Support Vector Machines - I

Prof. Dan A. Simovici

UMB
Linear Classification
Problem Setting

- the input space is $\mathcal{X} \subseteq \mathbb{R}^n$;
- the output space is $\mathcal{Y} = \{-1, 1\}$;
- concept sought: a function $f : \mathcal{X} \rightarrow \mathcal{Y}$;
- sample: a sequence $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ extracted from a distribution $\mathcal{D}$. 
Problem Statement

- the hypothesis space $H$ is $H \subseteq \mathcal{Y}^\mathcal{X}$;
- task: find $h \in H$ such that the generalization error

$$L_D(h) = P_{x \sim D}(h(x) \neq f(x))$$

is small.

The smaller the $\text{VCD}(H)$ the more efficient the process is. One possibility is the class of linear functions from $\mathcal{X}$ to $\mathcal{Y}$:

$$H = \{ x \mapsto \text{sign}(w'x + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R} \},$$

where

$$\text{sign}(a) = \begin{cases} 1 & \text{if } a \geq 0, \\ -1 & \text{if } a < 0. \end{cases}$$
A Fundamental Assumption: Linear Separability of $S$

If $S$ is linearly separable there are, in general, infinitely many hyperplanes that can do the separation.
SVMs seek the hyperplane with the maximum separation margin.
The distance of a point $x_0$ to a hyperplane $w'x + b = 0$

Equation of the line passing through $x_0$ and perpendicular on the hyperplane is

$$x - x_0 = tw;$$

Since $z$ is a point on this line that belongs to the hyperplane, to find the value of $t$ that corresponds to $z$ we must have $w'(x_0 + tw) + b = 0$, that is,

$$t = -\frac{w'x_0 + b}{\|w\|^2}$$
The distance of a point $x_0$ to a hyperplane $w'x + b = 0$

Thus, $z = x_0 - \frac{w'x_0 + b}{\|w\|^2}w$, hence the distance from $x_0$ to the hyperplane is

$$\|x_0 - z\| = \frac{|w'x_0 + b|}{\|w\|}.$$
We seek a hyperplane in $\mathbb{R}^n$ having the equation

$$\mathbf{w}' \mathbf{x} + b = 0,$$

where $\mathbf{w} \in \mathbb{R}^n$ is a vector normal to the hyperplane and $b \in \mathbb{R}$ is a scalar. A hyperplane $\mathbf{w}' \mathbf{x} + b = 0$ that does not pass through a point of $S$ is in canonical form relative to a sample $S$ if

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

Note that we may always assume that the separating hyperplane are in canonical form relative by $S$ by rescaling the coefficients of the equation that define the hyperplane (the components of $\mathbf{w}$ and $b$).
If the hyperplane $\mathbf{w}' \mathbf{x} + b = 0$ is in canonical form relative to the sample $S$, then the distance to the hyperplane to the closest points in $S$ (the margin of the hyperplane) is the same, namely,

$$
\rho = \min_{(x,y) \in S} \frac{|\mathbf{w}' \mathbf{x} + b|}{\| \mathbf{w} \|} = \frac{1}{\| \mathbf{w} \|}.
$$
For a canonical separating hyperplane we have

\[ |w'x + b| \geq 1 \]

for any point \((x, y)\) of the sample and

\[ |w'x + b| = 1 \]

for every support point. The point \((x_i, y_i)\) is classified correctly if \(y_i\) has the same sign as \(w'x_i + b\), that is, \(y_i(w'x_i + b) \geq 1\).

Maximizing the margin is equivalent to minimizing \(\|w\|\) or, equivalently, to minimizing \(\frac{1}{2} \|w\|^2\). Thus, in the separable case the SVM problem is equivalent to the following convex optimization problem:

- minimize \(\frac{1}{2} \|w\|^2\);
- subjected to \(y_i(w'x_i + b) \geq 1\) for \(1 \leq i \leq m\).
Why $\frac{1}{2} \| \mathbf{w} \|^2$?

Note that this objective function,

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

is differentiable!

We have $\nabla \left( \frac{1}{2} \| \mathbf{w} \|^2 \right) = \mathbf{w}$ and that

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

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

The Lagrangean of the optimization problem

- minimize $\frac{1}{2} \| \mathbf{w} \|^2$;
- subjected to $y_i (\mathbf{w}' \mathbf{x}_i + b) \geq 1$ for $1 \leq i \leq m$.

is

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \| \mathbf{w} \|^2 - \sum_{i=1}^{m} a_i \left( y_i (\mathbf{w}' \mathbf{x}_i + b) - 1 \right).$$
The Karush-Kuhn-Tucker Optimality Conditions

\[ \nabla_w L = w - \sum_{i=1}^{m} a_i y_i x_i = 0, \]

\[ \nabla_b L = -\sum_{i=1}^{m} a_i y_i = 0, \]

\[ a_i(y_i(w^\prime x_i + b) - 1) = 0 \text{ for all } i \]

imply

\[ w = \sum_{i=1}^{m} a_i y_i x_i = 0, \]

\[ \sum_{i=1}^{m} a_i y_i = 0, \]

\[ a_i = 0 \text{ or } y_i(w^\prime x_i + b) = 1 \text{ for } 1 \leq i \leq m. \]
the weight vector is a linear combination of the training vectors $x_1, \ldots, x_m$, where $x_i$ appears in this combination only if $a_i \neq 0$ (support vectors);

since $a_i = 0$ or $y_i(w'x_i + b) = 1$ for all $i$, if $a_i \neq 0$, then $y_i(w'x_i + b) = 1$ for the support vectors; thus, all these vectors lie on the marginal hyperplanes $w'x + b = 1$ or $w'x + b = -1$;

if non-support vector are removed the solution remains the same;

while the solution of the problem $w$ remains the same different choices may be possible for the support vectors.