S

Support Vector Machines in Neuroscience

Onur Seref
University of Florida, USA

O. Erhun Kundakcioglu
University of Florida, USA

Michael Bewernitz
University of Florida, USA

INTRODUCTION

In a typical *binary classification* problem, each pattern vector $\mathbf{x}_i \in \Re^d, i = 1,...,n$ belongs to one of two classes S^+ and S^-. A vector is given the label $y_i = 1$ or $y_i = -1$ if $\mathbf{x}_i \in S^+$ or $\mathbf{x}_i \in S^-$, respectively. The set of pattern vectors and their corresponding labels constitute the *training set*. The classification problem consists of determining which class new pattern vectors from the test set belong to. SVMs solve this problem by finding a hyperplane ($\mathbf{w}$, b) that separates the two classes in the training set from each other with the maximum margin.

The underlying optimization problem for the maximal margin classifier is only feasible if the two classes of pattern vectors are linearly separable. However, most of the real life classification problems are not linearly separable. Nevertheless, the maximal margin classifier encompasses the fundamental methods used in standard SVM classifiers. The solution to the optimization problem in the maximal margin classifier minimizes the bound on the generalization error (Vapnik, 1998). The basic premise of this method lies in the minimization of a convex optimization problem with linear inequality constraints, which can be solved efficiently by many alternative methods (Bennett & Campbell, 2000).

A hyperplane can be represented by $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$ where $\mathbf{w}$ is the normal vector and b is the offset parameter. There is an inherent degree of freedom in specifying a hyperplane as $(\lambda \mathbf{w}, b\ \lambda)$. A *canonical hyperplane* is the one from which the closest pattern vector has a functional distance of 1, that is, $\min_{i=1,...,m} \left| \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right| = 1$.

Consider two pattern vectors $\mathbf{x}^+$ and $\mathbf{x}^-$ belonging to classes S^+ and S^-, respectively. Assuming these pattern vectors are the closest to a canonical hyperplane, such that $\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle + b = 1$ and $\langle \mathbf{w} \cdot \mathbf{x}^- \rangle + b = -1$, it is easy to show that the margin between these pattern vectors and the hyperplane are both equal to $1/\|\mathbf{w}\|$. Maximizing this margin while satisfying the canonicality condition for the pattern vectors turns out to be the following optimization problem.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{(1a)}$$

subject to

$$y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) \geq 1 \quad \forall \mathrm{i} = 1, \ldots, n \qquad \text{(1b)}$$

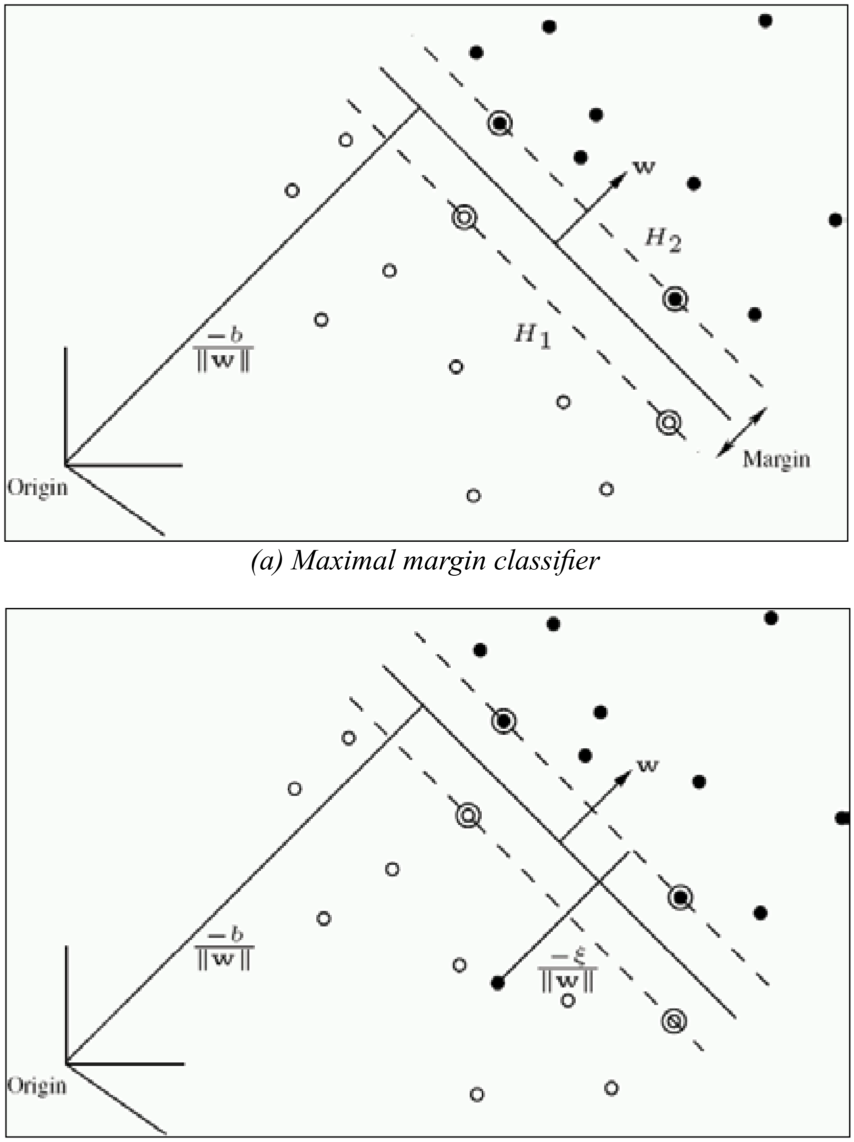
From the solution to (1), a new pattern vector $\mathbf{x}^*$ can be classified as positive if $\langle \mathbf{w} \cdot \mathbf{x}^* \rangle + b > 0$, and negative otherwise.

Most real life problems are composed of nonseparable data which is generally due to noise. In this case *slack variables* ξ_i are introduced for each pattern vector $\mathbf{x}_i$ in the training set. The slack variables allow misclassifications for each pattern vector with a penalty of $C/2$. In Figure 1-b, *soft margin classifier* is demonstrated that incurs penalty for misclassified pattern vectors. The maximum margin formulation can be augmented to soft margin formulation as follows:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 \qquad \text{(2a)}$$

subject to

Figure 1.



(a) Maximal margin classifier

(b) Soft margin classifier

$$y_i\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right) \geq 1-\xi_i \quad \forall i=1,\ldots,n \quad (2b)$$

In (2), nonnegativity of the slack variables is implied since the solution cannot be optimal when $\xi_i < 0$ for any pattern vector.

The 2-norm of the slack variables is penalized in the objective of (2). An alternative formulation involves penalization of the 1-norm slack variables in the objective where nonnegativity of the slack variables is forced (Cristianini & Shawe-Taylor, 2000).

Dual formulations for both 1-norm and 2-norm SVMs can be obtained using the optimization theory. The significance of the dual formulations is that they do not involve inequality constraints, and they allow the *kernel trick* to be introduced for nonlinear classification. The standard method to obtain the dual formulation of the SVM problem consists of two parts. First the *Lagrangian function* of the primal problem is derived. This function provides a lower bound for the solution of the primal problem. Next, the Lagrangian function is differentiated with respect to the primal variables and stationarity is imposed. Equivalent expressions for each