vectors belong to nonvectorial spaces. For example, the inputs may be sequences of symbols of different size (as in problems of bioinformatics or text classification). Therefore SVMs form a universal generalization engine that can be used for different problems of interest.

Two examples of Mercer kernels are the polynomial kernel of degree d

$$K(x_i, x_j) = ((x_i, x_j) + c)^d, \quad c \geq 0 \quad (2.30)$$

and the exponential kernel

$$K(x_i, x_j) = \exp \left\{ - \left(\frac{|x_i - x_j|}{\sigma} \right)^d \right\}, \quad \sigma > 0, \quad 0 \leq d \leq 2. \quad (2.31)$$

2.3.5 SVMs AND NONPARAMETRIC STATISTICAL METHODS

SVMs execute the idea of the structural risk minimization principle, where the choice of the appropriate element of the structure is defined by the constant C (and a kernel parameters). Therefore, theoretically, for any appropriate kernel (say for (2.31) by controlling parameters (which depends on the training data) one guarantees asymptotic convergence of the SVM solutions to the best possible solution [167].

In 1980 Devroye and Wagner proved that classical nonparametric methods of density estimation are also universally consistent [134]. That is, by controlling the parameter $\sigma_\ell = \sigma(\ell) > 0$ depending on the size ℓ of the training data, the following approximation of the density function

$$\bar{p}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{(2\pi)^{n/2} \sigma_\ell^n} \exp \left\{ - \left(\frac{|x_i - x|}{\sigma_\ell} \right)^2 \right\} \quad (2.32)$$

converges (in the uniform metric) to the desired density *with increasing* ℓ .

However, by choosing an appropriate parameter C of SVM, one controls the VC bound for any finite number of observations. One can also control these bounds by choosing the parameters of the kernels.

This section illustrates the practical advantage of this fact.

Let us use the nonparametric density estimation method to approximate the optimal (generative) decision rule for binary classification

$$p_1(x) - p_2(x) = 0, \quad (2.33)$$

where $p_1(x)$ is the density function of the vectors belonging to the first class and $p_2(x)$ is the density function of the vectors belonging to the second class. Here for notational simplicity we assume that the two classes are equally likely and that the number of training samples from the first and second class is the same. Using (2.32) the approximation (2.33) can be rewritten as follows.

$$\sum_{\{i: y_i=1\}} \exp \left\{ - \left(\frac{|x - x_i|}{\sigma} \right)^2 \right\} - \sum_{\{j: y_j=-1\}} \exp \left\{ - \left(\frac{|x - x_j|}{\sigma} \right)^2 \right\} = 0.$$

The SVM solution using the same kernel has the form

$$\sum_{\{i: y_i=1\}} \alpha_i \exp \left\{ - \left(\frac{|x - x_i|}{\sigma} \right)^2 \right\} - \sum_{\{j: y_j=-1\}} \alpha_j \exp \left\{ - \left(\frac{|x - x_j|}{\sigma} \right)^2 \right\} = 0.$$

Since our kernel is a positive definite function there exists a space Z where it defines an inner product (by the second part of Mercer's theorem). In Z space both solutions define separating hyperplanes

$$\sum_{\{i: y_i=1\}} (z_i, z) - \sum_{\{j: y_j=-1\}} (z_j, z) = 0$$

(the classical non-parametric solution) [152] and

$$\sum_{\{i: y_i=1\}} \alpha_i (z_i, z) - \sum_{\{j: y_j=-1\}} \alpha_j (z_j, z) = 0.$$

(the SVM solution). Figure 2.1 shows these solutions in Z space. The separating hyperplane obtained by nonparametric statistics is defined by the hyperplane orthogonal

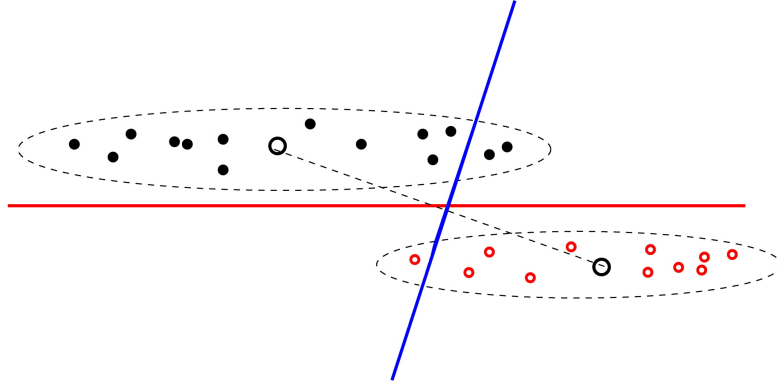


Figure 2.1: Classifications given by the classical nonparametric method and the SVM are very different.

to the line connecting the center of mass of two different classes. The SVM produces the optimal separating hyperplane.

In spite of the fact that both solutions converge asymptotically to the best one⁴ they are very different for a fixed number of training data since the SVM solution is optimal (for any number of observations it guarantees the smallest predictive loss), while the non-parametric technique is not.

This makes SVM a state-of-the-art technology in solving real-life problems.

2.4 AN EXTENSION OF SVMs: SVM+

In this section we consider a new algorithm called SVM+, which is an extension of SVM. SVM+ takes into account a known structure of the given data.

2.4.1 BASIC EXTENSION OF SVMs

Suppose that our data are the union of $t \geq 1$ groups:

$$(X, Y)_r = (x_{r_1}, y_{r_1}), \dots, (x_{r_{n_r}}, y_{r_{n_r}}), \quad r = 1, \dots, t.$$

Let us denote indices from the group r by

$$T_r = \{i_{n_1}, \dots, i_{n_r}\}, \quad r = 1, \dots, t.$$

⁴Note that nonparametric density estimate (2.32) requires dependence of σ from ℓ . Therefore, it uses different Z spaces for different ℓ .

Let inside one group the slacks be defined by some correcting function that belongs to a given set of functions

$$\xi_i = \xi_r(x_i) = \phi_r(x_i, w_r), \quad w_r \in W_r, \quad i \in T_r, \quad r = 1, \dots, t. \quad (2.34)$$

The goal is to define the decision function for a situation when sets of admissible correcting functions are restricted (when sets of admissible correcting functions are not restricted we are back to conventional SVM). By introducing groups of data and different sets of correcting functions for different groups one introduces additional information about the problem to be solved.

To define the correcting function $\xi(x) = \phi_r(x, w_r)$ for group T_r we map the input vectors $x_i, i \in T_r$ simultaneously into two different Hilbert spaces: into the space $z_i \in Z$ which defines the decision function (as we did for the conventional SVM) and into correcting function space $z_i^r \in Z_r$ which defines the set of correcting functions for a given group r . (Note that vectors of different groups are mapped into the same decision space Z but different correcting spaces Z_r .)

Let the inner products in the corresponding spaces be defined by the kernels

$$(z_i, z_j) = K(x_i, x_j), \quad \forall i, j$$

and

$$(z_i^r, z_j^r) = K_r(x_i, x_j), \quad i, j \in T_r, \quad r = 1, \dots, t. \quad (2.35)$$

Let the set of admissible correcting functions $\xi_r(x) = \phi_r(x, w_r), w_r \in W_r$, be linear in each Z_r space

$$\xi(x_i) = \phi_r(x, w_r) = [(w_r, z_i^r) + d_r] \geq 0, \quad i \in T_r, \quad r = 1, \dots, t. \quad (2.36)$$

As before our goal is to find the separating hyperplane in decision space Z ,

$$(w_0, z) + b_0 = 0$$

whose parameters w_0 and b_0 minimize the functional

$$R(w, w_1, \dots, w_t) = \frac{1}{2}(w, w) + C \sum_{r=1}^t \sum_{i \in T_r} ((w_r, z_i^r) + d_r), \quad (2.37)$$

subject to the constraints

$$y_i[(z_i, w) + b] \geq 1 - ((z_i^r, w_r) + d_r), \quad i \in T_r, \quad r = 1, \dots, t \quad (2.38)$$

and the constraints

$$(w_r, z_i^r) + d_r \geq 0, \quad i \in T_r, \quad r = 1, \dots, t. \quad (2.39)$$

Note that for set (2.36) the solution of this optimization problem does exist.

The corresponding Lagrangian is

$$L(w, w_1, \dots, w_t; \alpha, \mu) = \frac{1}{2}(w, w) + C \sum_{r=1}^t \sum_{i \in T_r} ((w_r, z_i^r) + d_r) \quad (2.40)$$

$$-\sum_{i=1}^{\ell} \alpha_i [y_i((w, z_i) + b) - 1 + d_r + (w_r, z_i^r)] - \sum_{i=1}^{\ell} \mu_i ((w_r, z_i^r) + d_r).$$

Using the same dual optimization technique as above one can show that the optimal separating hyperplane in Z space has the form

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i(z_i, z) + b_0 = 0,$$

where the coefficients $\alpha_i^0 \geq 0$ minimize the same quadratic form as before

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad (2.41)$$

subject to the conventional constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \quad (2.42)$$

and the new constraints

$$\sum_{i \in T_r} (\alpha_i + \mu_i) = |T_r| C, \quad r = 1, \dots, t \quad (2.43)$$

($|T_r|$ is the number of elements in T_r),

$$\sum_{i \in T_r} (\alpha_i + \mu_i) K_r(x_i, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j), \quad j \in T_r, \quad r = 1, \dots, t. \quad (2.44)$$

and constraints

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

When either

(1) There is no structure in the data: any vector belongs to its own group,

or

(2) There is no correlation between slacks inside all groups: $K_r(x_i, x_j)$ is an identity matrix for all r

$$K_r(x_i, x_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \quad (2.45)$$

then Equation (2.44) defines the box constraints as in conventional SVMs (in case (2) Equations (2.43) are satisfied automatically). Therefore the SVM+ model contains the classical SVM model as a particular case.

The advantage of the SVM+ is the ability to consider the global structure of the training data that the conventional SVM ignores.

This, however, requires solving a more general quadratic optimization problem to minimize in the space of 2ℓ nonnegative variables the same objective function subject to $(\ell + t + 1)$ linear constraints (instead of one minimizing this objective function in the space of ℓ variable subjects of one linear constraint and ℓ box constraints in the conventional SVM).

2.4.2 ANOTHER EXTENSION OF SVM: SVM_γ+

Consider another extension of SVM, the so-called SVM_γ+, which directly controls the capacity of sets of correcting functions.

Let us instead of objective function (2.37) consider the function

$$R(w, w_1, \dots, w_t) = \frac{1}{2}(w, w) + \frac{\gamma}{2} \sum_{r=1}^t (w_r, w_r) + C \sum_{r=1}^t \sum_{i \in T_r} ((w_r, z_i^r) + d_r), \quad (2.46)$$

where $\gamma > 0$ is some value. When γ approaches zero (2.46) and (2.37) coincide.

The SVM_γ+ solution minimizes functional (2.46) subject to the constraints (2.38) and (2.39). To solve this problem we construct the Lagrangian. Comparing it to (2.40), this Lagrangian has one extra term $\gamma/2 \sum (w_r, w_r)$. Repeating almost the same algebra as in the previous section we obtain that for the modified Lagrangian the dual space solution that defines the coefficients α_i^0 must maximize the functional

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

$$\frac{C}{\gamma} \sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i + \mu_i) K_r(x_i, x_j) - \frac{1}{2\gamma} \sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i + \mu_i)(\alpha_j + \mu_j) K_r(x_i, x_j)$$

subject to the constraints (2.42) and the constraints

$$\sum_{i \in T_r} (\alpha_i + \mu_i) = |T_r|C, \quad r = 1, \dots, t,$$

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

Note that when either:

(1) There is no structure (every training vector belongs to its own group),

or

(2) There is no correlation inside groups ((2.45) holds for all r) and $\gamma \rightarrow 0$

then the SVM_γ+ solution coincides with the conventional SVM solution.

This solution requires maximizing the quadratic objective function in the space of 2ℓ nonnegative variables subject to $t + 1$ equality constraints.

One can simplify the computation when using models of correcting functions (2.36) with $d_r = 0$, $r = 1, \dots, t$. In this case one has to maximize the functional $W(\alpha, \mu)$ over non-negative variables α_i, μ_i , $i = 1, \dots, \ell$ subject to one equality constraint (2.42).

2.4.3 LEARNING USING HIDDEN INFORMATION

SVM+ is an instrument for a new inference technology which can be called *learning using hidden information*. It allows one to extract additional information in situations where conditional technologies cannot be used.

WHAT INFORMATION CAN BE HIDDEN?

Consider the pattern recognition problem. Let one be given the training set

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

Suppose that one can add to this set additional information from two sources:

- (1) information that exists in *hidden classifications* of the training set and
- (2) information that exists in *hidden variables* of the training set.

The next two examples describe such situations.

EXAMPLE 1 (Information given in hidden classifications).

Suppose that one's goal is to find a rule that separates cancer patients from non cancer patients. One collects training data and assigns class $y_i = 1$, or $y_i = -1$, to patient x_i depending on the result of analysis of tissue taken during surgery. Analyzing the tissue, a doctor composes a report which not only concludes that the patient has a cancer (+1) or benign diagnosis (-1) but also that the patient belongs to a particular group (has a specific type of cancer or has a specific type of cell and so on). That is, the doctor's classification of the training data y_i^* is more detailed than the desired classification y_i . When constructing a classification rule $y = f(x)$, one can take into account information about y_i^* . This information can be used, for example, to create appropriate groups.

EXAMPLE 2 (Information given in hidden variables).

Suppose that one's goal is to construct a rule $y = f(x)$. However, for the training data along with the nonhidden variables x_i , one can determine the hidden variables x_i^* . The problem is using the data

$$(x_1, x_1^*, y_1), \dots, (x_\ell, x_\ell^*, y_\ell)$$

which contain both nonhidden and hidden variables and their classifications y_i , to construct a rule $y = f(x)$ (rather than a rule $y = f(x, x^*)$) that makes a prediction based on nonhidden variables. By using variables x for a decision space and variables x, x^* for a correcting space one can solve this problem.

EXAMPLE 3 (Special rule for selected features).

A particular case of the problem described in Example 2 is constructing a decision rule for selected features, using information about the whole set of features. In this problem, the selected features are considered as non hidden variables while the rest of the features are hidden variables.

THE GENERAL PROBLEM

How should one construct (a more accurate than conventional) rule $y = f(x)$ using the data

$$(x_1, x_1^*, y_1, y_1^*), \dots, (x_\ell, x_\ell^*, y_\ell, y_\ell^*)$$

instead of the data

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

To do this one can use the SVM+ method. Constructing the desired decision rule in the solution space, SVM+ uses two new ideas:

- (1) It uses structure on training data and
- (2) It uses several different spaces: (a) the solution space of nonhidden variables and (b) the correcting spaces of joint hidden and nonhidden variables.

SVM+ allows one to effectively use additional (hidden) information. The success of SVM+ depends on the quality of recovered hidden information.

The corresponding SVM+ technology requires the following three steps:

1. Use the data (x_i, x_i^*, y_i, y_i^*) for constructing a structure on the training set.
2. Use the kernel $K(x_i, x_j)$ for constructing a rule in the decision space, and
3. Use the kernels $K_r(x_i, x_i^*; x_j, x_j^*)$ in the correcting spaces.

Note that in the SVM+ method the idea of creating a structure on the training set differs from the classical idea of clustering of the training set.

2.5 GENERALIZATION FOR REGRESSION ESTIMATION PROBLEM

In this section we use the ε -insensitive loss function introduced in [140],

$$u_\varepsilon = \begin{cases} |u| - \varepsilon, & \text{if } |u| \geq \varepsilon \\ 0, & \text{if } |u| < \varepsilon. \end{cases}$$

This function allows one to transfer some properties of the SVM for pattern recognition (the accuracy and the sparsity) to the regression problem.

2.5.1 SVM REGRESSION

Consider the regression problem: given iid data

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

where $x \in X$ is a vector and $y \in (-\infty, \infty)$ is a real value, estimate the function in a given set of real-valued functions.

As before using kernel techniques we map input vectors x into the space of image vectors $z \in Z$ and approximate the regression by a linear function

$$y = (w, z) + b, \tag{2.47}$$

where w and b have to be defined. Our goal is to minimize the following loss,

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} |y_i - (w, z) - b|_\varepsilon. \tag{2.48}$$

To minimize the functional (2.48) we solve the following equivalent problem [140]:
Minimize the functional

$$R = \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \quad (2.49)$$

subject to the constraints

$$y_i - (w, z_i) - b \leq \varepsilon + \xi_i^*, \quad \xi_i^* \geq 0, \quad i = 1, \dots, \ell, \quad (2.50)$$

$$(w, z_i) + b - y_i \leq \varepsilon + \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell. \quad (2.51)$$

To solve this problem one constructs the Lagrangian

$$\begin{aligned} L = & \frac{1}{2}(w, w) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} \alpha_i [y_i - (w, z_i) - b + \varepsilon + \xi_i] \\ & - \sum_{i=1}^{\ell} \alpha_i^* [(w, z_i) + b - y_i + \varepsilon + \xi_i^*] - \sum_{i=1}^{\ell} (\beta_i \xi_i + \beta_i^* \xi_i^*) \end{aligned} \quad (2.52)$$

whose minimum over w , b , and ξ , ξ_i^* leads to the equations

$$w = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) z_i, \quad (2.53)$$

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0, \quad (2.54)$$

and

$$\alpha_i^* + \beta_i^* = C, \quad \alpha_i + \beta_i = C, \quad (2.55)$$

where α , α^* , β , $\beta^* \geq 0$ are the Lagrange multipliers. Putting (2.53) into (2.47) we obtain that in X space the desired function has the kernel form

$$y = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) K(x_i, x) + b. \quad (2.56)$$

To find the Lagrange multipliers one has to put the obtained equation back into the Lagrangian and maximize the obtained expression.

Putting (2.53), (2.54), and (2.55) back into (2.52) we obtain

$$W = - \sum_{i=1}^{\ell} \varepsilon (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j}^{\ell} (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) K(x_i, x_j). \quad (2.57)$$

To find α_i , α_i^* for the approximation (2.56) one has to maximize this functional subject to the constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i^* &= \sum_{i=1}^{\ell} \alpha_i, \\ 0 \leq \alpha_i \leq C, \quad 0 \leq \alpha_i^* \leq C, \quad i &= 1, \dots, \ell. \end{aligned}$$

2.5.2 SVM+ REGRESSION

Now let us solve the same regression problem of minimizing the functional (2.49) subject to the constraints (2.50) and (2.51) in the situation when the slacks ξ_i and ξ_i^* are defined by functions from the set described in Section 2.4:

$$\xi_i = \phi_r(x_i, w_r) = (w_r, z_i) - d_r \geq 0, \quad i \in T_r, \quad r = 1, \dots, t \quad (2.58)$$

$$\xi_i^* = \phi_r^*(x_i, w_r^*) = (w_r^*, z_i) - d_r^* \geq 0, \quad i \in T_r, \quad r = 1, \dots, t. \quad (2.59)$$

To find the regression we construct the Lagrangian similar to (2.52) where instead of slacks ξ_i and ξ_i^* we use their expressions (2.58) and (2.59).

Minimizing this Lagrangian over w, b (as before) and over $w_r, d_r, w_r^*, d_r^*, r = 1, \dots, t$ (instead of slacks ξ_i , and ξ_i^*) we obtain Equations (2.53) and (2.54) and the equations

$$\sum_{i \in T_r} (\alpha_i + \beta_i) z_i^r = C \sum_{i \in T_r} z_i^r, \quad \sum_{i \in T_r} (\alpha_i^* + \beta_i^*) z_i^r = C \sum_{i \in T_r} z_i^r, \quad r = 1, \dots, t, \quad (2.60)$$

$$\sum_{i \in T_r} (\alpha_i^* + \beta_i^*) = C|T_r|, \quad \sum_{i \in T_r} (\alpha_i + \beta_i) = C|T_r|, \quad r = 1, \dots, t \quad (2.61)$$

Putting these equations back into the Lagrangian we obtain

$$W = - \sum_{i=1}^{\ell} \varepsilon (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j}^{\ell} (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) K(x_i, x_j). \quad (2.62)$$

From (2.60) and (2.61) we obtain

$$\sum_{i \in T_r} (\alpha_i + \beta_i) K_r(x_j, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j), \quad r = 1, \dots, t, \quad j \in T_r, \quad (2.63)$$

$$\sum_{i \in T_r} (\alpha_i^* + \beta_i^*) K_r(x_i, x_j) = C \sum_{i \in T_r} K_r(x_i, x_j), \quad r = 1, \dots, t, \quad j \in T_r, \quad (2.64)$$

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

Therefore to estimate the SVM+ regression function (2.56) one has to maximize the functional (2.62) subject to the constraints (2.54), (2.61), (2.63), (2.64).

2.5.3 SVM_γ+ REGRESSION

Consider SVM_γ+ extension of regression estimation problem: Minimize the functional

$$R = \frac{1}{2} (w, w) + \frac{\gamma}{2} \left(\sum_{r=1}^t (w_r, w_r) + \sum_{r=1}^t (w_r^*, w_r^*) \right) + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \quad (2.65)$$

(instead of functional (2.49)) subject to constraints (2.50) and (2.51), where slacks ξ_i and ξ_i^* are defined by the correcting functions (2.58) and (2.59). The new objective function approaches (2.49) when γ approaches zero.

The same algebra of the Lagrange multiplier technique that was used above now implies that to find the coefficients α_i , α_i^* for approximation (2.56) one has to maximize the functional

$$\begin{aligned} W = & - \sum_{i=1}^{\ell} \varepsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) + \\ & \frac{C}{\gamma} \sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i + \beta_i) K_r(x_i, x_j) - \frac{1}{2\gamma} \sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i + \beta_i)(\alpha_j + \beta_j) K_r(x_i, x_j) + \\ & \frac{C}{\gamma} \sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i^* + \beta_i^*) K_r(x_i, x_j) - \frac{1}{2\gamma} \sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i^* + \beta_i^*)(\alpha_j^* + \beta_j^*) K_r(x_i, x_j) \end{aligned}$$

subject to the constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i^* &= \sum_{i=1}^{\ell} \alpha_i, \\ \sum_{i \in T_r} (\alpha_i + \beta_i) &= |T_r|C, \quad r = 1, \dots, t, \\ \sum_{i \in T_r} (\alpha_i^* + \beta_i^*) &= |T_r|C, \quad r = 1, \dots, t, \\ \alpha_i \geq 0, \quad \alpha_i^* \geq 0, \quad \beta_i \geq 0, \quad \beta_i^* \geq 0, \quad i &= 1, \dots, \ell. \end{aligned}$$

When either (1) there is no structure ($t = \ell$) or (2) there are no correlations ($K_r(x_i, x_j)$ has the form (2.45)) and $\gamma \rightarrow 0$ the solutions defined by SVM+ or SVM $_{\gamma}$ + regression coincide with the conventional SVM solution for regression.

2.6 THE THIRD GENERATION

In the mid-1990s the third generation of statistical learning theory (SLT) researchers appeared. They were well-educated, strongly motivated, and hard working PhD students from Europe. Many European universities allow their PhD students to work on their theses anywhere in the world, and several such students joined our department in order to work on their thesis. First came Bernhard Schölkopf, Volker Blanz, and Alex Smola from Germany, then Jason Weston from England, followed by Olivier Chapelle, Olivier Bousquet, and Andre Elisseeff from France, Pascal Vincent from Canada, and Corina Cortes (PhD student from Rochester university). At that time support vector technology had just started to develop. Later many talented young people followed this direction but these were the first from the third generation of researchers.

I would like to add to this group two young AT&T researchers of that time: Yoav Freund and Robert Schapire, who did not directly follow the line of statistical learning theory and developed Boosting technology that is close to the one discussed here [135, 136].

The third generation transformed both the area of machine learning research and the style of research. During a short period of time (less than ten years) they created a new direction in statistical learning theory: SVM and kernel methods. The format of this Afterword does not allow me to go into details of their work (there are hundreds of first-class articles devoted to this subject and it is very difficult to choose from them). I will just quote some of their textbooks [152–158], collective monographs and workshop materials [159–164]. Also I would like to mention the tutorial by Burges [165] which demonstrated the unity of theoretical and algorithmical parts of VC theory in a simple and convincing way.

The important achievement of the third generation was creating a large international SVM (kernel) community. They did it by accomplishing three things:

- (1) Constructing and supporting a special Website called Kernel Machine (www.kernel-machines.org).
- (2) Organizing eight machine learning workshops and five Summer Schools, where advanced topics relevant to empirical inference research were taught. These topics included:
 - Statistical learning theory,
 - Theory of empirical processes,
 - Functional analysis,
 - Theory of approximation,
 - Optimization theory, and
 - Machine learning algorithms.
 In fact they created the curriculum for a new discipline: *empirical inference science*.
- (3) Developing high-quality professional software for empirical inference problems that can be downloaded and used by anyone in the world.⁵

This generation took advantage of computer technology to change forever the style and atmosphere of data mining research: from the very hierarchical group structure of the 1970–1980s lead by old statistical gurus (with their *know-how* and dominating opinion) to an open new society (with widely available information, free technical tools, and open professional discussions).

Many of the third generation researchers of SLT became university professors. This Afterword is dedicated to their students.

⁵The three most popular software are:

- (1) *SVM-Light* developed by Thorsten Joachims (Germany) <http://svmlight.joachims.org/>,
- (2) *Lib-SVM* developed by Chin-Chang Chang and Chih-Jen Lin (Taiwan) <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>, and
- (3) *SVM-Torch* developed by Ronan Collobert (Switzerland) <http://www.torch.ch/>

2.7 RELATION TO THE PHILOSOPHY OF SCIENCE

By the end of the 1990s it became clear that there were strong ties between machine learning research and research conducted in the classical philosophy of induction. The problem of generalization (induction) always was one of the central problems in philosophy. Pattern recognition can be considered as the simplest problem of generalization (its drosophila fly: any idea of generalization has its reflection in this model). It forms a very good object for analysis and verification of a general inductive principle. Such analysis includes not only speculations but also experiments on computers.

Two main principles of induction were introduced in classical philosophy: the principle of simplicity (parsimony) formulated by the 14th century English monk Occam (Ocham), and the principle of falsifiability, formulated by the Austrian philosopher of the 20th century Karl Popper. Both of them have a direct reflection in statistical learning theory.

2.7.1 OCCAM'S RAZOR PRINCIPLE

The Occam's Razor (or parsimony) principle was formulated as follows:

Entities are not to be multiplied beyond necessity.

Such a formulation leaves two open questions:

- (1) What are the *entities*?
- (2) What does *beyond necessity* mean?

According to *The Concise Oxford Dictionary of Current English* [172] the word *entity* means

A thing's existence, as opposite to its qualities or relations; thing that has real existence.

So the number of entities is commonly understood to be the number of different parameters related to different physical (that which can be measured) features. The predictive rule is a function defined by these features.

The expression *not to be multiplied beyond necessity* has the following meaning: *not more than one needs to explain the observed facts.*

In accordance with such an interpretation the Occam's Razor principle can be reformulated as follows:

*Find the function from the set with the smallest number of free parameters that explains the observed facts.*⁶

⁶There exist wide interpretation of Occam's Razor principle as a request to minimize some functional (without specifying which). Such interpretation is too general to be useful since it depends on the definition of the functional. The original Occam formulation (assuming that entities are free parameters) is unambiguous and in many cases is a useful instrument of inference.

2.7.2 PRINCIPLES OF FALSIFIABILITY

To introduce the principles of falsifiability we need some definitions.

Suppose we are given a set of indicator functions $f(x, \alpha), \alpha \in \Lambda$. We say that the set of vectors

$$x_1, \dots, x_\ell, x_i \in X \quad (2.66)$$

cannot falsify the set of indicator functions $f(x, \alpha), \alpha \in \Lambda$ if all 2^ℓ possible separation of vectors (2.66) into two categories can be accomplished using functions from this set.

This means that on the data (2.66) one can obtain any classification (using functions from the admissible set). In other words, from these vectors one can obtain any possible law (given appropriate $y_i, i = 1, \dots, \ell$): the vectors themselves do not forbid (do not falsify) any possible law.

We say that the set of vectors (2.66) *falsifies* the set $f(x, \alpha), \alpha \in \Lambda$ if there exists such separation of the set (2.66) into two categories that cannot be obtained using an indicator function from the set $f(x, \alpha), \alpha \in \Lambda$.

Using the concept of falsifiability of a given set of functions by the given set of vectors, two different combinatorial definitions of the dimension of a given set of indicator functions were suggested: the VC dimension and the Popper dimension. These definitions lead to different concepts of falsifiability.

THE DEFINITION OF THE VC DIMENSION AND VC FALSIFIABILITY

The VC dimension is defined as follows (in *EDBED* it is called capacity. See Chapter 6, Sections 8 and A2:)

A set of functions $f(x, \alpha), \alpha \in \Lambda$ has VC dimension h if:

- (1) **there exist** h vectors that cannot falsify this set and
- (2) **any** $h + 1$ vectors falsify it.

The set of functions $f(x, \alpha), \alpha \in \Lambda$ is *VC falsifiable* if its VC dimension is finite and *VC nonfalsifiable* if its VC dimension is infinite.

The VC dimension of the set of hyperplanes in R^n is $n + 1$ (the number of free parameters of a hyperplane in R^n) since there exist $n + 1$ vectors that cannot falsify this set but any $n + 2$ vectors falsify it.

THE DEFINITION OF THE POPPER DIMENSION AND POPPER FALSIFIABILITY

The Popper dimension is defined as follows [137] (Section 38:)

A set of functions $f(x, \alpha), \alpha \in \Lambda$ has the Popper dimension h if:

- (1) **any** h vectors cannot falsify it and
- (2) **there exist** $h + 1$ vectors that can falsify this set.

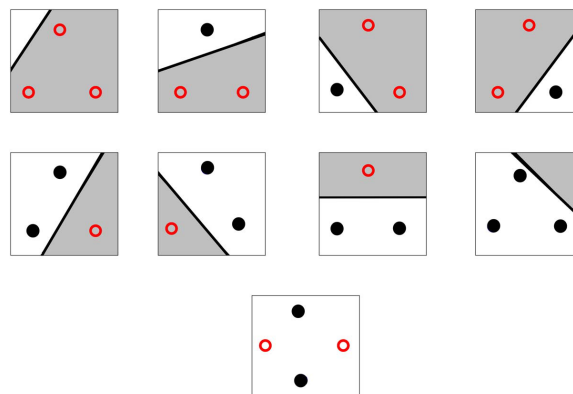


Figure 2.2: The VC dimension of the set of oriented lines in the plane is three since there exist three vectors that cannot falsify this set and any four vectors falsify it.

Popper called value h the degree of falsifiability or the dimension.

The set of functions $f(x, \alpha)$, $\alpha \in \Lambda$ is *Popper falsifiable* if its Popper dimension is finite and *Popper nonfalsifiable* if its Popper dimension is infinite.

Popper's dimension of the set of hyperplanes in R^n is at most two (independent of the dimensionality of the space n) since only two vectors that belong to the one-dimensional linear manifold can not falsify the set of hyperplanes in R^n and three vectors from this manifold falsify this set.

2.7.3 POPPER'S MISTAKES

In contrast to the VC dimension, the Popper concept of dimensionality does not lead to useful theoretical results for the pattern recognition model of generalization. The requirements of nonfalsifiability for any h vectors include, for example, the nonfalsifiability of vectors belonging to the line (one-dimensional manifold). Therefore, Popper's dimension will be defined by combinatorial properties restricted at most by the one-dimensional situation.

Discussing the concept of simplicity, Popper made several incorrect mathematical claims. This is the most crucial:

In an algebraic representation, the dimension of a set of curves depends upon the number of parameters whose value we can freely choose. We can therefore say that the number of freely determinable parameters of a set of curves by which a theory is represented is characteristic of the degree of falsifiability. [137, Section 43]

This is wrong for the Popper dimension. The claim is correct only in a restricted situation for the VC dimension, namely when the set of functions in R^n , $n > 2$ linearly depends on the parameters.

In other (more interesting) situations as in Denker's example with a set of $\theta(\{\sin ax\})$ functions (Section 2.2.1 and Section 2.7.5 below) and in the example of a separating hyperplane with the margin given in *EDBED* (Chapter 10, Section 5) that led to SVM technology, the considered set of functions depends nonlinearly upon the free parameters.

Popper did not distinguish the type of dependency on the parameters. Therefore he claimed that the set $\{\sin ax\}$ (with only one free parameter a) is a simple set of functions [137, Section 44]. However, the VC dimension of this set is infinite⁷ and therefore generalization using this set of functions is impossible.

It is surprising that the mathematical correctness of Popper's claims has never been discussed in the literature.⁸

2.7.4 PRINCIPLE OF VC FALSIFIABILITY

In terms of the philosophy of science, the structural risk minimization principle for the structure organized by the nested set with increasing VC dimension can be reformulated as follows:

Explain the facts using the function from the set that is easiest to falsify.

The mathematical consistency of SRM therefore can have the following philosophical interpretation:

Since one was able to find the function that separates the training data well, in the set of functions that is easy to falsify, these data are very special and the function which one chooses reflects the intrinsic properties of these data.⁹

It is possible, however, to organize the structure of nested elements on which capacity is defined by a more advanced measure than VC dimension (say, the Growth

⁷Since for any ℓ the set of values $x_1 = 2^{-1}, \dots, x_\ell = 2^{-\ell}$ cannot falsify $\{\theta(\sin ax)\}$. The desired classifications $y_1, \dots, y_\ell, y_i \in \{1, -1\}$ of this set provide the function $y = \theta(\sin a^*x)$ where the coefficient a^* is

$$a^* = \left(\pi \sum_{i=1}^{\ell} \frac{(1 - y_i)}{2} 2^i + 1 \right).$$

⁸Karl Popper's books were forbidden in the Soviet Union because of his criticism of communism. Therefore, I had no chance to learn about his philosophy until Gorbachev's time. In 1987 I attended a lecture on Popper's philosophy of science and learned about the falsifiability concept. After this lecture I became convinced that Popper described the VC dimension. (It was hard to imagine such a mistake.) Therefore in my 1995 and 1998 books I wrongly referred to Popper falsifiability as VC falsifiability. Only in the Spring of 2005 in the process of writing a philosophical article (see Corfield, Schölkopf, and Vapnik: "Popper, falsification and the VC dimension." Technical Report # 145, Max Planck Institute for Biological Cybernetics, Tübingen, 2005) did we check Popper's statements and realize my mistake.

⁹The *Minimum Message Length (MML)–Minimum Description Length (MDL)* principle [127, 128] that takes Kolmogorov's *algorithmic complexity* [129] into account can have the same interpretation. It is remarkable that even though the concepts of VC dimension and algorithmic complexity are very different, the MML-MDL principle leads to the same generalization bound for the pattern recognition problem that is given in *EDBED*. (See [139], Chapter 4, Section 4.6.)

function, or even better the VC entropy). This can lead to more advanced inference techniques (see Section 2.8 of this chapter).

Therefore the falsifiability principle is closely related to the VC dimension concept and can be improved by more refined capacity concepts.

2.7.5 PRINCIPLE OF PARSIMONY AND VC FALSIFIABILITY

The principle of simplicity was introduced as a principle of parsimony or a principle of economy of thought.

The definition of simplicity, however, is crucial since it can be very different. Here is an example. Which set of functions is simpler:

- (1) One that has the parametric form

$$f(x, \alpha), \alpha \in \Lambda, \text{ or}$$

- (2) One that has the parametric form

$$f(x, \alpha), \alpha \in \Lambda$$

and satisfies the constraint

$$\Omega(f) \leq C,$$

where $\Omega(f) \geq 0$ is some functional?

From a computational point of view, finding the desired function in situation 1 can be much simpler than in situation 2 (especially if the $\Omega(f) \leq C$ is a nonconvex set).

From an information theory point of view, however, to find the solution in situation 2 is simpler, since one is looking for the solution in a more restricted set of functions.

Therefore the inductive principle based on the (intuitive) idea of simplicity can lead to a contradiction. That is why Popper used the “degree of falsifiability” concept (Popper dimension) to characterize the simplicity:

The epistemological question which arise in connection with the concept of simplicity can all be answered if we equate this concept with degree of falsifiability. ([137], Section 43)

In the Occam’s Razor principle, the number of “entities” defines the simplicity. Popper incorrectly claimed the equality of Popper dimension to be the number of free parameters (entities), and considered the falsifiability principle to be a justification of the parsimony (Occam’s Razor) principle.

The principle of VC falsifiability does not coincide with the Occam’s Razor principle of induction, and this principle (but not Occam’s Razor) guarantee the generalization. VC dimension describes diversity of the set of functions. It does not refer either to the number of free parameters nor to our intuition of simplicity. Recall once again that Popper (and many other philosophers) had the intuition that $\{\theta(\sin ax)\}$ is the simple set of functions,¹⁰ while the VC dimension of this set is infinite.

¹⁰In the beginning of Section 44 [137] Popper wrote: “According to common opinion the sine-function is a simple one . . .”

The principle of VC falsifiability forms the necessary and sufficient conditions of consistency for the pattern recognition problem while there are pattern recognition algorithms that contradict the parsimony principle.¹¹

2.8 INDUCTIVE INFERENCE BASED ON CONTRADICTIONS

In my 1998 book, I discussed an idea of inference through contradictions [140, p.707]. In this Afterword, I introduce this idea as an algorithm for SVM. Sections 2.8.1 and 3.1.5 give the details of the algorithm. This section presents a simplified description of the general concept (see remark in Section 3.1.3 for details) of inductive inference through contradictions.

Suppose we are given a set of admissible indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$ and the training data. The vectors x from the training data split our admissible set of functions into a finite number of equivalence classes F_1, \dots, F_N . The equivalence class contains functions that have the same values on the training vectors x (separate them in the same way).

Suppose we would like to make a structure on the set of equivalence classes to perform SRM principle. That is, we would like to collect some equivalence classes in the first element of the structure, then add to them some other equivalence classes, constructing the second element, and so on. To do this we need to characterize every equivalence class by some value that describes our preference for it. Using such a measure, one can create the desired structure on the equivalence classes. When we constructed SVMs, we characterized the equivalence class by the size of the largest margin defined by the hyperplane belonging to this class.

Now let us consider a different characteristic. Suppose along with the training data we possess a set of vectors called *the Universum* or *the Virtual Universum*

$$x_1^*, \dots, x_k^*, x^* \in X. \quad (2.67)$$

The Universum plays the role of prior information in Bayesian inference. It describes our knowledge of the problem we are solving. However, there are important differences between the prior information in Bayesian inference and the prior information given by the Universum. In Bayesian inference, prior information is information about the relationship of the functions in the set of admissible functions to the desired one. With the Universum, prior information is information related to possible training and test vectors. For example, in the digit recognition problem it can be some vectors whose

¹¹The example of a machine learning algorithm that contradicts the parsimony principle is Boosting. This algorithm constructs so-called weak features (entities) which it linearly combines in a decision rule. Often this algorithm constructs some set of weak features and the corresponding decision rule that separates the training data with no mistakes but continues to add new weak features (new entities) to construct a better rule. With an increasing number of (unnecessary, i.e., those that have no effect on separating the training data) weak features, the algorithm improves its performance on the test data. One can show that with an increasing number of entities this algorithm increases the margin (as the SVM). The idea of this algorithm is to increase the number of entities (number of free parameters) in order to decrease the VC dimension [136].

images resemble a particular digit (say some artificial characters). It defines the style of the digit recognition task, and geometrically belongs to the same part of input space to which the training data belong.

We use the Universum to characterize the equivalence class. We say that a vector x^* is contradictory for the equivalence class F_s if there exists a function $f_1(x^*) \in F_s$ such that

$$f_1(x^*) > 0$$

and there also exists a function $f_2(x^*) \in F_s$ such that

$$f_2(x^*) < 0.$$

We will characterize our preference for an equivalence class by the number of contradictions that occur on the Universum: the more contradictions, the more preferable the equivalence class.¹² We construct structure on equivalence classes using these numbers.

When using the Universum to solve a classification problem based on SRM principle, we choose the function (say one that has the maximal margin) from the equivalence class that makes no (or a small number of) training mistakes and has the maximal number of contradictions on the Universum. In other words, for inductive inference, when constructing the structure for SRM, we replace the *maximal margin* score with the *maximal contradiction on Universum* (MCU) score and select maximal margin function from the chosen equivalence class.

The main problem with MCU inference is, how does one create the appropriate Universum? Note that since one uses Universum only for evaluation of sizes of equivalence classes, its elements do not need to have the same distribution as the training vectors.

2.8.1 SVMs IN THE UNIVERSUM ENVIRONMENT

The inference through contradictions can be implemented using SVM techniques as follows. Let us map both the training data and the Universum into Hilbert space

$$(y_1, z_1), \dots, (y_\ell, z_\ell) \tag{2.68}$$

$$z_1^*, \dots, z_u^*. \tag{2.69}$$

QUADRATIC OPTIMIZATION FRAMEWORK

In the quadratic optimization framework for an SVM, to conduct inference through contradictions means finding the hyperplane

$$(w^0, z) + b_0 = 0 \tag{2.70}$$

¹²A more interesting characteristic of an equivalence class would be the value of the VC entropy of the set of functions belonging to this equivalence class calculated on the Universum. This, however, leads to difficult computational problems. The number of contradictions can be seen as a characteristic of the entropy.

that minimizes the functional

$$R = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \theta(\xi_i) + C_2 \sum_{j=1}^u \theta(\xi_j^*), \quad C_1, C_2 > 0 \quad (2.71)$$

subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell \quad (2.72)$$

(related to the training data) and the constraints

$$|(w, z_j^*) + b| \leq a + \xi_j^*, \quad \xi_j^* \geq 0, \quad j = 1, \dots, u \quad (2.73)$$

(related to the Universum) where $a \geq 0$.

As before, for computational reasons we approximate the target function (2.71) by the function¹³

$$R = \frac{1}{2}(w, w) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{s=1}^u \xi_s^*, \quad C_1, C_2 > 0. \quad (2.74)$$

Using the Lagrange multipliers technique we determine that our hyperplane in feature space has the form

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i(z_i, z) + \sum_{s=1}^u (\mu_s^0 - \nu_s^0)(z_s^*, z) + b = 0, \quad (2.75)$$

where the coefficients $\alpha_i^0 \geq 0$, $\mu_s^0 \geq 0$, and $\nu_s^0 \geq 0$ are the solution of the following optimization problem: Maximize the functional

$$\begin{aligned} W(\alpha, \mu, \nu) = & \sum_{i=1}^{\ell} \alpha_i - a \sum_{s=1}^u (\mu_s + \nu_s) - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (z_i, z_j) \\ & - \sum_{i=1}^{\ell} \sum_{s=1}^u \alpha_i y_i (\mu_s - \nu_s)(z_i, z_s^*) - \frac{1}{2} \sum_{s,t=1}^u (\mu_s - \nu_s)(\mu_t - \nu_t)(z_s^*, z_t^*) \end{aligned} \quad (2.76)$$

subject to the constraint

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{s=1}^u (\mu_s - \nu_s) = 0 \quad (2.77)$$

and the constraints

$$0 \leq \alpha_i \leq C_1 \quad (2.78)$$

$$0 \leq \mu_s, \nu_s \leq C_2. \quad (2.79)$$

¹³One also can use a least squares technique by choosing ξ_i^2 and $(\xi_i^*)^2$ instead of ξ_i and ξ_i^* in objective function (2.74).

Taking into account Mercer's theorem, one can rewrite our separating function in input space as

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + \sum_{s=1}^u (\mu_s^0 - \nu_s^0) K(x_s^*, x) + b_0 = 0, \quad (2.80)$$

where the coefficients $\alpha_i^0 \geq 0$, $\mu_s^0 \geq 0$, and $\nu_s^0 \geq 0$ are the solution of the following optimization problem: Maximize the functional

$$\begin{aligned} W(\alpha, \mu, \nu) = & \sum_{i=1}^{\ell} \alpha_i - a \sum_{s=1}^u (\mu_s + \nu_s) - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ & - \sum_{i=1}^{\ell} \sum_{s=1}^u \alpha_i y_i (\mu_s - \nu_s) K(x_i, x_s^*) - \frac{1}{2} \sum_{s,t=1}^u (\mu_s - \nu_s)(\mu_t - \nu_t) K(x_s^*, x_t^*) \end{aligned} \quad (2.81)$$

subject to the constraints (2.77), (2.78), (2.79).

LINEAR OPTIMIZATION FRAMEWORK

To conduct inference based only on contradictions arguments (taking some function from the chosen equivalence class, not necessarily one with the largest margin) one has to find the coefficients α^0 , μ^0 , ν^0 in (2.80) using the following linear programming technique: Minimize the functional

$$W(\alpha, \mu, \nu) = \gamma \sum_{i=1}^{\ell} \alpha_i + \gamma \sum_{s=1}^u (\mu_s + \nu_s) + C_1 \sum_{i=1}^{\ell} \xi_i + C_2 \sum_{t=1}^u \xi_t^*, \quad \gamma \geq 0 \quad (2.82)$$

subject to the constraints

$$y_i \left[\sum_{j=1}^{\ell} \alpha_j y_j K(x_i, x_j) + \sum_{s=1}^u (\mu_s - \nu_s) K(x_i, x_s^*) + b \right] \geq 1 - \xi_i, \quad i = 1, \dots, \ell \quad (2.83)$$

and the constraints

$$\sum_{j=1}^{\ell} \alpha_j y_j K(x_t^*, x_j) + \sum_{s=1}^u (\mu_s - \nu_s) K(x_t^*, x_s^*) + b \leq a + \xi_t^*, \quad t = 1, \dots, k, \quad (2.84)$$

$$\sum_{j=1}^{\ell} \alpha_j y_j K(x_t^*, x_j) + \sum_{s=1}^u (\mu_s - \nu_s) K(x_t^*, x_s^*) + b \geq -a - \xi_t^*, \quad t = 1, \dots, u, \quad (2.85)$$

where $a \geq 0$. In the functional (2.82) the parameter $\gamma \geq 0$ controls the sparsity of the solution.

2.8.2 THE FIRST EXPERIMENTS AND GENERAL SPECULATIONS

In the summer of 2005, Ronan Collobert and Jason Weston conducted the first experiments on training SVM with Universum using the algorithm described in Section 2.8.1. They discriminated digit 8 from digit 5 from the MNIST database, using a conventional SVM and an SVM trained in three different Universum environments.

The following table shows for different sizes of training sets the performance of a conventional SVM and the SVMs trained using Universums U_1, U_2, U_3 (each containing 5000 examples). In all cases the parameter $a = .01$, the parameters C_1, C_2 , and the parameter of the Gaussian kernel were tuned using the tenfold cross-validation technique.

The Universums were constructed as follows:

U_1 : Selects random digits from the other classes (0,1,2,3,4,6,7,9).

U_2 : Creates an artificial image by first selecting a random 5 and a random 8, and then for each pixel of the artificial image choosing with probability 1/2 the corresponding pixel from the image 5 or from the image 8.

U_3 : Creates an artificial image by first selecting a random 5 and a random 8, and then constructing the mean of these two digits.

No. of train. examples	250	500	1000	2000	3000
Test Err. SVM (%)	2.83	1.92	1.37	0.99	0.83
Test Err. SVM+ U_1 (%)	2.43	1.58	1.11	0.75	0.63
Test Err. SVM+ U_2 (%)	1.51	1.12	0.89	0.68	0.60
Test Err. SVM+ U_3 (%)	1.33	0.89	0.72	0.60	0.58

The table shows that:

- (a) The Universum can significantly improve the performance of SVMs.
- (b) The role of knowledge provided by the Universum becomes more important with decreasing training size. However, even when the training size is large, the Universum still has a significant effect on performance.

We expect that advancing the understanding of the concept of a good Universum for the problem of interest will further boost the performance. This fact opens a new dimension in machine learning technology: How does one create a Virtual Universum for the problem of interest?

In trying to find an interpretation of the role of the Universum in machine learning, it is natural to compare it to the role of culture in the learning of humans, where knowledge about real life is concentrated not only in examples of reality but also in images that reflect this reality. To classify well, one uses inspiration from both sources.



<http://www.springer.com/978-0-387-30865-4>

Estimation of Dependences Based on Empirical Data

Vapnik, V.

2006, XVIII, 505 p., Hardcover

ISBN: 978-0-387-30865-4