Support Vector Machines

Stefanos Baros

Abstract

In this note, we present some insights and explanations surrounding Support Vector Machines, a popular and practically effective method for classification in Machine Learning. In doing so, we discuss its key theoretical aspects, advantages over other Machine Learning methods and practical application. The main results discussed here are from [1] and [2].

1 Hard-SVM

1.1 Basic idea

Support Vector Machines (SVM) are a useful tool for learning high-dimensional spaces. According to the fundamental theorem of learning [1], to learn halfspaces via Empirical Risk Minimization (ERM) requires a training set whose size depends on their VC dimension e.g., $\mathcal{X} \in \mathbb{R}^d$, the VC dimension is $d+1$. SVM on the other hand, require fewer samples in most practical cases to learn halfspaces. This is because SVM are searching for "large margin" separators and by restricting an algorithm to output a "large margin" separator can yield a small sample complexity even if the dimension of the space where $\mathcal{X}$ resides is large. Intuitively, "large margin" separators translate to small norm $\|w\|$ halfspaces. As SVM try to find halfspaces w with minimum norm the bound on the generalization error, which depends on $\|w\|$, will be tighter. Thus, fewer examples are needed with SVM to obtain an ε-accurate halfspace.

1.2 Hard-SVM

In the realizable case, all halfspaces w that perfectly separate data are ERM hypotheses as they result in $L_S(w) = 0$. Now, Hard-SVM tries to find the halfspace that, not only separates the data, but also gives the maximum margin. For a training set, this is defined to be the minimal distance between a point in the training set and the hyperplane. Formally, a distance of a point from the plane is given as follows.

Claim 1 ([1]). *The distance between a point x and the hyperplane, defined by (w, b), where $\|w\| = 1$ is $|\langle w, x\rangle + b|$.*

Given that, the *Hard-SVM problem* can be stated as:

$$\begin{aligned} \underset{w,b:\|w\|=1}{\operatorname{argmax}} \min_{i\in[m]} \quad & |\langle w, x_i\rangle + b| \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b)) > 0, \forall i. \end{aligned} \tag{1}$$

The above problem is equivalent, assuming we are in the separable case, to:

$$\underset{w,b:\|w\|=1}{\operatorname{argmax}} \min_{i\in[m]} y_i(\langle w, x_i\rangle + b). \tag{2}$$

Remark 1. *If one halfspace does not separate samples perfectly there would be at least one i for which $y_i(\langle w, x_i\rangle + b) < 0$. Thus, for this halfspace, the quantity $\min_{i\in[m]} \quad y_i(\langle w, x_i\rangle + b)$, would be negative. As the algorithm tries to find a halfspace w that leads to maximum margin i.e., maximum quantity $\min_{i\in[m]} \quad y_i(\langle w, x_i\rangle + b)$, (realizable case), it will exclude such halfspaces. The algorithm therefore will try to find the optimal halfspace by implicitly searching for w for which $y_i(\langle w, x_i\rangle + b) > 0, \forall i$. Hence, the inequality constraint in Problem (1) can be dropped. Finally, $y_i(\langle w, x_i\rangle + b) = |\langle w, x_i\rangle + b|$ always holds. It follows that Problem (1) and (2) are equivalent.*

1.3 Hard-SVM as a quadratic optimization problem

The Hard-SVM problem can be stated more elegantly as a *quadratic optimization problem* which takes as input a training set $S = (x_1, y_1), ..., (x_m, y_m)$ and yields output the vectors $\hat{w} = \frac{w_0}{\|w_0\|}, \hat{b} = \frac{b_0}{\|w_0\|}$.

$$\begin{aligned} (w_0, b_0) = \underset{w,b}{\operatorname{argmin}} \quad & \|w\|^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b)) \geq 1, \ \forall i. \end{aligned} \tag{3}$$

Problem (3) is equivalent to Problems (1) and (2). To see this, consider a solution $(\hat{w}, \hat{b})$ of Problem (3). For this pair, it holds:

$$y_i(\langle \hat{w}, x_i\rangle + \hat{b})) = \frac{1}{\|w_0\|} y_i(\langle w_0, x_i\rangle + b_0)) \geq \frac{1}{\|w_0\|}. \tag{4}$$

The last inequality follows by taking into account the constraint in Problem (3). One can easily realize that, the solution $(\hat{w}, \hat{b})$ corresponds to the halfspace with the largest possible margin $\gamma = \frac{1}{\|w_0\|}$ as the above optimization problem seeks to find w_0 for which $\|w_0\|$ is the minimum. Further, $\|\hat{w}\| = 1$ and thus follows that $(\hat{w}, \hat{b})$ is the optimal solution of the Problems (1) and (2) as well. With that, we conclude that the three problems are equivalent.

1.4 Sample complexity of Hard-SVM

Here, we make an additional assumption, that the training set is separable with a margin of at least γ. In other words, we lower bound the margin in the training set. Mathematically this can be written as:

$$\max_{(w,b):\|w\|=1} \min_{i\in[m]} y_i(\langle w, x_i\rangle + b) \geq \gamma \tag{5}$$

We now define (γ, ρ)- separability as a property of a given distribution D.

Definition 1 ((γ, ρ)- separability, [1])**.** Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. We say that D is separable with a (γ, ρ)-margin if there exists $(w^\star, b^\star)$ such that $\|w^*\| = 1$ and with probability 1 over the choice $(x, y) \sim D$ we have that $y(\langle w^\star, x\rangle + b^\star \geq \gamma)$ and $\|x\| \leq \rho$.

Intuitively, the above definition can be explained as follows. Any point $(x, y) \sim D$ we pick from D, with probability 1, will result in $y(\langle w^\star, x\rangle + b^\star) \geq \gamma$ and $\|x\| \leq \rho$, for at least one hyperplane $(w^\star, b^\star)$ where $\|w^\star\| = 1$. Therefore, D is separable if there exists a hyperplane so that these conditions hold for a particular distribution D.

The following theorem holds for a distribution D that satisfies the γ, ρ-separability assumption.

Theorem 1 ([1])**.** *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (γ, ρ)- separability with margin assumption using a homogeneous halfspace i.e., $b = 0$. Then, with probability $(1 - \delta)$ over the choice of a training set of size m, the $0 - 1$ error of the output of Hard-SVM is at most*

$$\sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}. \tag{6}$$

The importance of the above result is that, when the separability assumption holds, the sample complexity of Hard-SVM only depends on $(\rho/\gamma)^2$ and is independent of d, the dimension of the Euclidean space where x_i reside. This SVM property will be very valuable when we wish to learn a halfspace efficiently by mapping the training data into a high-dimensional space using Kernel methods.

2 Soft-SVM

2.1 Basic idea

Hard-SVM assumes that the training set is linearly separable. In many practical scenarios, this is not the case and Soft-SVM can be used instead. Soft-SVM is a relaxation of Hard-SVM and can be applied whenever the training set is not linearly separable i.e., when the assumption $y_i(\langle w, x_i\rangle + b) > 0$, $\forall i$, does not hold.

2.2 Soft-SVM

Soft-SVM introduces nonnegative slack variables ξ_i that measure by how much the constraint $y_i(\langle w, x_i\rangle + b) \geq 1$ is being violated in the Hard-SVM Problem (3). It is worthwhile noting here that Soft-SVM not only allows for slight violations but also for misclassifications of the training instances. Mathematically the *Soft-SVM problem* can be stated as follows.

$$\begin{aligned} \min_{w,b,\xi} \quad & (\lambda\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i) \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b)) \geq 1 - \xi_i,\ \forall i, \\ & \xi_i \geq 0. \end{aligned} \tag{7}$$

The Soft-SVM problem jointly minimizes the norm of w, that corresponds to maximizing the margin, and the average of the violations ξ_i. The tradeoff among the two is controlled by the constant λ. The Soft-SVM problem tries to balance between large margin and large violations.

2.3 Soft-SVM as a Regularized Loss Minimization problem

Interestingly, Problem (7) can be recast into a RLM problem. To carry out this, we first eliminate the slack variables ξ_i in the above formulation by relying on the following observations:

- When $1 - y_i(\langle w, x_i\rangle + b)) \leq 0$, instance x_i is labeled correctly by w so the best assignment for ξ_i, given that $\xi_i \geq 1 - y_i(\langle w, x_i\rangle + b))$, is $\xi_i = 0$.
- When $1 - y_i(\langle w, x_i\rangle + b)) > 0$, instance x_i has smaller than desired margin so the best assignment for ξ_i, given that $\xi_i \geq 1 - y_i(\langle w, x_i\rangle + b))$, is $\xi_i = 1 - y_i(\langle w, x_i\rangle + b))$.

From this analysis, it is evident that ξ_i results in the same output as the *hinge loss* function. It follows that:

$$\xi_i = \ell^{hinge}\big((w,b),(x_i,y_i)\big) = \max\{0, 1 - y_i(\langle w, x_i\rangle + b))\}. \tag{8}$$

We can then restate Problem (7) as:

$$\min_{w,b} \quad \lambda\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\ell^{hinge}\big((w,b),(x_i,y_i)\big), \tag{9}$$

or more compactly, as:

$$\min_{w,b} \quad \lambda\|w\|^2 + L_S^{hinge}(w,b). \tag{10}$$

The Soft-SVM problem can therefore be formulated as a Regularized Loss Minimization problem that involves the hinge loss function. In general, one should remember that, Soft-SVM has a bias towards low-norm separators so it tries to find the maximum possible margin while minimizing the total distance of violated instances from the margin.

2.4 Sample Complexity of Soft-SVM

By reducing the Soft-SVM problem to an RLM problem enables us to leverage all the theoretical machinery developed for RLM. In our case, we know that the hinge loss function is convex and $\|x\|$-Lipschitz. Letting $\|x\| \leq \rho$ and employing the main RLM result gives us:

$$\mathbb{E}_{S \sim D^m}[L_D^{hinge}(A(S))] \leq L_D^{hinge}(u) + \lambda \|u\|^2 + \frac{2\rho^2}{\lambda m}. \tag{11}$$

Further, letting $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$, the halfspaces have bounded norm $\|u\| \leq B$ and recalling that, the hinge loss is a *surrogate convex function* for the $0-1$ loss function, allows us to arrive at:

$$\mathbb{E}_{S \sim D^m}[L_D^{0-1}(A(S))] \leq \mathbb{E}_{S \sim D^m}[L_D^{hinge}(A(S))] \leq \min_{w:\|w\| \leq B} L_D^{hinge}(w) + \sqrt{\frac{8\rho^2 B^2}{\lambda m}}. \tag{12}$$

We now state a few remarks associated with this result.

Remark 2. *We can control the sample complexity of learning a halfspace as a function of the norm of that halfspace B and the norm of x, independently of the Euclidean dimension of x. So we don't really care whether $x \in \mathbb{R}^2, \mathbb{R}^3, ..., \in \mathbb{R}^n$. This is one of the key advantages of the SVM method.*

Remark 3. *With SVM, we introduce a bias — we prefer large-margin halfspaces. This can decrease our estimation error, the second term in* (12) *which now depends on B, ρ, λ, m instead of the VC dimension d. This implies that we can also control the generalization error variance, the difference between the generalization L_D and empirical error L_S. The downside of this, is that this bias might increase the approximation error, the first term in* (12)*, which is given with respect to the hinge loss. This is because we restrict our hypothesis class by searching for halfspaces with small norm $\|w\|$.*

3 Duality

Many of the properties of SVM are obtained by considering the dual function of the SVM problem. Further, the name "Support Vector Machine" stems from the fact that the solution of the Hard-SVM problem w_0 is supported by i.e., is in the linear space of, the examples that are exactly at distance $\frac{1}{\|w_0\|}$ from the separating plane.

We now derive the dual problem for the Hard-SVM problem in order to reveal some of these properties. The *primal problem* can be stated as:

$$\begin{aligned} \min_{w} \quad & \frac{\|w\|^2}{2} \\ \text{s.t.} \quad & y_i(\langle w, x_i + b \rangle) \geq 1, \ \forall i. \end{aligned} \tag{13}$$

To formulate the *dual problem*, we construct the *Lagrangian* as:

$$\mathcal{L}(w,b,a) = \frac{\|w\|^2}{2} + \sum_{i=1}^{m} a_i(1 - y_i(\langle w, x_i\rangle + b)). \tag{14}$$

The *dual function* can be computed as [3]:

$$g(a) = \inf_{w} \mathcal{L}(w,b,a) = \inf_{w}(\frac{\|w\|^2}{2} + \sum_{i=1}^{m} a_i(1 - y_i(\langle w, x_i\rangle + b))). \tag{15}$$

The *KKT conditions* for this problem can be computed as:

$$\nabla_w \mathcal{L} = 0 \Rightarrow w = \sum_{i=1}^{m} y_i x_i a_i = 0, \tag{16}$$

$$\nabla_b \mathcal{L} = 0 \Rightarrow \sum_{i=1}^{m} y_i a_i = 0, \tag{17}$$

$$a_i(1 - y_i(\langle w, x_i\rangle + b))), \quad \forall i. \tag{18}$$

It is evident from (16) that the solution w is lies in the linear subspace defined by the instance vectors, the “support vectors”, x_i. To obtain the dual problem we plug w derived above into the Lagrangian (14). That yields:

$$g(a) = \sum_{j=1}^{m} a_j - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \langle x_i, x_j\rangle, \tag{19}$$

and the *dual problem*:

$$\begin{aligned} \max_{a} \quad & \sum_{j=1}^{m} a_j - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \langle x_i, x_j\rangle, \\ \text{s.t.} \quad & \forall i,\ a_i \geq 0,\ \sum_{i=1}^{m} y_i a_i = 0. \end{aligned} \tag{20}$$

It is important to note here that strong duality holds and the primal problem (13) and the dual problem (20) are equivalent. The dual problem only involves the inner product among the instances x_i and not the vectors themselves. This will be very useful consider SVM with Kernel methods.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[3] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.