Support Vector Machines - I

Prof. Dan A. Simovici

UMB

Problem Setting

- the input space is $\mathcal{X} \subseteq \mathbb{R}^n$;
- the output space is $\mathcal{Y} = \{-1, 1\}$;
- concept sought: a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$;
- sample: a sequence $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ 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

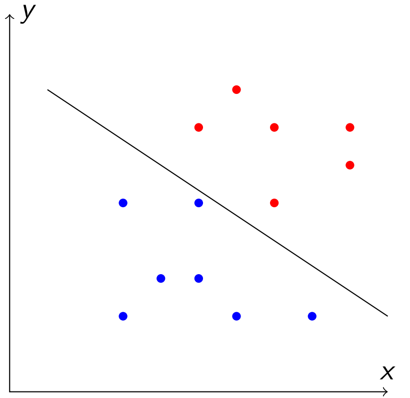
$$R(h) = P_{x \sim \mathcal{D}}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$$

  is small.

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

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

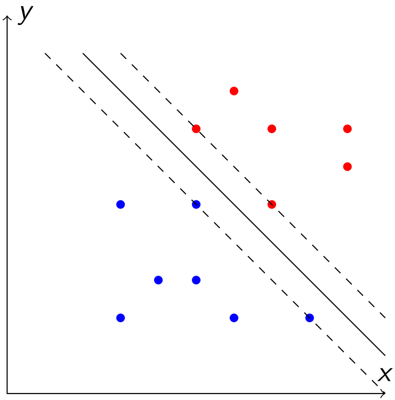
A Fundamental Assumption: Linear Separability of S



If S is linearly separable there are, in general, infinitely many hyperplanes that can do the separation.

Solution returned by SVMs

SVMs seek the hyperplane with the maximum separation margin.



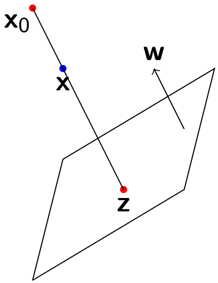
The distance of a point $\boldsymbol{x}_0$ to a hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$

Equation of the line passing through $\boldsymbol{x}_0$ and perpendicular on the hyperplane is

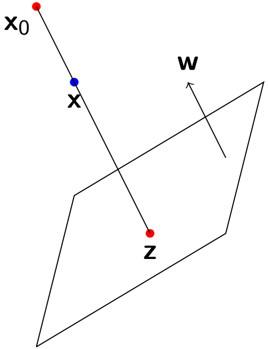
$$\boldsymbol{x} - \boldsymbol{x}_0 = t\boldsymbol{w};$$

Since $\boldsymbol{z}$ is a point on this line that belongs to the hyperplane, to find the value of t that corresponds to $\boldsymbol{z}$ we must have $\boldsymbol{w}'(\boldsymbol{x}_0 + t\boldsymbol{w}) + b = 0$, that is,

$$t = -\frac{\boldsymbol{w}'\boldsymbol{x}_0 + b}{\parallel \boldsymbol{w} \parallel^2}$$



The distance of a point $\boldsymbol{x}_0$ to a hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$



Thus, $\boldsymbol{z} = \boldsymbol{x}_0 - \frac{\boldsymbol{w}'\boldsymbol{x}_0 + b}{\|\boldsymbol{w}\|^2}\boldsymbol{w}$, hence the distance from $\boldsymbol{x}_0$ to the hyperplane is

$$\| \boldsymbol{x}_0 - \boldsymbol{z} \| = \frac{|\boldsymbol{w}'\boldsymbol{x}_0 + b|}{\| \boldsymbol{w} \|}.$$

Primal Optimization Problem

We seek a hyperplane in $\mathbb{R}^n$ having the equation

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

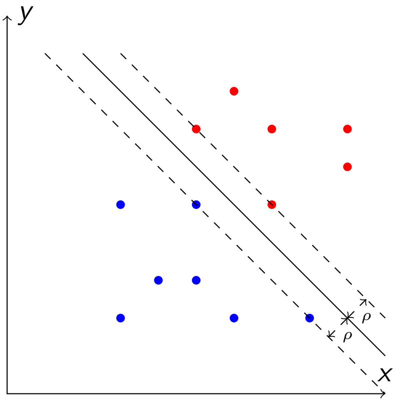
where $\boldsymbol{w} \in \mathbb{R}^n$ is a vector normal to the hyperplane and $b \in \mathbb{R}$ is a scalar. A hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$ that does not pass through a point of S is in canonical form relative to a sample S if

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

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 $\boldsymbol{w}$ and b).

If the hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$ 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,

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



Canonical Separating Hyperplane

For a canonical separating hyperplane we have

$$|\boldsymbol{w}'\boldsymbol{x} + b| \geqslant 1$$

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

$$|\boldsymbol{w}'\boldsymbol{x} + b| = 1$$

for every support point. The point $(\boldsymbol{x}_i, y_i)$ is classified correctly if y_i has the same sign as $\boldsymbol{w}'\boldsymbol{x}_i + b$, that is, $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geqslant 1$.
Maximizing the margin is equivalent to minimizing $\parallel \boldsymbol{w} \parallel$ or, equivalently, to minimizing $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$. Thus, in the separable case the SVM problem is equivalent to the following convex optimization problem:

- minimize $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$;
- subjected to $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geqslant 1$ for $1 \leqslant i \leqslant m$.

Why $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$?

Note that this objective function,

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

is differentiable!
We have $\nabla \left(\frac{1}{2} \parallel \boldsymbol{w} \parallel^2 \right) = \boldsymbol{w}$ and that

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

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

Support Vectors

The Lagrangean of the optimization problem

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

is

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

The Karush-Kuhn-Tucker Optimality Conditions

$$\begin{aligned} \nabla_{\boldsymbol{w}} L &= \boldsymbol{w} - \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i = 0, \\ \nabla_b L &= -\sum_{i=1}^{m} a_i y_i = 0, \\ a_i(y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) - 1) &= 0 \text{ for all } i \end{aligned}$$

imply

$$\begin{aligned} \boldsymbol{w} &= \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i = 0, \\ \sum_{i=1}^{m} a_i y_i &= 0, \\ a_i &= 0 \text{ or } y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1 \text{ for } 1 \leqslant i \leqslant m. \end{aligned}$$

Consequences of the KKT Conditions

- the weight vector is a linear combination of the training vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$, where $\mathbf{x}_i$ appears in this combination only if $a_i \neq 0$ (support vectors);
- since $a_i = 0$ or $y_i(\mathbf{w}'\mathbf{x}_i + b) = 1$ for all i, if $a_i \neq 0$, then $y_i(\mathbf{w}'\mathbf{x}_i + b) = 1$ for the support vectors; thus, all these vectors lie on the marginal hyperplanes $\mathbf{w}'\mathbf{x} + b = 1$ or $\mathbf{w}'\mathbf{x} + b = -1$;
- if non-support vector are removed the solution remains the same;
- while the solution of the problem $\mathbf{w}$ remains the same different choices may be possible for the support vectors.