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).
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 \( t \):

\[
f_t : \mathbb{R}^n \rightarrow \{-1, 1\}.
\]

(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 \( \phi \) is called the confidence term and is defined as:

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

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.