SUPPORT VECTOR MACHINES Part 0

Prof. Dan A. Simovici

UMB

Support Vector Machines

Support Vector Machines (SVM) were developed by Vapnik in [1] and [2]. SVMs seek to find a hypothesis H for which one can guarantee the lowest probability of error for a sample

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

where $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$ for $1 \leqslant i \leqslant m$.

What do SVMs Learn?

SVMs learn linear threshold functions of the type

$$H(\mathbf{x}) = sign(\mathbf{w}'\mathbf{x} + b) = \begin{cases} 1 & \text{if } \mathbf{w}'\mathbf{x} + b > 0, \\ -1 & \text{otherwise} \end{cases}$$

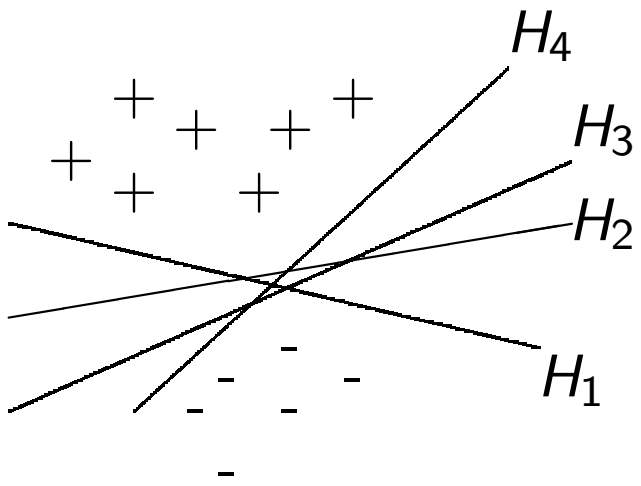
Each such function corresponds to a hyperplane $\mathbf{w}'\mathbf{x} + b = 0$.

Linear Hard-Margin SVMs

Suppose that the training data can be separated by at least one hyperplane H':

- all positive training examples are on one side of the hyperplane;
- all negative examples are on the other side of H'.

If H' has the equation $\mathbf{w}'\mathbf{x} + b = 0$, this is equivalent to $y_i(\mathbf{w}'\mathbf{x}_i + b) > 0$ for $1 \leqslant i \leqslant m$.



Let $\mathbf{w}'\mathbf{x} + b = 0$ be a hyperplane H in $\mathbb{R}^n$.
The vector $\mathbf{w}$ is orthogonal to H, so the line that passes through $\mathbf{x}_0$ and is orthogonal to the hyperplane is

$$\mathbf{x} - \mathbf{x}_0 = a\mathbf{w}.$$

The intersection of this line with the hyperplane is

$$\mathbf{w}'(\mathbf{x}_0 + a\mathbf{w}) + b = 0,$$

so

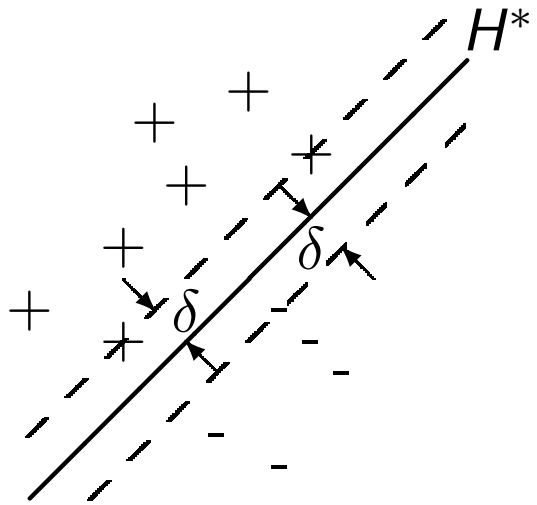
$$a = -\frac{\mathbf{w}'\mathbf{x}_0 + b}{\| \mathbf{w} \|^2}.$$

Consequently, the projection of $\mathbf{x}$ on the hyperplane is

$$\mathbf{x}_0 - \frac{\mathbf{w}'\mathbf{x}_0 + b}{\| \mathbf{w} \|^2}\mathbf{w}.$$

and the distance from $\mathbf{x}_i$ to H is $\frac{|\mathbf{w}'\mathbf{x}_0 + b|}{\|\mathbf{w}\|}$.

- in general, if the data is linearly separable, there could be several hyperplanes that do the separation;
- the best separating hyperplane is the one for which the distance to the closest examples is the largest;



- this largest distance is the margin δ.

- for each separable training set there exists only one hyperplane with maximum margin;
- the example closest to this hyperplane are the support vectors; they have distance to the hyperplane equal to δ.

Finding the Best Hyperplane as an Optimization Problem

We require the stronger conditions

$$y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1$$

for $1 \leqslant i \leqslant m$ instead of $y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 0$. These conditions imply $|\mathbf{w}'\mathbf{x}_i + b| \geqslant 1$ for $1 \leqslant i \leqslant m$.
If $\delta = \min_i \frac{|\mathbf{w}'\mathbf{x}_i + b|}{\|\mathbf{w}\|}$, then by the previous restrictions, $\delta \geqslant \frac{1}{\|\mathbf{w}\|}$. To maximize δ we need to minimize $\| \mathbf{w} \|$.
We have the optmization problem to maximize $\frac{1}{2} \| \mathbf{w} \|^2$ with the restrictions $y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1$ for $1 \leqslant i \leqslant m$.


C. Cortes and V. N. Vapnik.
Support-vector networks.
Machine Learning Journal, 20:273–297, 1995.


V. N. Vapnik.
Statistical Learning Theory.
Wiley, Chichester, UK, 1998.