Support-Vector Machines - Part 1

Scribes: Wilfredo J. Molina and An Vu

1 Introduction

Support-vector machines (SVMs), also called support-vector networks, are supervised learning methods designed to solve binary classification problems [1,2]. Given a set of labeled training data, each marked as belonging to either one of two categories, an SVM algorithm computes an optimal hyperplane that assigns each new data point to one category or the other, making it a non-probabilistic binary classifier. The optimal criterion for the hyperplane consists in determining the hyperplane achieving the widest possible gap between the two categories. In addition to performing linear classification, SVMs can also perform a non-linear classification using what is called the *kernel trick* – a method that implicitly maps input data into an appropriate feature space where feature vectors are linearly separable.

We start by considering the linear SVM case.

2 Linear SVM

Let $\mathscr{H}$ be a Hilbert space over $\mathbb{R}$ and $X = \{x_1, x_2, \ldots, x_n\} \subset \mathscr{H}$. In the simplest case, one can consider $\mathscr{H} = \mathbb{R}^d$.

We consider the binary classification problem where $x_i \in X$ belongs to either of two classes with labels $y_i \in \{-1, +1\}$.

Observe that any hyperplane in $\mathscr{H}$ has the form

$$H_{w,b} := \{x \in \mathscr{H} : \langle w, x \rangle + b = 0\}$$

Published: February 5, 2020

where $w \in \mathscr{H}$ and $b \in \mathbb{R}$. Geometrically, the vector $\frac{w}{\|w\|}$ can be identified with the unit normal vector to the hyperplane and b with the offset or distance of the hyperplane from the origin (cf. 1).

Accordingly, we can define a *decision function* $f(x) = \operatorname{sgn}(\langle w,x\rangle + b)$ that takes values in the set $\{-1,+1\}$ depending on x falling on either side of the hyperplane $H_{w,b}$.

To develop the theory of SVMs, we start by assuming that the points $X \in \mathscr{H}$ are linearly separable, that is, there exists an hyperplane $H_{w,b}$ separating the two classes.

Definition 1. Let $D = \{(x_i,y_i) \subset \mathscr{H} \times \{-1,+1\} : i = 1,\dots,n\}$. The set D is said to be *linearly separable* if there exist $w \in \mathscr{H}$ and $b \in \mathbb{R}$ such that

$$y_i\,(\langle w,x\rangle + b) > \delta \quad \forall i = 1,\dots n,$$

for some $\delta > 0$. In this case, $H_{w,b}$ is said to be a *separating hyperplane*.

Clearly, if a set of points is linearly separable, there are in general multiple separating hyperplanes. Among those hyperplanes, the SVM approach seeks to find the one with the maximum margin of separation between any (training) point and the hyperplane.

Definition 2. The *optimal separating hyperplane* for a set $D = \{(x_i,y_i) \subset \mathscr{H} \times \{-1,+1\} : i = 1,\dots,n\}$ is the solution of:

$$\max_{w\in\mathscr{H},b\in\mathbb{R}} \min\{\|x_i - x\| : x \in \mathscr{H}, \langle w,x\rangle + b = 0, \text{ and } i = 1,2,\dots,n\} \tag{1}$$

It is easy to see that any hyperplane $H_{w,b}$ can be rescaled by multiplying w and b by the same non-zero constant λ so that $H_{w,b} = H_{\lambda w,\lambda b}$ for all nonzero. We can remove this unnecessary degree of freedom. by rescaling w and b so that the point(s) closest to the hyperplane satisfy $|\langle w,x_i\rangle + b| = 1$.

Definition 3. The hyperplane $H_{w,b}$ is said to be in *canonical form* with respect to $X = \{x_i \in \mathscr{H} : i = 1,\dots,n\}$ if $\min_i |\langle w,x_i\rangle + b| = 1$.

If $H_{w,b}$ is a canonical hyperplane, then a vector in $x_i \in X$ is said to be a *support vector* if it belongs to either one of the hyperplanes H_{-1} or H_1, where $H_k := \{x \in \mathscr{H} : \langle w,x\rangle + b = k\}$.

If $x \in H_{-1}$ and $x' \in H_1$, then

$$\begin{aligned}
2 &= \left|\langle w,x\rangle + b - \left(\langle w,x'\rangle + b\right)\right| \\
2 &= \left|\langle w,x\rangle - \langle w,x'\rangle\right| \\
2 &= \left|\langle w,x - x'\rangle\right| \implies \left|\left\langle \frac{w}{\|w\|}, x - x'\right\rangle\right| = \frac{2}{\|w\|}.
\end{aligned}$$

This calculation shows that the distance between H_{-1} and H_1 is $2/\|w\|$ and that, as a consequence, the distance between $H_{w,b}$ and a support vector is $\frac{1}{\|w\|}$.

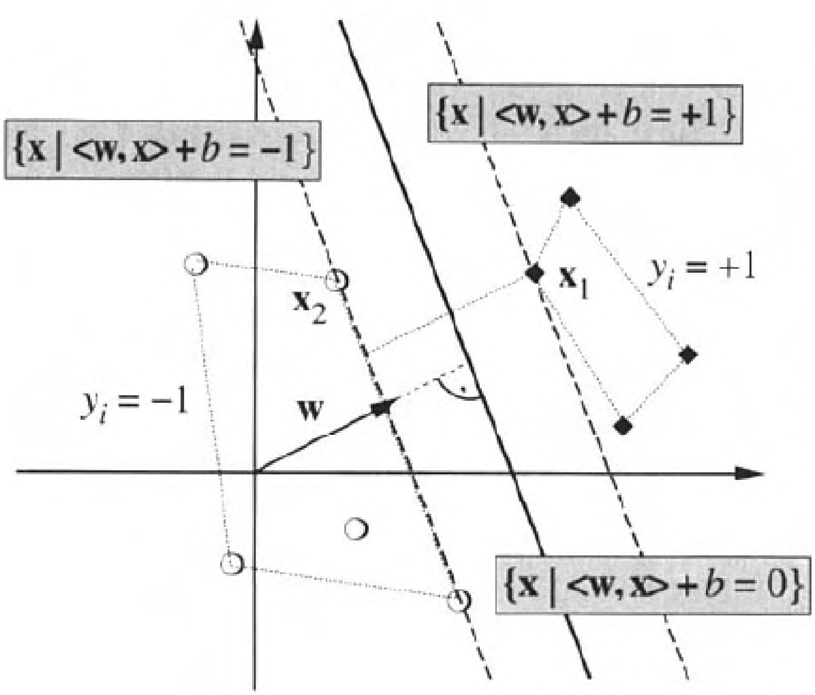

Fig. 1 Binary classification problem. Two set of points, illustrated as balls and diamonds, are separated by an hyperplane illustrated as a solid line. Rescaling the vector w and the constant b, the point(s) x closest to the hyperplane satisfy the condition $|\langle w,x\rangle + b| = 1$. This normalization gives the canonical form of the hyperplane satisfying $y_i(\langle w,x_i\rangle + b) \geq 1$ for all points x_i. The optimal separating hyperplane, as shown in the figure, maximizes the margin, that is, the distance of the closest points to the hyperplane.

Alternatively, using the distance formula from a point to a plane, one can derive that $d(x, H_{w,b}) = \frac{|\langle w,x_i\rangle + b|}{\|w\|}$ and, thus, for a support vector, this distance is $\frac{1}{\|w\|}$.

To construct the optimal separating hyperplane for a set D, we need to find the w and b that *maximize* the margin (1) under the assumption that $y_i(\langle w,x_i\rangle + b) \geq 1$ for all $i = 1,\ldots,n$. By the observation we just made on the size of the margin, this is equivalent to solving the following optimization problem

$$\begin{aligned} &\min_{w \in \mathscr{H}} \tau(w) = \frac{1}{2}\|w\|^2 \\ &\text{subject to:} \quad y_i(\langle w,x_i\rangle + b) \geq 1, \quad \forall i = 1,\ldots,n. \end{aligned} \tag{2}$$

This is a constrained quadratic optimization problem with $n+1$ parameters. Since it is quadratic, it has a single global minimum.

Using the method of Lagrange multipliers, problem (2) can be reformulated. Using the non-negative constants $\alpha_1, \dots, \alpha_n$, we define the Lagrangian

$$L(w,b,\alpha) := \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\langle w, x_i \rangle + b\right) - 1\right).$$

Then problem (2) can be solved as

$$\begin{aligned} &\min_{w \in \mathscr{H}, b \in \mathbb{R}} L(w,b,\alpha) \\ &\text{subject to } \alpha_i \geq 0, \text{ for all } i = 1, \dots n. \end{aligned} \tag{3}$$

The problem (3) above is called the *primal constrained optimization problem*. Here the Lagrangian L has to be minimized with respect to the *primal variables* w and b, while maximized with respect to the *dual variables* α_i. Hence the solution is a saddle point of the multivariate function. Hence, at this saddle point, the derivatives of L with respect to two primal variables must be zero. A direct calculation gives:

$$\begin{aligned} \frac{\partial L}{\partial b} = 0 \quad &\Longrightarrow \quad \sum_i \alpha_i y_i = 0, \\ \frac{\partial L}{\partial w} = 0 \quad &\Longrightarrow \quad w = \sum_i \alpha_i y_i x_i. \end{aligned} \tag{4}$$

Also, $b = y_j - \sum_i \alpha_i y_i \left\langle x_i, x_j \right\rangle$, where j is such that x_j is a support vector.

Eq. (4) shows that the solution w is a linear combination of the support vectors, that is, those vectors x_i for which $\alpha_i > 0$.

By substituting equations (4) into the Lagrangian formulation and eliminating the primal variables w and b, problem (3) is reformulated (and is equivalent) to the following problem *dual optimization problem*:

$$\begin{aligned} &\max_{\alpha_1, \dots \alpha_n} L(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle \\ &\text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \text{ for all } i = 1, \dots, n. \end{aligned} \tag{5}$$

Similarly, using equation (4) the decision function can be written as

$$f(x) = \operatorname{sgn}\left(\sum_i \alpha_i y_i \langle x_i, x \rangle + b\right).$$

We remark is that the advantage of the dual formulation is that the decision function is evaluated in terms of inner products between the input x and the support vectors x_i. This observation is critical for the extension of the SVM approach to nonlinear classification problems.

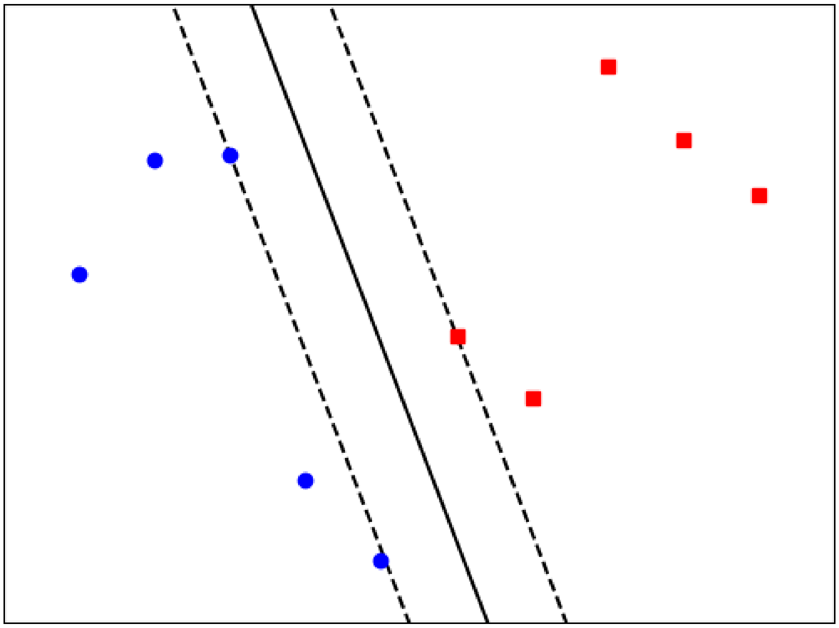

Fig. 2 A set of points in the plane belonging to one of two classes, denoted as blue circles and red squares, is separated by a line using the SVM algorithm (`https://tinyurl.com/lsvmprimal`). The dotted lines are the lines of equation $H_k = \{x \in \mathscr{H} : \langle w,x \rangle + b = k\}$, where $k = \pm 1$, and the points lying on those lines are the support vectors.

2.1 Numerical implementation

See links https://tinyurl.com/lsvmprimal and https://tinyurl.com/lsvmdual for a Python implementation of (3) and (5), respectively.

References

1. Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.
2. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.