Support Vector Machines - I

Prof. Dan A. Simovici

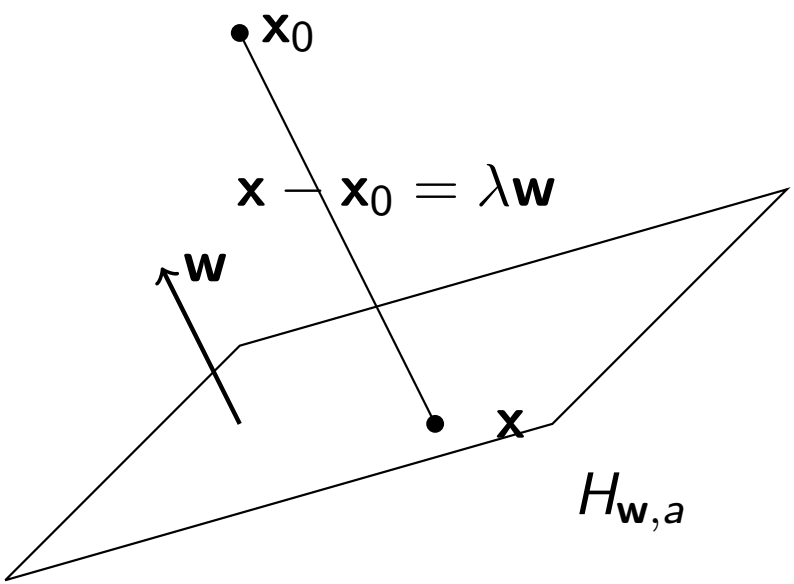
UMB

History

- Support vector machines (SVMs) were introduced in statistical learning theory by V. N. Vapnik.
- One of the intial success stories of SVMs was the handwritten digit recognition. The results obtained with SVMs show superior classification performance comparable with the best classifiers developed in machine learning.
- Although intended initially for classifying data where classes are linearly separable, using techniques from functional analysis, SVMs manage to successfuly classify data where classes are separated by non-linear boundaries.

The distance between a point $\mathbf{x}_0 \in \mathbb{R}^n$ and a hyperplane $H_{\mathbf{w},a}$ defined by the equation $\mathbf{w}'\mathbf{x} = a$ is given by:

$$d(H_{\mathbf{w},a}, \mathbf{x}_0) = \frac{|\mathbf{w}'\mathbf{x}_0 - a|}{\| \mathbf{w} \|}.$$



Data Samples

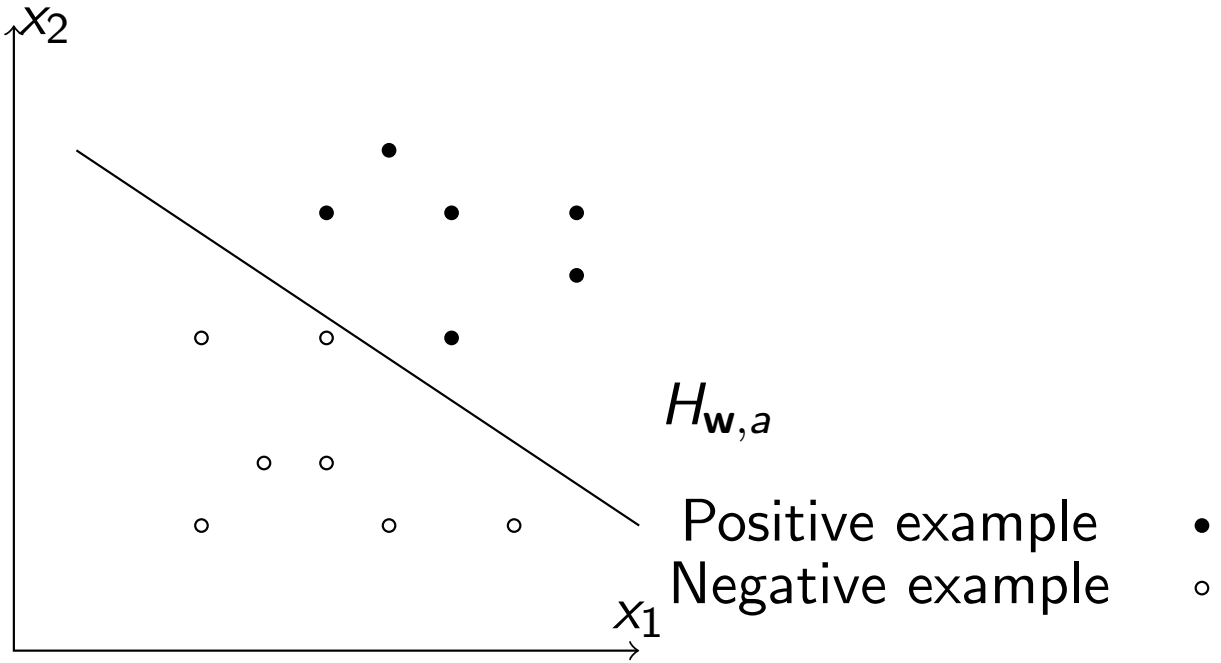
A data sample of size m is a sequence

$$\mathbf{s} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)),$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ belong to $\mathbb{R}^n$ and $y_i \in \{-1, 1\}$ for $1 \leqslant i \leqslant m$. The positive examples of $\mathbf{s}$ are those pairs $(\mathbf{x}_i, y_i)$ such that $y_i = 1$; the remaining pairs are the negative examples.

The Task of a Linear Classifier

The task of a linear classifier is to construct a hyperplane $H_{\mathbf{w},a}$ starting from the sample $\mathbf{s}$ such that for each positive example $(\mathbf{x}_i, 1)$ we have $\mathbf{x}_i \in H^{>0}_{\mathbf{w},a}$ and for each negative example we have $(\mathbf{x}_i, -1) \in H^{<0}_{\mathbf{w},a}$. If $\mathbf{s}$ is a linearly separable sample there are, in general, infinitely many hyperplanes that can do the separation.



Hyperplanes in Canonical Form

Definition

A hyperplane $H_{\mathbf{w}',a}$ that does not pass through a point of the sample $\mathbf{s}$ is in canonical form relative to $\mathbf{s}$ if

$$\min_{(\mathbf{x},y)\in S} |\mathbf{w}'\mathbf{x} - a| = 1.$$

We may always assume that the separating hyperplane is in canonical form relative by $\mathbf{s}$ by rescaling the coefficients of the equation that define the hyperplane (a and the components of $\mathbf{w}$).

If the hyperplane $\mathbf{w}'\mathbf{x} = a$ is in canonical form relative to the sample $\mathbf{s}$, then the distance to the hyperplane to the closest points in $\mathbf{s}$ (the margin of the hyperplane) is the same, namely,

$$\rho = \min_{(\mathbf{x},y)\in S} \frac{|\mathbf{w}'\mathbf{x} - a|}{\| \mathbf{w} \|} = \frac{1}{\| \mathbf{w} \|}.$$

Points that are closest to the separating hyperplane are referred to as support vectors.

For a canonical separating hyperplane we have

$$|\mathbf{w}'\mathbf{x} - a| \geqslant 1$$

for any point $(\mathbf{x}, y)$ of the sample, and

$$|\mathbf{w}'\mathbf{x} - a| = 1$$

for every support point. The point $(\mathbf{x}_i, y_i)$ is classified correctly if y_i has the same sign as $\mathbf{w}'\mathbf{x}_i - a$, that is, $y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1$.

Maximizing the margin is equivalent to minimizing $\| \mathbf{w} \|$ or, equivalently, to minimizing $\frac{1}{2} \| \mathbf{w} \|^2$. Thus, if $\mathbf{s}$ is separable, the SVM problem is equivalent to the following convex optimization problem:

$$\begin{aligned} &\textit{minimize } \tfrac{1}{2} \| \mathbf{w} \|^2 \\ &\qquad \textit{subject to } y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1 \textit{ for } 1 \leqslant i \leqslant m. \end{aligned}$$

Note that this objective function,

$$\frac{1}{2} \parallel \mathbf{w} \parallel^2 = \frac{1}{2}(w_1^2 + \cdots + w_n^2)$$

is differentiable; furhermore, we have $\nabla \left(\frac{1}{2} \parallel \mathbf{w} \parallel^2 \right) = \mathbf{w}$ and the Hessian of this function is

$$H_{\frac{1}{2}\|\mathbf{w}\|^2} = \mathbf{I}_n,$$

which shows that $\frac{1}{2} \parallel \mathbf{w} \parallel^2$ is a convex function of $\mathbf{w}$.

The Lagrangian of the primal optimization problem

$$\begin{aligned} &\textit{minimize } \tfrac{1}{2} \parallel \mathbf{w} \parallel^2 \\ &\qquad \textit{subject to } y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1 \textit{ for } 1 \leqslant i \leqslant m. \end{aligned}$$

is

$$L(\mathbf{w}, a, \mathbf{u}) = \frac{1}{2} \parallel \mathbf{w} \parallel^2 - \sum_{i=1}^{m} u_i \left(y_i(\mathbf{w}'\mathbf{x}_i - a) - 1 \right),$$

where u_i are the Lagrange multipliers for $1 \leqslant i \leqslant m$.

To compute the dual objective function g we impose the Karush-Kuhn-Tucker optimality conditions on the Lagrangian L:

$$\begin{aligned}
\frac{\partial L}{\partial w_j} &= w_j - \sum_{i=1}^{m} u_i y_i (\mathbf{x}_i)_j = 0 \\
\frac{\partial L}{\partial a} &= \sum_{i=1}^{m} u_i y_i = 0, \\
u_i(y_i(\mathbf{w}'\mathbf{x}_i + b) - 1) &= 0 \text{ for all } i,
\end{aligned}$$

which imply

$$\begin{aligned}
\mathbf{w} &= \sum_{i=1}^{m} u_i y_i \mathbf{x}_i = 0, \\
\sum_{i=1}^{m} u_i y_i &= 0, \\
u_i &= 0 \text{ or } y_i(\mathbf{w}'\mathbf{x}_i - a) = 1 \text{ for } 1 \leqslant i \leqslant m.
\end{aligned}$$

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{m} u_i y_i \mathbf{x}_i = 0, \\ \sum_{i=1}^{m} u_i y_i &= 0, \\ u_i &= 0 \text{ or } y_i(\mathbf{w}'\mathbf{x}_i - a) = 1 \text{ for } 1 \leqslant i \leqslant m. \end{aligned}$$

Conclusions:

- The weight vector is a linear combination of the training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$.
- $\mathbf{x}_i$ effectively occurs in the linear combination that defines the weight vector only if $u_i \neq 0$, that is, if $\mathbf{x}_i$ is a support vector.

Since $u_i = 0$ or $y_i(\mathbf{w}'\mathbf{x}_i - a) = 1$ for all i, if $u_i \neq 0$, then $y_i(\mathbf{w}'\mathbf{x}_i - a) = 1$ for the support vectors; thus, all these vectors lie on the marginal hyperplanes $\mathbf{w}'\mathbf{x} - a = 1$ or $\mathbf{w}'\mathbf{x} - a = -1$. If non-support vector are removed the solution remains the same; however, while the solution of the problem remains the same different choices may be possible for the support vectors.

The dual probem can now be stated as follows:

$$\begin{aligned} &\textit{maximize } g(\mathbf{u}) = \sum_{i=1}^{m} u_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} u_i u_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \\ &\qquad \textit{subject to } u_i \geqslant 0 \textit{ for } 1 \leqslant i \leqslant m \textit{ and } \sum_{i=1}^{m} u_i y_i = 0. \end{aligned}$$

The dual objective function $g(\mathbf{u})$ depends on the vector of Lagrange multipliers $\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$. Constraints are affine, so the strong duality holds; therefore, the primal and the dual problems are equivalent.
The solution $\mathbf{u}$ of the dual problem can be used directly to determine the classifying function returned by the SVM as:

$$h(\mathbf{x}) = sign(\mathbf{w}'\mathbf{x} - a) = sign\left(\sum_{i=1}^{m} u_i y_i (\mathbf{x}_i' \mathbf{x}) - a \right).$$

Since support vectors lie on the marginal hyperplanes, for every support vector $\mathbf{x}_i$ we have $\mathbf{w}'\mathbf{x}_i - a = y_i$, so

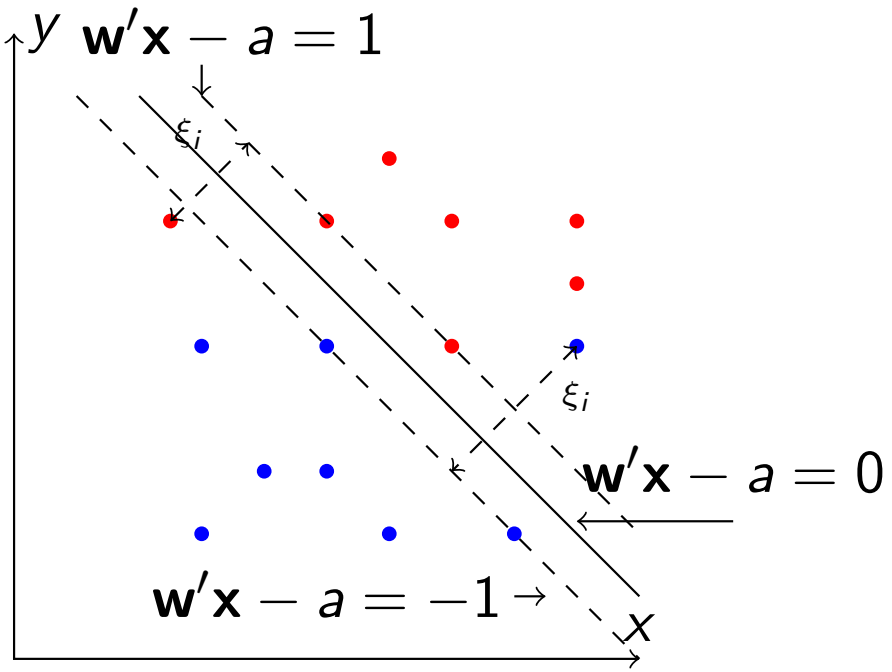
$$a = \sum_{j=1}^{m} u_j y_j (\mathbf{x}_j' \mathbf{x}) - y_i.$$

If data is not separable the conditions $y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1$ cannot all hold (for $1 \leqslant i \leqslant m$). Instead, we impose a relaxed version, namely

$$y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1 - \xi_i,$$

where ξ_i are new variables known as slack variables.
A slack variable ξ_i measures the amount by which $\mathbf{x}_i$ violates the desired inequality $y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1$.

Objects missclassified and slack variables



Outliers

- A vector $\mathbf{x}_i$ is an outlier if $\mathbf{x}_i$ is not positioned correctly on the side of the appropriate hyperplane.
- A vector $\mathbf{x}_i$ with $0 < y_i(\mathbf{w}'\mathbf{x}_i - a) < 1$ is still an outlier even if it is correctly classified by the hyperplane $\mathbf{w}'\mathbf{x} - a = 0$ but is misplaced relative to the shifted separating hyperplane.
- If we omit the outliers the data is correctly separated by the hyperplane $\mathbf{w}'\mathbf{x} - a = 0$ with a soft margin $\rho = \frac{1}{\|\mathbf{w}\|}$.
  The total slack due to outliers can be estimated as $\sum_{i=1}^{m} \xi_i$. We seek a hyperplane with a large margin (even though this may lead to more outliers).

The optimization problem for the non-separable data is:

$$\begin{aligned} &\textit{minimize } \tfrac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{m} \xi_i \\ &\qquad \textit{subject to } y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1 - \xi_i \textit{ and } \xi_i \geqslant 0 \textit{ for } 1 \leqslant i \leqslant m. \end{aligned}$$

The parameter C is determined in the process of cross-validation.
This is a convex optimization problem with affine constraints. As in the separable case constraints are affine and thus, qualified, the objective function and the affine constraints are convex and differentiable, so the KKT conditions apply.

If $u_i \geqslant 0$ are Lagrange multipliers associated with the constraints $y_i(\mathbf{w}'\mathbf{x}_i - a) \geqslant 1 - \xi_i$ and $v_i \geqslant 0$ for $1 \leqslant i \leqslant m$ are Lagrange multipliers associated with the non-negativity constraints of the slack variables for $1 \leqslant i \leqslant m$, then the Lagrangian is defined as:

$$\begin{aligned} L(\mathbf{w}, a, \xi_1, \ldots, \xi_m, \mathbf{u}, \mathbf{v}) \quad = \quad & \tfrac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{m} \xi_i \\ & - \sum_{i=1}^{m} u_i [y_i(\mathbf{w}'\mathbf{x}_i - a) - 1 + \xi_i] \\ & - \sum_{i=1}^{n} v_i \xi_i . \end{aligned}$$

where $\mathbf{v}$ is the vector whose components are v_i.

Karush-Kahn-Tucker Conditions

The KKT conditions are:

KKT Condition	Consequence
$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} u_i y_i \mathbf{x}_i = 0$	$\mathbf{w} = \sum_{i=1}^{m} u_i y_i \mathbf{x}_i$
$\nabla_a L = \sum_{i=1}^{m} u_i y_i = 0$	$\sum_{i=1}^{m} u_i y_i = 0$
$\nabla_{\xi_i} L = C - u_i - v_i = 0$ for $1 \leqslant i \leqslant m$	$u_i + v_i = C$ for $1 \leqslant i \leqslant m$
$u_i[y_i(\mathbf{w}'\mathbf{x}_i - a) - 1 + \xi_i] = 0$ for $1 \leqslant i \leqslant m$	$u_i = 0$ or $y_i(\mathbf{w}'\mathbf{x}_i - a) = 1 - \xi_i$ for $1 \leqslant i \leqslant m$
$v_i \xi_i = 0$ for $1 \leqslant i \leqslant m$	$v_i = 0$ or $\xi_i = 0$ for $1 \leqslant i \leqslant m$

Consequences of KKT:

- $\mathbf{w}$ is a linear combination of the training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$, where $\mathbf{x}_i$ appears in the combination only if $u_i \neq 0$;
- if $u_i \neq 0$, then $y_i(\mathbf{w}'\mathbf{x}_i - a) = 1 - \xi_i$;
- if $\xi_i = 0$, then $y_i(\mathbf{w}'\mathbf{x}_i - a) = 1$ and $\mathbf{x}_i$ lies on marginal hyperplane as in the separable case; otherwise, $\mathbf{x}_i$ is an outlier;
- if $\mathbf{x}_i$ is an outlier, $v_i = 0$ and $u_i = C$ or $\mathbf{x}_i$ is located on the marginal hyperplane.
- $\mathbf{w}$ is unique; the support vectors are not.

The objective function of the dual problem is obtained by substituting $\mathbf{w}$ and incorporating the consequences of the KKT conditions:

$$
\begin{aligned}
& g(\mathbf{u}, \mathbf{v}) \\
&= \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m u_i \xi_i - \sum_{i=1}^m v_i \xi_i \\
&\quad - \sum_{i=1}^m u_i [y_i (\mathbf{w}' \mathbf{x}_i - a) - 1] \\
&= \frac{1}{2} \parallel \mathbf{w} \parallel^2 + \sum_{i=1}^m (C - u_i - v_i) \xi_i - \sum_{i=1}^m u_i [y_i (\mathbf{w}' \mathbf{x}_i - a) - 1] \\
&= \frac{1}{2} \left\| \sum_{i=1}^m u_i y_i \mathbf{x}_i \right\|^2 - \sum_{i=1}^m \sum_{j=1}^m u_i u_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \\
&\quad + \sum_{i=1}^m u_i y_i a + \sum_{i=1}^m u_i \\
&= \sum_{i=1}^m u_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m u_i u_j y_i y_j \mathbf{x}_i' \mathbf{x}_j .
\end{aligned}
$$

Note that g depends only on $\mathbf{u}$ and we have:

$$g(\mathbf{u}) = \sum_{i=1}^{m} u_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} u_i u_j y_i y_j \mathbf{x}_i' \mathbf{x}_j.$$

has exactly the same form as in the separable case. Also, observe that $0 \leqslant u_i \leqslant C$ for $1 \leqslant i \leqslant m$ because both u_i and v_i are non-negative and $u_i + v_i = C$.
The dual optimization problem for the non-separable case becomes:

$$\begin{aligned} &\textit{maximize } \textstyle\sum_{i=1}^{m} u_i - \frac{1}{2} u_i u_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \\ &\quad \textit{subject to } 0 \leqslant u_i \leqslant C \textit{ and } \textstyle\sum_{i=1}^{m} u_i y_i = 0 \\ &\quad \textit{for } 1 \leqslant i \leqslant m. \end{aligned}$$

The objective function is concave and differentiable and the solution can be used to determine the separating hyperplane. As in the separable case, the hyperplane depends only on the inner products between the vectors and not directly on the vectors themselves.