Chapter 2
Support Vector Learning and Optimization

East is east and west is west and never the twain shall meet.
The Ballad of East and West by Rudyard Kipling

2.1 Goals of This Chapter

The kernel-based methodology of SVMs [Vapnik and Chervonenkis, 1974], [Vapnik, 1995a] has been established as a top ranking approach for supervised learning within both the theoretical and red practical research environments. This very performing technique suffers nevertheless from the curse of an opaque engine [Huysmans et al, 2006], which is undesirable for both theoreticians, who are keen to control the modeling, and the practitioners, who are more than often suspicious of using the prediction results as a reliable assistant in decision making.

A concise view on a SVM is given in [Cristianini and Shawe-Taylor, 2000]:

> A system for efficiently training linear learning machines in kernel-induced feature spaces, while respecting the insights of generalization theory and exploiting optimization theory.

The right placement of data samples to be classified triggers corresponding separating surfaces within SVM training. The technique basically considers only the general case of binary classification and treats reductions of multi-class tasks to the former. We will also start from the general case of two-class problems and end with the solution to several classes.

If the first aim of this chapter is to outline the essence of SVMs, the second one targets the presentation of what is often presumed to be evident and treated very rapidly in other works. We therefore additionally detail the theoretical aspects and mechanism of the classical approach to solving the constrained optimization problem within SVMs.

Starting from the central principle underlying the paradigm (Sect. 2.2), the discussion of this chapter pursues SVMs from the existence of a linear decision function (Sect. 2.3) to the creation of a nonlinear surface (Sect. 2.4) and ends with the treatment for multi-class problems (Sect. 2.5).

C. Stoean and R. Stoean, *Support Vector Machines and Evolutionary Algorithms for Classification*, Intelligent Systems Reference Library 69,
DOI: 10.1007/978-3-319-06941-8_2,

2.2 Structural Risk Minimization

SVMs act upon a fundamental theoretical assumption, called the principle of structural risk minimization (SRM) [Vapnik and Chervonenkis, 1968].

Intuitively speaking, the SRM principle asserts that, for a given classification task, with a certain amount of training data, generalization performance is solely achieved if the accuracy on the particular training set and the capacity of the machine to pursue learning on any other training set without error have a good balance. This request can be illustrated by the example found in [Burges, 1998]:

> A machine with too much capacity is like a botanist with photographic memory who, when presented with a new tree, concludes that it is not a tree because it has a different number of leaves from anything she has seen before; a machine with too little capacity is like the botanist's lazy brother, who declares that if it's green, then it's a tree. Neither can generalize well.

We have given a definition of classification in the introductory chapter and we first consider the case of a binary task. For convenience of mathematical interpretation, the two classes are labeled as -1 and 1; henceforth, $y_i \in \{-1,1\}$.

Let us suppose the set of functions $\{f_t\}$, of generic parameters t:

$$f_t : \mathbb{R}^n \to \{-1,1\}. \tag{2.1}$$

The given set of m training samples can be labeled in 2^m possible ways. If for each labeling, a member of the set $\{f_t\}$ can be found to correctly assign those labels, then it is said that the collection of samples is shattered by that set of functions [Cherkassky and Mulier, 2007].

Definition 2.1. [Burges, 1998] The Vapnik-Chervonenkis (VC) - dimension h for a set of functions $\{f_t\}$ is defined as the maximum number of training samples that can be shattered by it.

Proposition 2.1. *(Structural Risk Minimization principle) [Vapnik, 1982]*

For the considered classification problem, for any generic parameters t and for $m > h$, with a probability of at least $1-\eta$, the following inequality holds:

$$R(t) \leq R_{emp}(t) + \phi\left(\frac{h}{m}, \frac{\log(\eta)}{m}\right),$$

where $R(t)$ is the test error, $R_{emp}(t)$ is the training error and ϕ is called the confidence term and is defined as:

$$\phi\left(\frac{h}{m}, \frac{\log(\eta)}{m}\right) = \sqrt{\frac{h\left(\log\frac{2m}{h}+1\right) - \log\frac{\eta}{4}}{m}}.$$

The SRM principle affirms that, for a high generalization ability, both the training error and the confidence term must be kept minimal; the latter is minimized by reducing the VC-dimension.

2.3 Support Vector Machines with Linear Learning

When confronted with a new classification task, the first reasonable choice is to try and separate the data in a linear fashion.

2.3.1 Linearly Separable Data

If training data are presumed to be linearly separable, then there exists a linear hyperplane H:

$$H : w \cdot x - b = 0, \tag{2.2}$$

which separates the samples according to their classes [Haykin, 1999]. w is called the weight vector and b is referred to as the bias.

Recall that the two classes are labeled as -1 and 1. The data samples of class 1 thus lie on the positive side of the hyperplane and their negative counterparts on the opposite side.

Proposition 2.2. *[Haykin, 1999]*

Two subsets of n-dimensional samples are linearly separable iff there exist $w \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$ *such that for every sample* $i = 1, 2, ..., m$*:*

$$\begin{cases} w \cdot x_i - b > 0, y_i = 1 \\ w \cdot x_i - b \leq 0, y_i = -1 \end{cases} \tag{2.3}$$

An insightful picture of this geometric separation is given in Fig. 2.1.

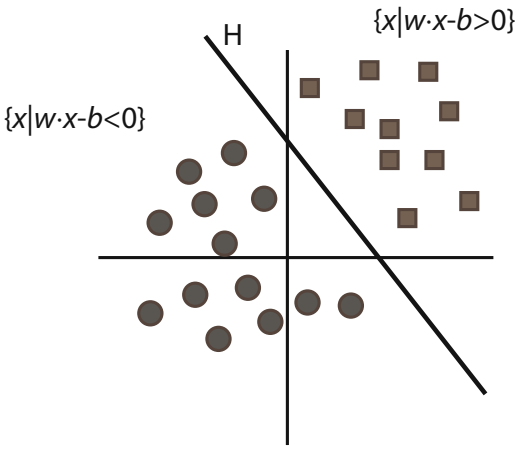

Fig. 2.1 The positive and negative samples, denoted by squares and circles, respectively. The decision hyperplane between the two corresponding separable subsets is H.

It is further resorted to a stronger statement for linear separability, where the positive and negative samples lie behind a corresponding supporting hyperplane.

Proposition 2.3. *[Bosch and Smith, 1998] Two subsets of n-dimensional samples are linearly separable iff there exist* $w \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$ *such that, for every sample* $i = 1, 2, ..., m$*:*

$$\begin{cases} w \cdot x_i - b \geq 1, y_i = 1 \\ w \cdot x_i - b \leq -1, y_i = -1 \end{cases} \tag{2.4}$$

An example for the stronger separation concept is given in Fig. 2.2.

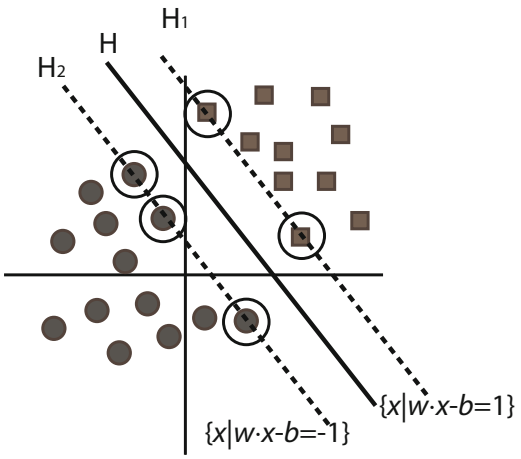

Fig. 2.2 The decision and supporting hyperplanes for the linearly separable subsets. The separating hyperplane H is the one that lies in the middle of the two parallel supporting hyperplanes H_1, H_2 for the two classes. The support vectors are circled.

Proof. (we provide a detailed version – as in [Stoean, 2008] – for a gentler flow of the connections between the different conceptual statements)

Suppose there exist w and b such that the two inequalities hold.

The subsets given by $y_i = 1$ and $y_i = -1$, respectively, are linearly separable since all positive samples lie on one side of the hyperplane given by

$$w \cdot x - b = 0,$$

from:

$$w \cdot x_i - b \geq 1 > 0 \text{ for } y_i = 1,$$

and simultaneously:

$$w \cdot x_i - b \leq -1 < 0 \text{ for } y_i = -1,$$

so all negative samples lie on the other side of this hyperplane.

Now, conversely, suppose the two subsets are linearly separable. Then, there exist $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that, for $i = 1, 2, ..., m$:

$$\begin{cases} w \cdot x_i - b > 0, y_i = 1 \\ w \cdot x_i - b \leq 0, y_i = -1 \end{cases}$$

Since:

$$\min \{w \cdot x_i | y_i = 1\} > \max \{w \cdot x_i | y_i = -1\},$$

let us set:

$$p = \min \{w \cdot x_i | y_i = 1\} - \max \{w \cdot x_i | y_i = -1\}$$

and make:

$$w' = \frac{2}{p} w$$

and

$$b' = \tfrac{1}{p} (\min \{w \cdot x_i | y_i = 1\} + \max \{w \cdot x_i | y_i = -1\})$$

Then:

$$\begin{aligned} \min \left\{ w' \cdot x_i | y_i = 1 \right\} = & \\ & = \frac{2}{p} \min \{w \cdot x_i | y_i = 1\} \\ & = \frac{1}{p} (\min \{w \cdot x_i | y_i = 1\} + \max \{w \cdot x_i | y_i = -1\} + \\ & \min \{w \cdot x_i | y_i = 1\} - \max \{w \cdot x_i | y_i = -1\}) \\ & = \frac{1}{p} (\min \{w \cdot x_i | y_i = 1\} + \max \{w \cdot x_i | y_i = -1\} + p) \\ & = b' + 1 \end{aligned}$$

and

$$\begin{aligned} \max \left\{ w' \cdot x_i | y_i = -1 \right\} = & \\ & = \frac{2}{p} \max \{w \cdot x_i | y_i = -1\} \\ & = \frac{1}{p} (\min \{w \cdot x_i | y_i = 1\} + \max \{w \cdot x_i | y_i = -1\} - p) \\ & = b' - 1 \end{aligned}$$

Consequently, there exist $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that:

$$w \cdot x_i \geq b + 1 \Rightarrow w \cdot x_i - b \geq 1 \text{ when } y_i = 1$$

$$\text{and } w \cdot x_i \leq b - 1 \Rightarrow w \cdot x_i - b \leq -1 \text{ when } y_i = -1$$

□

Definition 2.2. The support vectors are the training samples for which either the first or the second line of (2.4) holds with the equality sign.

In other words, the support vectors are the data samples that lie closest to the decision surface. Their removal would change the found solution. The supporting hyperplanes are those denoted by the two lines in (2.4), if equalities are stated instead.

Following the geometrical separation statement (2.4), SVMs hence have to determine the optimal values for the coefficients w and b of the decision hyperplane that linearly partitions the training data. In a more succinct formulation, from (2.4), the optimal w and b must then satisfy for every $i = 1, 2, ..., m$:

$$y_i(w \cdot x_i - b) - 1 \geq 0 \tag{2.5}$$

In addition, according to the SRM principle (Proposition 2.1), separation must be performed with a high generalization capacity. In order to also address this point, in the next lines, we will first calculate the margin of separation between classes.

The distance from one random sample z to the separating hyperplane is given by:

$$\frac{|w \cdot z - b|}{\|w\|}. \tag{2.6}$$

Let us subsequently compute the same distance from the samples z_i that lie closest to the separating hyperplane on either side of it (the support vectors, see Fig. 2.2). Since z_i are situated closest to the decision hyperplane, it results that either $z_i \in H_1$ or $z_i \in H_2$ (according to Def. 2.2) and thus $|w \cdot z_i - b| = 1$, for all i.

Hence:

$$\frac{|w \cdot z_i - b|}{\|w\|} = \frac{1}{\|w\|} \text{ for all } i = 1, 2, ..., m. \tag{2.7}$$

Then, the margin of separation becomes equal to [Vapnik, 2003]:

$$\frac{2}{\|w\|}. \tag{2.8}$$

Proposition 2.4. *[Vapnik, 1995b]*

Let r be the radius of the smallest ball

$$B_r(a) = \{x \in \mathbb{R}^n | \, \|x - a\| < r\}, a \in \mathbb{R}^n$$

containing the samples $x_1, ..., x_m$ and let

$$f_{w,b} = sgn(w \cdot x - b)$$

be the hyperplane decision functions.

Then the set $\left\{ f_{w,b} | \, \|w\| \leq A \right\}$ has a VC-dimension h (as from Definition 2.1) satisfying

$$h < r^2 A^2 + 1$$

In other words, it is stated that, since $\|w\|$ is inversely proportional to the margin of separation (from (2.8)), by requiring a large margin (i.e., a small A), a small VC-dimension is obtained. Conversely, by allowing separations with small margin, a much larger class of problems can be potentially separated (i.e., there exists a larger class of possible labeling modes for the training samples, from the definition of the VC-dimension).

The SRM principle requests that, in order to achieve high generalization of the classifier, training error and VC-dimension must be both kept small. Therefore, hyperplane decision functions must be constrained to maximize the margin, i.e.,

$$\text{minimize } \frac{\|w\|^2}{2}, \tag{2.9}$$

and separate the training data with as few exceptions as possible.

From (2.5) and (2.9), it follows that the resulting optimization problem is (2.10) [Haykin, 1999]:

$$\begin{cases} \text{find } w \text{ and } b \text{ as to minimize } \dfrac{\|w\|^2}{2} \\ \text{subject to } y_i(w \cdot x_i - b) \geq 1, \text{ for all } i = 1, 2, ..., m \end{cases} \tag{2.10}$$

The reached constrained optimization problem is called the primal problem (PP).

2.3.2 Solving the Primal Problem

The original solving of the PP (2.10) requires the a priori knowledge of several fundamental mathematical propositions described in the subsequent lines.

Definition 2.3. A function $f : C \to \mathbb{R}$ is said to be convex if
$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$, for all $x, y \in C$ and $\alpha \in [0,1]$.

Proposition 2.5. *For a function $f : (a,b) \to \mathbb{R}$, $(a,b) \subseteq \mathbb{R}$, that has a second derivative in (a,b), a necessary and sufficient condition for its convexity on that interval is that the second derivative $f''(x) \geq 0$, for all $x \in (a,b)$.*

Proposition 2.6. *If two functions are convex, the composition of the functions is convex.*

Proposition 2.7. *The objective function in PP* (2.10) *is convex [Haykin, 1999].*

Proof. (detailed as in [Stoean, 2008])
Let $h = f \circ g$, where $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ and $g : \mathbb{R}^n \to \mathbb{R}$, $g(w) = \|w\|$.

1. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 \Rightarrow f'(x) = 2x \Rightarrow f''(x) = 2 \geq 0 \Rightarrow f$ is convex.

2. $g : \mathbb{R}^n \to \mathbb{R}$, $g(w) = \|w\|$

We appeal to two well-known properties of a norm:

1. $\|\alpha v\| = |\alpha| \, \|v\|$
2. $\|v + w\| \leq \|v\| + \|w\|$

Let $v, w \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

$$g(\alpha v + (1-\alpha)w) = \|\alpha v + (1-\alpha)w\| \leq |\alpha| \, \|v\| + |1-\alpha| \, \|w\| =$$
$$\alpha \, \|v\| + (1-\alpha) \, \|w\| = \alpha g(v) + (1-\alpha) g(w)$$

$\Rightarrow g$ is convex.

Following Proposition 2.6 $\Rightarrow h$ is convex. □

Since constraints in PP (2.10) are linear in w, the following proposition arises.

Proposition 2.8. *The feasible region for a constrained optimization problem is convex if the constraints are linear.*

At this point, we have all the necessary information to outline the classical solving of the PP inside SVMs (2.10). The standard method of finding the optimal solution with respect to the defined constraints resorts to an extension of the Lagrange multipliers method. This is described in detail in what follows.

Since the objective function is convex and constraints are linear, the Karush-Kuhn-Tucker-Lagrange (KKTL) conditions can be stated for PP [Haykin, 1999] .

This is based on the argument that, since constraints are linear, the KKTL conditions are guaranteed to be necessary. Also, since PP is convex (convex objective function + convex feasible region), the KKTL conditions are at the same time sufficient for global optimality [Fletcher, 1987].

First, the Lagrangian function is constructed:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y_i (w \cdot x_i - b) - 1 \right], \tag{2.11}$$

where variables $\alpha_i \geq 0$ are the Lagrange multipliers.

The solution to the problem is determined by the KKTL conditions for every sample $i = 1, 2, ..., m$ [Burges, 1998]:

$$\begin{cases} \dfrac{\partial L(w,b,\alpha)}{\partial w} = 0 \\ \dfrac{\partial L(w,b,\alpha)}{\partial b} = 0 \\ \alpha_i \left[y_i (w \cdot x_i - b) - 1 \right] = 0 \end{cases}$$

Application of the KKTL conditions yields [Haykin, 1999]:

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{2.12}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.13}$$

$$\alpha_i \left[y_i (w \cdot x_i - b) - 1 \right] = 0, i = 1,2,...,m \tag{2.14}$$

We additionally refer to the separability statement and the conditions for positive Lagrange multipliers for every $i = 1,2,...,m$:

$$y_i (w \cdot x_i - b) - 1 \geq 0$$

$$\alpha_i \geq 0$$

We have to solve the particular PP in (2.10). Generally speaking, given the PP:

$$\begin{cases} \text{minimize } f(x) \\ \text{subject to } \begin{cases} g_1(x) \geq 0 \\ ... \\ g_m(x) \geq 0 \end{cases} \end{cases}, \tag{2.15}$$

the Lagrange multipliers are $\alpha = (\alpha_1^*, ..., \alpha_m^*)$, $\alpha_i^* \geq 0$, such that:

$$\inf_{g_1(x) \geq 0, ..., g_m(x) \geq 0} f(x) = \inf_{x \in \mathbb{R}^n} L(x, \alpha^*),$$

where L is the Lagrangian function:

$$L(x,\alpha) = f(x) + \sum_{j=1}^{m} \alpha_j g_j(x), x \in \mathbb{R}^n, \alpha \in \mathbb{R}^m$$

Then, one can resort to the dual function [Haykin, 1999]:

$$q(\alpha) = \inf_{x \in \mathbb{R}^n} L(x,\alpha)$$

This naturally leads to the dual problem (DP) :

$$\begin{cases} \text{maximize } q(\alpha) \\ \text{subject to } \alpha \geq 0 \end{cases} \tag{2.16}$$

The optimal primal value is $f^* = \inf_{g_1(x) \geq 0, ..., g_r(x) \geq 0} f(x) = \inf_{x \in \mathbb{R}^n} \sup_{\alpha \geq 0} L(x,\alpha)$.
The optimal dual value is $g^* = \sup_{\alpha \geq 0} q(\alpha) = \sup_{\alpha \geq 0} \inf_{x \in \mathbb{R}^n} L(x,\alpha)$.
There is always that $q^* \leq f^*$.

But, if there is convexity in the PP, then:

1. $q^* = f^*$
2. Optimal solutions of the DP are multipliers for the PP.

Further on, (2.11) is expanded and one obtains [Haykin, 1999]:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i y_i w \cdot x_i + b\sum_{i=1}^{m} \alpha_i y_i + \sum_{i=1}^{m} \alpha_i \tag{2.17}$$

The third term on the right-hand side of the expansion is zero from (2.13). Moreover, from (2.12), one obtains:

$$\frac{1}{2}\|w\|^2 = w \cdot w = \sum_{i=1}^{m} \alpha_i y_i w \cdot x_i = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

Therefore, (2.17) changes to:

$$L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

According to the duality concepts, by setting $Q(\alpha) = L(w,b,\alpha)$, one obtains the DP:

$$\begin{cases} \text{find } \{\alpha_i\}_{i=1,2,\ldots,m} \text{ as to maximize } Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ \text{subject to } \begin{cases} \sum_{i=1}^{m} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \end{cases} \end{cases} \tag{2.18}$$

The optimum Lagrange multipliers are next determined by setting the gradient of Q to zero and solving the resulting system.

Then, the optimum vector w can be computed from (2.12) [Haykin, 1999]:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

As b is concerned, it can be obtained from any of the equalities of (2.14), when $\alpha_i \neq 0$. Then:

$$y_i(w \cdot x_i - b) - 1 = 0 \Rightarrow$$

$$y_i(\sum_{j=1}^{m} \alpha_j y_j x_j \cdot x_i - b) = 1 \Rightarrow$$

$$\sum_{j=1}^{m} \alpha_j y_j x_j \cdot x_i - b = y_i \Rightarrow$$

$$b = \sum_{j=1}^{m} \alpha_j y_j x_j \cdot x_i - y_i$$

Note that we have equalled $1/y_i$ to y_i above, since y_i can be either 1 or -1.

Although the value for b can be thus directly derived from only one such equality when $\alpha_i \neq 0$, it is nevertheless safer to compute all the b values and take their mean as the final result.

In the reached solution to the constrained optimization problem, those points for which $\alpha_i > 0$ are the support vectors and they can also be obtained as the output of the SVM.

Finally, the class for a test sample x' is predicted based on the sign of the decision function with the found coefficients w and b applied to x' and the inequalities in (2.4):

$$class(x') = sgn(w \cdot x' - b)$$

2.3.3 Linearly Nonseparable Data

Since real-world data are not linearly separable, it is obvious that a linear separating hyperplane is not able to build a partition without any errors. However, a linear separation that minimizes training error can be tried as a solution to the classification problem [Haykin, 1999].

The separability statement can be relaxed by introducing slack variables $\xi_i \geq 0$ into its formulation [Cortes and Vapnik, 1995]. This can be achieved by observing the deviations of data samples from the corresponding supporting hyperplanes, which designate the ideal condition of data separability. These variables may then indicate different nuanced digressions (Fig. 2.3), but only a $\xi_i > 1$ signifies an error of classification.

Minimization of training error is achieved by adding the indicator of an error (slack variable) for every training data sample into the separability statement and, at the same time, by minimizing their sum.

For every sample $i = 1, 2, ..., m$, the constraints in (2.5) subsequently become:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \tag{2.19}$$

where $\xi_i \geq 0$.

Simultaneously with (2.19), the sum of misclassifications must be minimized:

$$\text{minimize } C \sum_{i=1}^{m} \xi_i. \tag{2.20}$$

$C > 0$ is a parameter of the methodology and is employed for the penalization of errors.

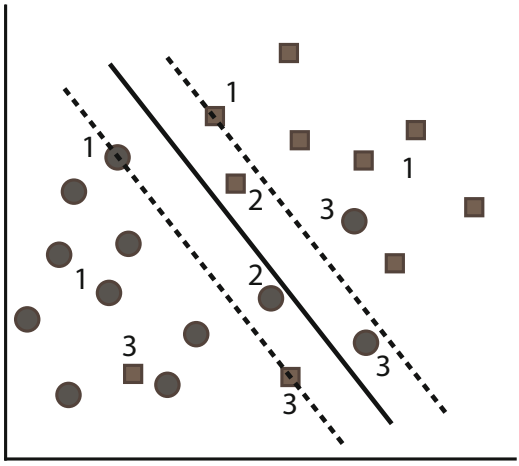

Fig. 2.3 Different data placements in relation to the separating and supporting hyperplanes. Corresponding indicators of errors are labeled by 1, 2 and 3: correct placement, $\xi_i = 0$ (label 1), margin position, $\xi_i < 1$ (label 2) and classification error, $\xi_i > 1$ (label 3).

Therefore, the optimization problem changes to (2.21):

$$\begin{cases} \text{find } w \text{ and } b \text{ as to minimize } \frac{\|w\|^2}{2} + C\sum_{i=1}^{m} \xi_i, C > 0 \\ \text{subject to } y_i(w \cdot x_i - b) \geq 1 - \xi_i, \xi_i \geq 0, \text{ for all } i = 1, 2, ..., m \end{cases} \tag{2.21}$$

This formulation still obeys the SRM principle as the VC-dimension is once more minimized and separation of training data with as few exceptions as possible is again achieved, both through (2.19) and (2.20).

From the formulation in (2.11), the Lagrangian function changes in the following way [Burges, 1998], where variables α_i and μ_i, $i = 1, 2, ..., m$, are the Lagrange multipliers:

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i - b) - 1 + \xi_i] - \sum_{i=1}^{m} \mu_i \xi_i,$$

The introduction of the μ_i multipliers is related to the inclusion of the ξ_i variables in the relaxed formulation of the PP.

Application of the KKTL conditions to this new constrained optimization problem leads to the following lines: [Burges, 1998]:

$$\frac{\partial L(w, b, \xi, \alpha, \mu)}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{2.22}$$

$$\frac{\partial L(w,b,\xi,\alpha,\mu)}{\partial b} = \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.23}$$

$$\frac{\partial L(w,b,\xi,\alpha,\mu)}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow \alpha_i + \mu_i = C \tag{2.24}$$

The KKTL conditions also require that, for every $i = 1,2,...,m$, the subsequent equalities hold:

$$\alpha_i \left[y_i (w \cdot x_i - b) - 1 + \xi_i \right] = 0 \tag{2.25}$$

$$\mu_i \xi_i = 0 \tag{2.26}$$

We additionally refer to the relaxed separability statement and the conditions for positive slack variables ξ_i and Lagrange multipliers α_i and μ_i for every $i = 1,2,...,m$:

$$y_i (w \cdot x_i - b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0$$

After term by term expansion, the Lagrangian function is then transformed to:

$$L(w,b,\xi,\alpha,\mu) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \mu_i \xi_i$$

From (2.26), the last term of the Lagrangian becomes zero and following (2.24) and expanding the third term, one obtains:

$$\begin{aligned} L(w,b,\xi,\alpha,\mu) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j + \sum_{i=1}^{m} (\alpha_i + \mu_i) \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j + \sum_{i=1}^{m} \alpha_i \xi_i + \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \end{aligned}$$

Consequently, the following corresponding DP is obtained:

$$\begin{cases} \text{find } \{\alpha_i\}_{i=1,2,\ldots,m} \text{ as to maximize } Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ \text{subject to } \begin{cases} \sum_{i=1}^{m} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases}, C > 0 \end{cases} \tag{2.27}$$

The second constraint is obtained from (2.24) and the condition that $\mu_i \geq 0$, for every sample $i = 1,2,\ldots,m$.

The optimum value for w is again computed as:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Coefficient b of the hyperplane can be determined as follows [Haykin, 1999]. If the values α_i obeying the condition $\alpha_i < C$ are considered, then from (2.24) it results that for those i $\mu_i \neq 0$. Subsequently, from (2.26) we derive that $\xi_i = 0$, for those certain i. Under these circumstances, from (2.25) and (2.22), one obtains the same formulation as in the separable case:

$$y_i(w \cdot x_i - b) - 1 = 0 \Rightarrow b = \sum_{j=1}^{m} \alpha_j y_j x_j \cdot x_i - y_i.$$

It is again better to take b as the mean value resulting from all such equalities.

Those points that have $0 < \alpha_i < C$ are the support vectors.

2.4 Support Vector Machines with Nonlinear Learning

If a linear hyperplane is not able to provide satisfactory results for the classification task, then is it possible that a nonlinear decision surface can do the separation? The answer is affirmative and is based on the following result.

Theorem 2.1. *[Cover, 1965] A complex pattern classification problem cast in a high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space.*

The above theorem states that an input space can be mapped into a new feature space where it is highly probable that data are linearly separable provided that:

1. The transformation is nonlinear.
2. The dimensionality of the feature space is high enough.

The initial space of training data samples can thus be nonlinearly mapped into a higher dimensional feature space, where a linear decision hyperplane can be subsequently built. The decision hyperplane achieves an accurate separation in the feature space which corresponds to a nonlinear decision function in the initial space (see Fig. 2.4).

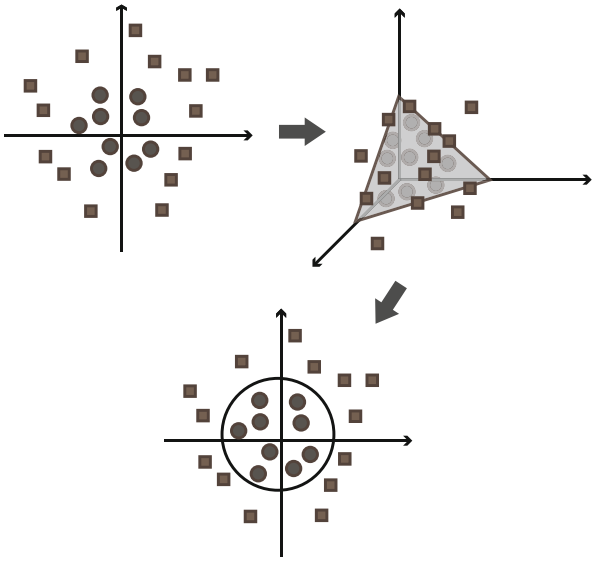

Fig. 2.4 The initial data space with squares and circles (up left) is nonlinearly mapped into the higher dimensional space, where the objects are linearly separable (up right). This corresponds to a nonlinear surface discriminating in the initial space (down).

The procedure therefore leads to the creation of a linear separating hyperplane that minimizes training error as before, but this time performs in the feature space. Accordingly, a nonlinear map $\Phi : \mathbb{R}^n \to H$ is considered and data samples from the initial space are mapped by Φ into H.

In the standard solving of the SVM optimization problem, vectors appear only as part of scalar products; the issue can be thus further simplified by substituting the dot product by a kernel, which is a function with the property that [Courant and Hilbert, 1970]:

$$K(x,y) = \Phi(x) \cdot \Phi(y), \tag{2.28}$$

where $x, y \in \mathbb{R}^n$.

SVMs require that the kernel is a positive (semi-)definite function in order for the standard solving approach to find a solution to the optimization problem [Boser et al, 1992]. Such a kernel is one that satisfies Mercer's theorem from functional analysis and is therefore required to be a dot product in some space [Burges, 1998].

Theorem 2.2. *[Mercer, 1908]*

Let K(x,y) be a continuous symmetric kernel that is defined in the closed interval $a \leq x \leq b$ *and likewise for y. The kernel K(x,y) can be expanded in the series*

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \Phi(x)_i \Phi(y)_i$$

with positive coefficients, $\lambda_i > 0$ for all i. For this expansion to be valid and for it to converge absolutely and uniformly, it is necessary that the condition

$$\int_a^b \int_a^b K(x,y)\psi(x)\psi(y)dxdy \geq 0$$

holds for all $\psi(\cdot)$ for which

$$\int_a^b \psi^2(x)dx < \infty$$

Restricting the kernel to be positive (semi-)definite has two drawbacks [Mierswa, 2006b]. On the one hand, it is difficult to check Mercer's condition for a newly constructed kernel. On the other hand, kernels that fail to meet the conditions of the theorem might have proven to achieve a better separation of the training samples.

When applying SVMs for a classification task, there are a couple of classical kernels that had been demonstrated to meet Mercer's condition [Vapnik, 1995b]:

- the polynomial kernel of degree p: $K(x,y) = (x \cdot y)^p$
- the radial basis function kernel: $K(x,y) = e^{-\sigma\|x-y\|^2}$,

where p and σ are parameters of the SVM.

One may state the DP in this new case by simply replacing the dot product between data points with the chosen kernel, as below:

$$\begin{cases} \text{find } \{\alpha_i\}_{i=1,2,\ldots,m} \text{ as to maximize } Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{subject to } \begin{cases} \sum_{i=1}^{m} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases}, C > 0 \end{cases} \tag{2.29}$$

As generally one is not able to construct the mapping Φ from the kernel K, the value for the optimum vector w cannot always be determined explicitly from:

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

Consequently, one usually has to directly determine the class for a new data sample x', as follows:

$$class(x') = sgn(w \cdot \Phi(x') - b)$$

Therefore, by replacing w with $\sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$, one gets:

$$\begin{aligned} class(x') &= sgn(w \cdot \Phi(x') - b) \\ &= sgn(\sum_{i=1}^{m} \alpha_i y_i \Phi(x) \cdot \Phi(x_i) - b) \\ &= sgn(\sum_{i=1}^{m} \alpha_i y_i K(x, x_i) - b) \end{aligned}$$

One is left to determine the value of b. This is done by replacing the dot product by the kernel in the formula for the linear case, *i.e.* when $0 < \alpha_i < C$:

$$b = \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i) - y_i,$$

and taking the mean of all the values obtained for b.

2.5 Support Vector Machines for Multi-class Learning

Multi-class SVMs build several two-class classifiers that separately solve the corresponding tasks. The translation from multi-class to two-class is performed through different systems, among which one-against-all, one-against-one or decision directed acyclic graph are the most commonly employed.

Resulting SVM decision functions are considered as a whole and the class for each sample in the test set is decided by the corresponding system [Hsu and Lin, 2004] .

2.5.1 One-Against-All

The one-against-all technique [Hsu and Lin, 2004] builds k classifiers. Every i^{th} SVM considers all training samples labeled with i as positive and all the remaining ones as negative.

The aim of every i^{th} SVM is thus to determine the optimal coefficients w and b of the decision hyperplane to separate the samples with outcome i from all the other samples in the training set, such that (2.30) :

$$\begin{cases} \text{find } w^i \text{ and } b^i \text{ as to minimize } \dfrac{\|w^i\|^2}{2} + C \sum_{j=1}^{m} \xi_j^i \\ \text{subject to } y_j(w^i \cdot x_j - b^i) \geq 1 - \xi_j^i, \xi_j^i \geq 0, \text{ for all } j = 1, 2, ..., m. \end{cases} \tag{2.30}$$

Once the all hyperplanes are determined following the classical SVM solving as in the earlier pages, the class for a test sample x' is given by the category that has the maximum value for the learning function, as in (2.31):

$$class(x') = argmax_{i=1,2,...,k}(w^i \cdot \Phi(x')) - b^i) \tag{2.31}$$

2.5.2 One-Against-One and Decision Directed Acyclic Graph

The one-against-one technique [Hsu and Lin, 2004] builds $\frac{k(k-1)}{2}$ SVMs. Every i^{th} machine is trained on data from every two classes, i and j, where samples labelled with i are considered positive while those in class j are taken as negative.

The aim of every SVM is hence to determine the optimal coefficients of the decision hyperplane to discriminate the samples with outcome i from the samples with outcome j, such that (2.32) :

$$\begin{cases} \text{find } w^{ij} \text{ and } b^{ij} \text{ as to minimize } \dfrac{\|w^{ij}\|^2}{2} + C\sum_{l=1}^{m} \xi_l^{ij}, \\ \text{subject to } y_l(w^{ij} \cdot x_l - b^{ij}) \geq 1 - \xi_l^{ij}, \xi_l^{ij} \geq 0, \text{ for all } l = 1,2,...,m \end{cases} \tag{2.32}$$

When the hyperplanes of the $\frac{k(k-1)}{2}$ SVMs are found, a voting method is used to determine the class for a test sample x'. For every SVM, the class of x' is computed by following the sign of its resulting decision function applied to x'. Subsequently, if the sign says x' is in class i, the vote for the i-th class is incremented by one; conversely, the vote for class j is increased by unity. Finally, x' is taken to belong to the class with the largest vote. In case two classes have an identical number of votes, the one with the smaller index is selected.

Classification within the decision directed acyclic graph technique [Platt et al, 2000] is done in an identical manner to that of one-against-one.

For the second part, after the hyperplanes of the $\frac{k(k-1)}{2}$ SVMs are discovered, the following graph system is used to determine the class for a test sample x (Fig. 2.5). Each node of the graph has an attached list of classes and considers the first and last elements of the list. The list that corresponds to the root node contains all k classes. When a test instance x is evaluated, one descends from node to node, in other words, eliminates one class from each corresponding list, until the leaves are reached.

The mechanism starts at the root node which considers the first and last classes. At each node, i vs j, we refer to the SVM that was trained on data from classes i and j. The class of x is computed by following the sign of the corresponding decision function applied to x. Subsequently, if the sign says x is in class i, the node is exited via the right edge; conversely, we exit through the left edge. We thus eliminate the wrong class from the list and proceed via the corresponding edge to test the first and last classes of the new list and node. The class is given by the leaf that x eventually reaches.

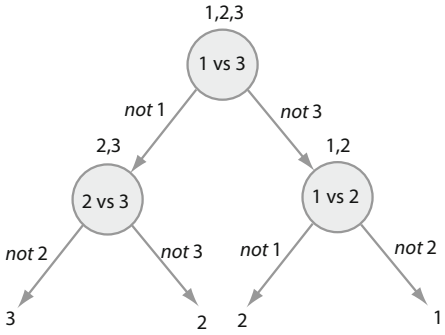

Fig. 2.5 An example of a 3-class problem labeled by a decision directed acyclic graph

2.6 Concluding Remarks

SVMs provide a very interesting and efficient vision upon classification. They pursue a geometrical interpretation of the relationship between samples and decision surfaces and thus manage to formulate a simple and natural optimization task.

On the practical side, when applying the technique for the problem at hand, one should first try a linear SVM (with possibly some errors) and only after this fails, turn to a nonlinear model; there, a radial kernel should generally do the trick.

Although very effective (as demonstrated by their many applications, like those described in [Kramer and Hein, 2009], [Kandaswamy et al, 2010], [Li et al, 2010], [Palmieri et al, 2013], to give only a few examples of their diversity), the standard solving of the reached optimization problem within SVMs is both intricate, as seen in this chapter, and constrained: the possibilities are limited to the kernels that obey Mercer's theorem. Thus, nonstandard possibly better performing decision functions are left aside. However, as a substitute for the original solving, direct search techniques (like the EAs) do not depend on the condition whether the kernel is positive (semi-) definite or not.