

Support Vector Machines

The Support Vector Machine (SVM) is an algorithm for learning linear classifiers. Given a linearly separable training set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$, SVM outputs the linear classifier corresponding to the unique solution $\mathbf{w}^* \in \mathbb{R}^d$ of the following convex optimization problem with linear constraints

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 \quad t = 1, \dots, m. \end{aligned} \tag{1}$$

Geometrically, \mathbf{w}^* corresponds to the **maximum margin separating hyperplane**. For every linearly separable set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$, the maximum margin is defined by

$$\gamma^* = \max_{\mathbf{u}: \|\mathbf{u}\|=1} \min_{t=1, \dots, m} y_t \mathbf{u}^\top \mathbf{x}_t$$

and the vector \mathbf{u}^* achieving the maximum margin is the maximum margin separator.

Theorem 1. *For every linearly separable set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$, the maximum margin separator \mathbf{u}^* satisfies $\mathbf{u}^* = \gamma^* \mathbf{w}^*$, where \mathbf{w}^* is the unique solution of (1).*

PROOF. Note that \mathbf{u}^* is the solution of the following optimization problem

$$\begin{aligned} \max_{\mathbf{u} \in \mathbb{R}^d, \gamma > 0} \quad & \gamma^2 \\ \text{s.t.} \quad & \|\mathbf{u}\|^2 = 1 \\ & y_t \mathbf{u}^\top \mathbf{x}_t \geq \gamma \quad t = 1, \dots, m. \end{aligned}$$

Indeed, \mathbf{u} maximizing the margin γ is the same \mathbf{u} maximizing γ^2 because the function $f(\gamma) = \gamma^2$, is monotone for $\gamma > 0$. Dividing by $\gamma > 0$ both sides of each constraint $y_t \mathbf{u}^\top \mathbf{x}_t \geq \gamma$, we obtain the equivalent constraint $y_t (\mathbf{u}^\top \mathbf{x}_t) / \gamma \geq 1$. Introducing $\mathbf{w} = \mathbf{u} / \gamma$, and noting that $\|\mathbf{w}\|^2 = 1 / \gamma^2$ because of the constraint $\|\mathbf{u}\|^2 = 1$, we obtain the equivalent problem

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d, \gamma > 0} \quad & \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \gamma^2 \|\mathbf{w}\|^2 = 1 \\ & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 \quad t = 1, \dots, m. \end{aligned}$$

Now observe that the constraint $\gamma^2 \|\mathbf{w}\|^2 = 1$ is redundant and can be eliminated. Indeed, for all $\mathbf{w} \in \mathbb{R}^d$ we can find $\gamma > 0$ such that the constraint is satisfied. Multiplying the objective function by $\frac{1}{2}$, we obtain

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 \quad t = 1, \dots, m \end{aligned}$$

concluding the proof. □

We have thus shown the equivalence between the problem of maximizing the margin of \mathbf{u} while keeping the norm $\|\mathbf{u}\|$ constant, and the problem of minimizing the norm $\|\mathbf{w}\|$ while keeping the margin of \mathbf{w} constant.

The following result helps us compute the form of the optimal solution \mathbf{w}^* .

Lemma 2 (Fritz John optimality condition). *Consider the problem*

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & f(\mathbf{w}) \\ \text{s.t.} \quad & g_t(\mathbf{w}) \leq 0 \quad t = 1, \dots, m \end{aligned}$$

where the functions f, g_1, \dots, g_m are all differentiable. If \mathbf{w}_0 is an optimal solution, then there exists a nonnegative vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that

$$\nabla f(\mathbf{w}_0) + \sum_{t \in I} \alpha_t \nabla g_t(\mathbf{w}_0) = \mathbf{0}$$

where $I = \{1 \leq t \leq m : g_t(\mathbf{w}_0) = 0\}$.

By applying the Fritz John optimality condition to the SVM objective with $f(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2$ and $g_t(\mathbf{w}) = 1 - y_t \mathbf{w}^\top \mathbf{x}_t$ we obtain

$$\mathbf{w}^* - \sum_{t \in I} \alpha_t y_t \mathbf{x}_t = \mathbf{0} .$$

Hence, the optimal solution has form

$$\mathbf{w}^* = \sum_{t \in I} \alpha_t y_t \mathbf{x}_t$$

where I denotes the set of training examples (\mathbf{x}_t, y_t) such that $y_t (\mathbf{w}^*)^\top \mathbf{x}_t = 1$. These \mathbf{x}_t are called **support vectors**, and are all those training points for which the margin of \mathbf{w}^* is exactly 1. If we removed all training examples except for the support vectors, the SVM solution would not change.

We now move on to consider the case of a training set that is not linearly separable. How should we change the SVM objective? Consider the following formulation

$$\begin{aligned} \min_{(\mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R}^{d+m}} \quad & \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{t=1}^m \xi_t \\ \text{s.t.} \quad & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 - \xi_t \quad t = 1, \dots, m \\ & \xi_t \geq 0 \quad t = 1, \dots, m. \end{aligned}$$

The components of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)$ are called **slack variables** and measure how much each margin constraint is violated by a potential solution \mathbf{w} . The average of these violations is then added to the objective function. Finally, a regularization parameter $\lambda > 0$ is introduced to balance the two terms.

We now consider the constraints involving the slack variables ξ_t . That is, $\xi_t \geq 1 - y_t \mathbf{w}^\top \mathbf{x}_t$ and $\xi_t \geq 0$. In order to minimize each ξ_t , we can set

$$\xi_t = \begin{cases} 1 - y_t \mathbf{w}^\top \mathbf{x}_t & \text{if } y_t \mathbf{w}^\top \mathbf{x}_t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

To see this, fix $\mathbf{w} \in \mathbb{R}^d$. If the constraint $y_t \mathbf{w}^\top \mathbf{x}_t \geq 1$ is satisfied by \mathbf{w} , then ξ_t can be set to zero. Otherwise, if the constraint is not satisfied by \mathbf{w} , then we set ξ_t to the smallest value such that the constraint becomes satisfied, namely $1 - y_t \mathbf{w}^\top \mathbf{x}_t$. Summarizing, $\xi_t = [1 - y_t \mathbf{w}^\top \mathbf{x}_t]_+$, which is exactly the hinge loss $h_t(\mathbf{w})$ of \mathbf{w} .

The SVM problem can then be re-formulated as $\min_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w})$, where

$$F(\mathbf{w}) = \frac{1}{m} \sum_{t=1}^m h_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 .$$

We now show that, even when the training set is not linearly separable, the solution \mathbf{w}^* belongs to the subspace defined by linear combinations of training points multiplied by their labels.

Theorem 3. *The minimizer \mathbf{w}^* of F can be written as a linear combination of $y_1 \mathbf{x}_1, \dots, y_m \mathbf{x}_m$.*

PROOF. By contradiction, assume

$$\mathbf{w}^* = \sum_{t=1}^m \alpha_t y_t \mathbf{x}_t + \mathbf{u} \tag{2}$$

where $\mathbf{u} \in \mathbb{R}^d$ is the component of \mathbf{w}^* orthogonal to the subspace spanned by $\mathbf{x}_1, \dots, \mathbf{x}_m$. Therefore,

$$y_t \mathbf{u}^\top \mathbf{x}_t = 0 \quad t = 1, \dots, m. \tag{3}$$

Now, let $\mathbf{v} = \mathbf{w}^* - \mathbf{u}$. First, $\|\mathbf{v}\|^2 \leq \|\mathbf{w}^*\|^2$ because in (2) we wrote \mathbf{w}^* as a sum of two orthogonal components and we removed one of them, and so its length decreased. Second,

$$h_t(\mathbf{v}) = [1 - y_t \mathbf{v}^\top \mathbf{x}_t]_+ = [1 - y_t (\mathbf{w}^* - \mathbf{u})^\top \mathbf{x}_t]_+ = [1 - y_t (\mathbf{w}^*)^\top \mathbf{x}_t + y_t \mathbf{u}^\top \mathbf{x}_t]_+ = h_t(\mathbf{w}^*)$$

using (3). Therefore $F(\mathbf{v}) \leq F(\mathbf{w}^*)$, contradicting the optimality of \mathbf{w}^* . Hence $\mathbf{u} = \mathbf{0}$ and the proof is concluded. \square

Note that, as in the linearly separable case, \mathbf{w}^* generally depends on a subset of support vectors. However, unlike the linearly separable case, these support vectors also include the training points associated with positive slack variables.

We proceed by showing how F can be minimized using Online Gradient Descent (OGD). First, observe that

$$F(\mathbf{w}) = \frac{1}{m} \sum_{t=1}^m \ell_t(\mathbf{w})$$

where $\ell_t(\mathbf{w}) = h_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2$ is a strongly convex function. Indeed, $\frac{\lambda}{2} \|\mathbf{w}\|^2$ is λ -strongly convex, and h_t is convex (and also piecewise linear). This implies that their sum is λ -strongly convex. We can then apply the OGD algorithm for strongly convex functions to the set of losses ℓ_1, \dots, ℓ_m . This instance of OGD, which is known as **Pegasos**, can be described as follows.

Parameters: number T of rounds, regularization coefficient $\lambda > 0$

Input: Training set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$

Set $\mathbf{w}_1 = \mathbf{0}$

For $t = 1, \dots, T$

1. Draw uniformly at random an element $(\mathbf{x}_{Z_t}, y_{Z_t})$ from the training set

2. Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell_{Z_t}(\mathbf{w}_t)$

Output: $\bar{\mathbf{w}} = \frac{1}{T}(\mathbf{w}_1 + \dots + \mathbf{w}_T)$.

We now move on to analyze Pegasos. Let $(\mathbf{x}_{Z_1}, y_{Z_1}), \dots, (\mathbf{x}_{Z_T}, y_{Z_T})$ the sequence of training set examples that were drawn at random in step 1 of the algorithm, and let $\ell_{Z_1}, \dots, \ell_{Z_T}$ the corresponding sequence of loss functions. Namely, $\ell_{Z_t}(\mathbf{w}) = h_{Z_t}(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2$ where $h_{Z_t}(\mathbf{w}) = [1 - y_{Z_t} \mathbf{w}^\top \mathbf{x}_{Z_t}]_+$.

Let \mathbf{w}^* be the optimal SVM solution,

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} \left(\frac{1}{m} \sum_{t=1}^m h_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right). \quad (4)$$

For every realization s_1, \dots, s_T of the random variables Z_1, \dots, Z_T , OGD analysis for strongly convex losses immediately gives

$$\frac{1}{T} \sum_{t=1}^T \ell_{s_t}(\mathbf{w}_t) \leq \frac{1}{T} \sum_{t=1}^T \ell_{s_t}(\mathbf{w}^*) + \frac{G^2}{2\lambda T} \ln(T+1) \quad (5)$$

where $G = \max_{t=1, \dots, T} \|\nabla \ell_{s_t}(\mathbf{w}_t)\|$ is also a random variable.

In order to show how this result can be used to bound $F(\bar{\mathbf{w}})$, we use the following fact

$$\mathbb{E}[\ell_{Z_t}(\mathbf{w}_t) \mid Z_1, \dots, Z_{t-1}] = \frac{1}{m} \sum_{s=1}^m \ell_s(\mathbf{w}_t) = F(\mathbf{w}_t). \quad (6)$$

In other words, conditioned on the first $t-1$ random draws (which determine \mathbf{w}_t), the expected value of $\ell_{Z_t}(\mathbf{w}_t)$ is equal to $F(\mathbf{w}_t)$. We also use the fact that for every pair of random variables

X, Y the following holds $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]]$. Hence, we can write

$$\begin{aligned}
\mathbb{E}[F(\bar{\mathbf{w}})] &= \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^T \mathbf{w}_t\right)\right] \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T F(\mathbf{w}_t)\right] \quad \text{using Jensen inequality, since } F \text{ is convex} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\ell_{Z_t}(\mathbf{w}_t) | Z_1, \dots, Z_{t-1}]\right] \quad \text{using (6)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{Z_t}(\mathbf{w}_t)\right] \quad \text{using } \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]] \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{Z_t}(\mathbf{w}^*)\right] + \frac{\mathbb{E}[G^2]}{2\lambda T}(\ln T + 1) \quad \text{using (5)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\ell_{Z_t}(\mathbf{w}^*) | Z_1, \dots, Z_{t-1}]\right] + \frac{\mathbb{E}[G^2]}{2\lambda T}(\ln T + 1) \quad \text{using } \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]] \\
&= F(\mathbf{w}^*) + \frac{\mathbb{E}[G^2]}{2\lambda T} \ln(T + 1) \quad \text{using (6)}.
\end{aligned}$$

We thus obtained

$$\mathbb{E}[F(\bar{\mathbf{w}})] \leq F(\mathbf{w}^*) + \frac{\mathbb{E}[G^2]}{2\lambda T}(\ln T + 1). \quad (7)$$

Therefore, if $\mathbb{E}[G^2]$ can be upper bounded by a constant, the average $\bar{\mathbf{w}}$ of the vectors generated by OGD converges (in expectation with respect to the random draw of the elements from the training set) to \mathbf{w}^* with rate $\frac{\ln T}{T}$. With a bit more work, one can show that $\bar{\mathbf{w}}$ converges to \mathbf{w}^* not only in expectation but also in probability.

We now bound G for every realization s_1, \dots, s_T of the random variables Z_1, \dots, Z_T . We have $\nabla \ell_{s_t}(\mathbf{w}_t) = -y_{s_t} \mathbf{x}_{s_t} \mathbb{I}\{h_{s_t}(\mathbf{w}_t) > 0\} + \lambda \mathbf{w}_t$. Let $\mathbf{v}_t = y_{s_t} \mathbf{x}_{s_t} \mathbb{I}\{h_{s_t}(\mathbf{w}_t) > 0\}$. Because $\eta_t = 1/(\lambda t)$, the update rule for \mathbf{w}_t takes the following simple form,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell_t(\mathbf{w}_t) = \mathbf{w}_t + \eta_t \mathbf{v}_t - \eta_t \lambda \mathbf{w}_t = \left(1 - \frac{1}{t}\right) \mathbf{w}_t + \frac{1}{\lambda t} \mathbf{v}_t.$$

Let $X = \max_{s=1, \dots, m} \|\mathbf{x}_s\|$. Since $\|\nabla \ell_{s_t}(\mathbf{w}_t)\| \leq \|\mathbf{v}_t\| + \lambda \|\mathbf{w}_t\| \leq X + \lambda \|\mathbf{w}_t\|$, we are left with the task of computing an upper bound for $\|\mathbf{w}_t\|$. In order to do so, we look at the recurrence

$$\mathbf{w}_{t+1} = \left(1 - \frac{1}{t}\right) \mathbf{w}_t + \frac{1}{\lambda t} \mathbf{v}_t.$$

As one can easily show by induction, \mathbf{w}_{t+1} can be written as a linear combination of $\mathbf{v}_1, \dots, \mathbf{v}_t$. In order to determine the coefficients of this linear combination, we fix $s \leq t$ and observe that \mathbf{v}_s is added to the sum with coefficient $1/(\lambda s)$. When \mathbf{w}_{t+1} is computed, the coefficient of \mathbf{v}_s has become

$$\frac{1}{\lambda s} \prod_{r=s+1}^t \left(1 - \frac{1}{r}\right) = \frac{1}{\lambda s} \prod_{r=s+1}^t \frac{r-1}{r} = \frac{1}{\lambda t}.$$

We thus obtain a simple expression for \mathbf{w}_{t+1} ,

$$\mathbf{w}_{t+1} = \frac{1}{\lambda t} \sum_{s=1}^t \mathbf{v}_s . \quad (8)$$

Because \mathbf{w}_{t+1} is an average of \mathbf{v}_s divided by λ , we finally have $\|\mathbf{w}_{t+1}\| \leq \frac{1}{\lambda} \max_s \|\mathbf{v}_s\| \leq \frac{1}{\lambda} X$. This allows us to conclude that $\|\nabla \ell_t(\mathbf{w}_t)\| \leq X + \lambda \|\mathbf{w}_t\| \leq 2X$. Substituting this bound for G in (7) we get

$$\mathbb{E}[F(\bar{\mathbf{w}})] \leq F(\mathbf{w}^*) + \frac{2X^2}{\lambda T} \ln(T+1) .$$

Theorem 3 states that the solution \mathbf{w}^* to the SVM problem can be written as

$$\mathbf{w}^* = \sum_{s \in S} y_s \alpha_s \mathbf{x}_s$$

where $\alpha_s > 0$ and $S \equiv \{t = 1, \dots, m : h_t(\mathbf{w}^*) > 0\}$. An important consequence of this result is that we can solve the problem (4) in a RKHS \mathcal{H}_K , where the objective function F becomes

$$F_K(g) = \frac{1}{m} \sum_{t=1}^m h_t(g) + \frac{\lambda}{2} \|g\|_K^2 \quad g \in \mathcal{H}_K$$

with $h_t(g) = [1 - y_t g(\mathbf{x}_t)]_+$. In \mathcal{H}_K , the SVM solution can therefore be written as

$$\sum_{s \in S} y_s \alpha_s K(\mathbf{x}_s, \cdot)$$

which is clearly an element of the RKHS

$$\mathcal{H}_K \equiv \left\{ \sum_{i=1}^N \alpha_i K(\mathbf{x}_i, \cdot) : \mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d, \alpha_1, \dots, \alpha_N \in \mathbb{R}, N \in \mathbb{N} \right\}$$

As we did for the Perceptron, we can run Pegasos in the RKHS \mathcal{H}_K . The predictors generated by kernel Pegasos are of the form

$$g_{t+1} = \frac{1}{\lambda t} \sum_{r=1}^t f_r$$

with $f_r = y_{s_r} K(\mathbf{x}_{s_r}, \cdot) \mathbb{I}\{h_{s_r}(g_r) > 0\}$, where s_1, \dots, s_t is the realized sequence of examples drawn from the training set.