The data to be classified is formally written as

\[ \Theta = \begin{Bmatrix} (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \end{Bmatrix} \]

\[ x_i \in \mathbb{R}^m \]

\[ y_i \in \{-1, 1\} \]

(Equation 1)

The nonlinear feature map \( \phi(x) : x \subset \mathbb{R}^m \rightarrow \mathbb{R}^d \) is never explicitly used in the calculation. Vapnik [9] suggests the form of the hyperplane \( f(x) \in F \) to be chosen from a family of functions with sufficient capacity. In particular, \( F \) contains functions for the linearly and non-linearly separable hyperplane having the following forms:

\[ f(x) = \sum_{i=1}^{n} w_i x_i + b \] (Equation 2)

\[ f(x) = \sum_{i=1}^{n} w_i \phi(x) + b \] (Equation 3)

Now for separation in feature space, we would like to obtain the hyperplane with the following properties:

\[ f(x) = \sum_{i=1}^{n} w_i \phi(x) + b \]

\[ f(x) > 0 \quad \forall i : y_i = +1 \]

\[ f(x) < 0 \quad \forall i : y_i = -1 \]

(Equation 4)

The conditions in equation Equation 4 can be described by a strict linear discriminant function, so that for each element pair in \( \Theta \) we require:

\[ y_i \left( \sum_{i=1}^{n} w_i \phi(x) + b \right) \geq 1 \]

(Equation 5)

The distance from the hyper-plane to points lying closest to it is given geometrically as \( \frac{1}{||w||} \).

The soft-margin minimization problem relaxes the strict discriminant in equation 5 by introducing slack variables, \( \xi_i \), and is formulated as:

\[
\min_{w, \xi} \frac{1}{2} \sum_{i=1}^{n} w_i^2 + C \sum_{i=1}^{n} \xi_i
\]

s.t \( y_i \left( \sum_{i=1}^{n} w_i \phi(x) + b \right) \geq 1 + \xi_i \)

\( \xi_i > 0 \quad \forall i = 1 .. n \) (Equation 6)

The constant C is selected so as to compromise between the minimization of training error and prevention of over-fitting. Applying Lagrangian Theory, the following dual problem in terms of Lagrange multipliers \( \alpha_i \) is usually solved

\[
\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)
\]

\[ \rho = \left\{ a \mid 0 \leq a_i \leq C \sum_{i=1}^{n} \alpha_i y_i = 0 \right\} \]

(Equation 7)

The explicit use of the nonlinear function \( \phi(.) \), has been circumvented by the use of a kernel function, defined formally as the dot products of the nonlinear functions

\[ K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \]

(Equation 8)

Kernels can be chosen according to Mercer’s theorem. In all our experiments we use polynomial kernel with degree \( d = 2 \) given by

\[ K(x_i, x_j) = (1 + x_i \cdot x_j)^d \] (Equation 9)

This was chosen based on preliminary experiments involving fewer protein chains. The SVM classifier is given by:

\[ f(x) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right) \] (Equation 10)