# Support Vector Machines

Calvin Chi

July 21, 2020

## Contents

## 1 Introduction

The following is a note on support vector machines (SVM) that blends perspectives and derivations introduced by Andrew Ng's Stanford CS229 course on machine learning and Laurent El Ghaoui's UC Berkeley EE127 course on convex optimization [1]. Briefly, SVM is a supervised learning method that fits a separating hyperplane to discrimiante between samples of different classes such that the gap between the closest data points and the hyperplane is large. This is depicted in Figure 2.

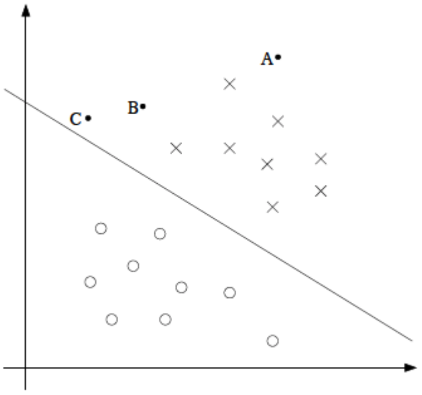

**Figure 1:** Fitting a hard-margin SVM.

This note emphasizes on arriving at the basic ideas behind SVM from ideas that are intuitive and obvious, but is not meant to be comprehensive. Where appropriate, links to relevant resources to cover the rest of SVM will be provided.

# 2 Hard-margin SVM

## 2.1 Primal problem

A hard-margin SVM fits a separating hyperplane assuming the classes are linearly separable with some hyperplane $w^\top x + b = 0$. Suppose such a hyperplane defined by $w, b$ exists. Let $\gamma^{(i)}$ be the distance between samples $x^{(i)} \in \mathbb{R}^n$ and the hyperplane/decision boundary. A maximal margin decision boundary achieves maximal $\gamma = \min_{i=1,\dots,m} |\gamma^{(i)}|$, which is equivalent to stating $|\gamma^{(i)}| \geq \gamma$ for all $i = 1, \dots, m$.

What is $\gamma^{(i)}$ as a function of $w, b, x^{(i)}$? Let $x'$ be a point on the decision boundary such that the line segment between $x^{(i)}$ and $x'$ is perpendicular to the boundary, and hence has length that is the shortest distance between $x^{(i)}$ and $x'$. Then $x' = x^{(i)} - \gamma^{(i)} \frac{w}{||w||_2}$, satisfying

$$w^\top x' + b = 0 \Rightarrow w^\top \left( x^{(i)} - \gamma^{(i)} \frac{w}{||w||_2} \right) + b = 0 \Rightarrow \gamma^{(i)} = \left( \frac{w}{||w||_2} \right)^\top x^{(i)} + \frac{b}{||w||_2}$$

Depending on which side of the boundary $x^{(i)}$ is on, $\gamma^{(i)}$ can be either positive or negative. Out of mathematical convenience, if we define $y^{(i)} \in \{-1, +1\}$ such that $y^{(i)} = +1$ if $w^\top x^{(i)} + b > 0$ and $y^{(i)} = -1$ otherwise, we can express

$$\gamma = y^{(i)} \left( \left( \frac{w}{||w||_2} \right)^\top x^{(i)} + \frac{b}{||w||_2} \right)$$

If we additionally impose the constraint $||w||_2 = 1$, then the problem of maximizing the minimum distance $\gamma$ can be stated as

$$\begin{aligned} &\max_{\gamma w, b} \quad \gamma \\ &y^{(i)}(w^\top x^{(i)} + b) \geq \gamma \quad i = 1, \dots, m \\ &||w||_2 = 1 \end{aligned}$$

However, the constraint that $||w||_2 = 1$ makes the problem nonconvex because the set of feasible values of $w$ is not convex[1]. If we rewrite the problem without explicitly stating the $||w||_2 = 1$ constraint by "absorbing" it into $\hat{\gamma}^{(i)} = \gamma^{(i)} ||w||_2 \Leftrightarrow \gamma^{(i)} = \hat{\gamma}^{(i)} / ||w||_2$. Then after defining $\hat{\gamma} = \min_{i=1,\dots,m} \hat{\gamma}^{(i)}$, we can rewrite the problem as

$$\begin{aligned} &\max_{\hat{\gamma}, w, b} \quad \frac{\hat{\gamma}}{||w||_2} \\ &y^{(i)}(w^\top x^{(i)} + b) \geq \hat{\gamma} \quad i = 1, \dots, m \end{aligned}$$

It turns out that $\hat{\gamma}$ can be set to any value via scaling (i.e. $c\hat{\gamma}$) without changing the prediction rule that $\hat{y}^{(i)} = \text{sgn}(w^\top x^{(i)} + b)$, since $\text{sgn}(w^\top x^{(i)} + b) = \text{sgn}(cw^\top x^{(i)} + cb)$. Thus, we can multiply both sides of the inequality constraint by a constant such that $w, b$ are redefined by a scaling constant and $\hat{\gamma} = 1$ after scaling. This rescaling also amounts to multiplying both the numerator and denominator of the objective function by the constant. Then the problem transforms into

$$\begin{aligned} &\max_{w, b} \quad \frac{1}{||w||_2} \\ &y^{(i)}(w^\top x^{(i)} + b) \geq 1 \quad i = 1, \dots, m \end{aligned}$$

after setting $w := cw$ and $b := cb$. The problem is equivalent to

[1] This is not to be confused with the set $||w||_2 \leq 1$, which is a convex set

$$
\begin{aligned}
&\min_{w,b} \quad \frac{1}{2}||w||_2^2 \\
&y^{(i)}(w^\top x^{(i)} + b) \geq 1 \quad i = 1, \ldots, m
\end{aligned}
$$

We refer to this problem formulation directly from the problem definition as the primal problem. The primal problem reduces to a quadratic program and could be solved with a quadratic solver.

## 2.2 Geometry of primal objective

There is a geometric interpretation to the primal optimization problem that leads to the idea of a margin in SVM. Points $x^{(i)}$ fulfilling the inequality constraint with equality either satisfy $w^\top x^{(i)} + b = 1$ or $w^\top x^{(i)} + b = -1$ for a given pair $w, b$. Points satisfying $w^\top x^{(i)} + b = \pm 1$ are said to fall on the margin of the SVM, with the decision boundary lying in the middle in between.

Recall that the choice of $\pm 1$ is not necessary but is a convention. With $w, b$ defining a hyperplane such that the closest points on either side are equidistant to it and satisfy either $w^\top x^{(i)} + b = c$ or $w^\top x^{(i)} + b = -c$. These equations can equivalently be expressed as $(w/c)^\top x^{(i)} + b/c = 1$ and $(w/c)^\top x^{(i)} + b/c = -1$.

The distance between the two margins $d$ can be found as the projection of any $x_1 - x_0$, where $w^\top x_1 + b = 1$ and $w^\top x_0 + b = -1$, onto $w$. Starting from definition of projection

$$
\begin{aligned}
d &= \frac{w^\top}{||w||_2}(x_1 - x_0) \\
&= \frac{1}{||w||_2}(w^\top x_1 - w^\top x_0) \\
&= \frac{1}{||w||_2}((w^\top x_1 + b) - (w^\top x_0 + b)) \\
&= \frac{2}{||w||_2}
\end{aligned}
$$

Hence, finding $w$ to minimize $||w||_2^2$ maximizes the distance between the margins.

## 2.3 Dual problem

Although the primal problem could be solved with a quadratic program, it turns out that the dual problem naturally leads to the application of kernels that can map the current feature space into a new feature space where classification becomes easier. This is because once the parameters of the dual problem are found, prediction involves the dot product between a test sample with select training samples called support vectors.

To establish that SVM can be implemented using either the primal or dual formulations, we need to first establish that the optimal values of the primal and dual problem are the same ($i.e.p^* = d^*$). Since in the primal problem the inequality constraint involves an affine function of $w, b$ and the objective is convex, we can apply the weak Slater's condition to assert that $p^* = d^*$.
We now start from the Lagrangian function to derive the dual problem formulation.

$$
\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||_2^2 - \sum_{i=1}^{m} \alpha_i (y^{(i)}(w^\top x^{(i)} + b) - 1)
$$

Since $\mathcal{L}(w, b, \alpha)$ is convex in $w, b$, the $w, b$ minimizing $\mathcal{L}(w, b, \alpha)$ can be determined with the first-order derivative condition.

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow w^* = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}$$

$$\nabla_b \mathcal{L}(w,b,\alpha) = \sum_{i=1}^m \alpha_i y^{(i)} = 0$$

where $\alpha_i \geq 0$ for all $i = 1, \dots, m$. Now $\mathcal{L}(\alpha, w^*, b^*)$ is the dual objective, and the dual problem involves solving

$$\begin{aligned} &\max_\alpha \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^\top x^{(j)} \\ &\alpha_i \geq 0, \quad i = 1, \dots, m \\ &\sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned}$$

which is another quadratic program. Once $\alpha^* \in \mathbb{R}^m$ is solved, $w^* \in \mathbb{R}^n$ is solved, then $b^*$ can be found by first considering that

$$\max_{i:y^{(i)}=-1} (w^*)^\top x^{(i)} + b = -1 \qquad \min_{i:y^{(i)}=1} (w^*)^\top x^{(i)} + b = 1$$

Then

$$\begin{aligned} &\max_{i:y^{(i)}=-1} (w^*)^\top x^{(i)} + b + \min_{i:y^{(i)}=1} (w^*)^\top x^{(i)} + b = 0 \\ &\Rightarrow b^* = -\frac{\max_{i:y^{(i)}=-1} (w^*)^\top x^{(i)} + \min_{i:y^{(i)}=1} (w^*)^\top x^{(i)}}{2} \end{aligned}$$

Assume that $w^*, \alpha^*, b^*$ have all been solved. Then to return to how solving the dual problem leads to the natural application of the kernel trick, notice that prediction involves the dot product between $x$ and samples in the training set.

$$w^\top x + b = \sum_{i=1}^m \alpha_i y^{(i)} (x^{(i)})^\top x + b$$

A great explanation for how dot products lead to the application of the kernel trick can be found in note 3 of Stanford's CS229 lecture notes. The fact that $p^* = d^*$ implies that the Karush-Kuhn-Tucker (KKT) conditions are also satisfied, which explains that prediction actually only uses training samples on the margins called "support vectors". Let $g(w,b) = 1 - y^{(i)}(w^\top x^{(i)} + b)$ denote the inequality constraint in our primal problem, then one of the KKT conditions is that $\alpha_i^* g_i(w^*, b^*) = 0$ for all $i = 1, \dots, m$. Thus, after solving the dual problem, if $\alpha_i^* > 0$, then this implies that $g_i(w^*, b^*) = 0 \Rightarrow y^{(i)}(w^\top x^{(i)} + b) = 1$. In other words, $\alpha_i > 0$ only corresponds to $x^{(i)}$ on the SVM margin. So prediction really only involves the dot product between test sample $x$ and support vectors.

## 3 Soft-margin SVM

The hard-margin SVM is impractical for two reasons. One, the assumption that classes are linearly separable is often violated in real-life situations. Two, even if the classes are linearly separable, the hard-margin SVM would be very sensitive to outliers due to having to ensure every sample lies on the correct side of the hyperplane. This is best illustrated in Figure 2.

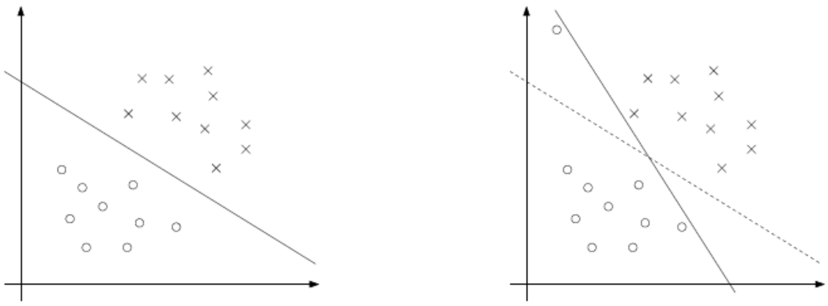

**Figure 2:** Fitting a hard-margin SVM with outliers.

Starting with the primal problem for the separable case

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||_2^2 \\ & y^{(i)}(w^\top x^{(i)} + b) \geq 1 \quad i = 1, \ldots, m \end{aligned}$$

a modification to accommodate the non-separable scenario balances the objectives of maintaining a large margin while allowing misclassification. Introduce $s^{(i)} \geq 0$ as a variable such that if $x^{(i)}$ is misclassified with $y^{(i)} = +1$, then a value can be assigned to $s^{(i)}$ such that $w^\top x_i + b + s^{(i)} = 1$. For misclassified $x_i$ with $y^{(i)} = -1$, $s^{(i)}$ can similarly be assigned a value such that $w^\top x_i + b - s^{(i)} = -1$. The two scenarios are expressed along with $y^{(i)}$ below

$$y^{(i)}(w^\top x^{(i)} + b - s^{(i)}) = 1, \quad y^{(i)}(w^\top x^{(i)} + b + s^{(i)}) = 1$$

which can be rewritten as one equation[2].

$$y^{(i)}(w^\top x^{(i)} + b) = 1 - s^{(i)}$$

To find a hyperplane minimizing misclassification, the quantity $s^{(i)}$ should be minimized. To incorporate $s^{(i)}$ into the original inequality constraint, we allow $s^{(i)}$ to have the freedom to over-correct such that $y^{(i)}(w^\top x^{(i)} + b) \geq 1 - s^{(i)}$. However, since $s^{(i)}$ is a quantity to be minimized, the inequality will be met with equality when $s^{(i)} > 0$.

The optimization problem that balances both objectives of minimizing misclassification error while maximizing the margin now becomes

$$\begin{aligned} \min_{w,b,s} \quad & \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{m} s^{(i)} \\ & y^{(i)}(w^\top x^{(i)} + b) \geq 1 - s^{(i)} \quad i = 1, \ldots, m \\ & s^{(i)} \geq 0 \quad i = 1, \ldots, m \end{aligned}$$

The parameter term $C \in \mathbb{R}$ controls the balance between the two objectives, with a larger $C$ leading to better classification on the training dataset. Note the placement of $C$ with $\sum_i s^{(i)}$ is

[2] Since if $x^{(i)}$ is misclassified such that $w^\top x^{(i)} + b < 0$, then $y^{(i)} = +1$, and $1 - s^{(i)}y^{(i)} = 1 - s^{(i)}$. Otherwise, if $x^{(i)}$ is misclassified such that $w^\top x^{(i)} + b > 0$, then $y^{(i)} = -1$, and $1 + s^{(i)}y^{(i)} = 1 - s^{(i)}$.

more of a convention, since $C$ could be placed with $\frac{1}{2}||w||_2^2$ as well.

Increasing the hyperparameter $C$ reinforces this objective to achieve a low bias, high variance classifier[3]. In contrast, decreasing $C$ increases the relative contribution of $\frac{1}{2}||w||_2^2$ to the total loss, which achieves a high bias, low variance classifier. Just like the hard-margin SVM problem, the soft-margin SVM has a dual problem formulation

$$\begin{aligned}\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^\top x^{(j)}\\ 0 \le \alpha_i \le C, \quad i = 1, \ldots, m\\ \sum_{i=1}^{m} \alpha_i y^{(i)} = 0\end{aligned}$$

whose detailed derivation can be found in Stanford CS229's convex optimization notes.

# 4 Relationship with hinge loss

It turns out that the term $C\sum_i s^{(i)}$ in the primal problem of the soft-margin SVM is related to the hinge loss, which penalizes misclassified samples more as they are further away from the decision boundary. We can build the intuition for the hinge loss by starting with one of the simplest losses for binary classification - the zero-one loss.

$$G(z) = \begin{cases} 1 & \text{if } z < 0 \\ 0 & \text{if } z \ge 0 \end{cases}$$

The zero-one loss is graphically depicted in Figure 3.

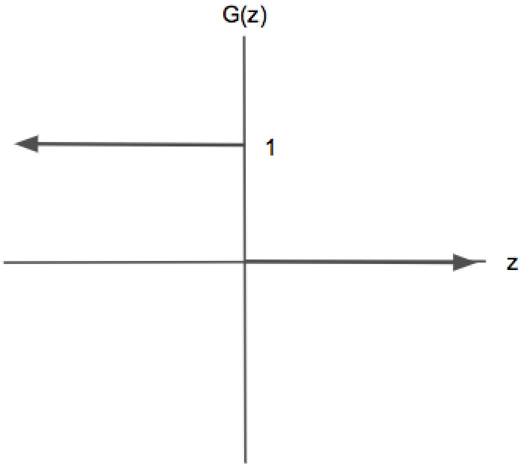

**Figure 3:** 1-0 loss.

By denoting $y \in \{+1, -1\}$, sample $x^{(i)}$ is correctly satisfied if and only if

$$y^{(i)}(w^\top x^{(i)} + b) \ge 0$$

[3] In the sense that the fitted hyperplane is variable across fits to different samples of a population, in the attempt to minimize misclassification.

The zero-one loss for $m$ samples is

$$L(w,b) = \sum_{i=1}^{m} G\left[y^{(i)}(w^\top x^{(i)} + b)\right] = \sum_{i=1}^{m} G(z^{(i)})$$

However, this loss treats all misclassified samples the same, regardless of how far away they are from the hyperplane. Additionally, the loss function $L(w,b)$ is not convex and is hard to optimize[4].

A loss function that penalizes more severely misclassified samples is the hinge loss

$$H(z) = \max(0, 1-z)$$

which is graphically depicted in Figure 4.

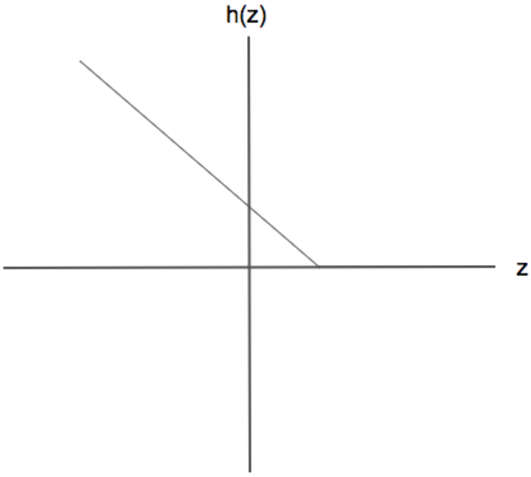

**Figure 4:** Hinge loss.

Thus, the hinge loss for $m$ samples is

$$L(w,b) = \sum_{i=1}^{m} \max(0, 1 - y^{(i)}(w^\top x^{(i)} + b))$$

The objective is convex because the sum of convex functions is convex and the point-wise maximum of convex functions is convex. The term $1 - y^{(i)}(w^\top x^{(i)} + b)$ is the affine map

$$1 - y^{(i)}(w^\top x^{(i)} + b) = 1 - \begin{bmatrix} y_i(x^{(i)})^\top & y_i \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix}$$

Finally, by convexity under convex composition of affine maps, $\max(0, 1 - y^{(i)}(w^\top x^{(i)} + b))$ is a convex function.

To introduce regularization, one can introduce the $\ell 2$ norm to arrive at

$$L(w,b) = \sum_{i=1}^{m} \max(0, 1 - y_i(w^T x_i + b)) + \lambda ||w||_2^2$$

[4] To see why $G(z)$ is not convex, for any point $z_1 < 0$ and $z_2 > 0$, the resulting line segment $\overline{z_1 z_2}$ is not strictly above $G(z)$ for $z \in [z_1, z_2]$, violating the definition of a convex function $\lambda G(z_1) + (1-\lambda)G(z_2) \geq G(\lambda z_1 + (1-\lambda)z_2)$ for $\lambda \in [0,1]$.

which is equivalent to the primal optimization problem of the soft-margin SVM

$$
\begin{aligned}
\min_{w,b,s} \quad & \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{m} s^{(i)} \\
& y^{(i)}(w^\top x^{(i)} + b) \geq 1 - s^{(i)} \quad i = 1, \ldots, m \\
& s^{(i)} \geq 0 \quad i = 1, \ldots, m
\end{aligned}
$$

since minimizing $s^{(i)}$'s under the constraint

$$y^{(i)}(w^\top x^{(i)} + b) \geq 1 - s^{(i)} \Leftrightarrow s^{(i)} \geq 1 - y^{(i)}(w^\top x^{(i)} + b)$$

with non-negativity of $s^{(i)}$ constraint is equivalent to minimizing $\max(0, 1 - y^{(i)}(w^\top x^{(i)} + b))$. Additionally, we can see that introducing the $||w||_2^2$ term as $\ell 2$ penalty to reduce variance and increase bias leads to increasing the margin of the SVM.

# References

[1] A. Ng, "Cs229 lecture notes," *CS229 Lecture notes*, vol. 1, no. 1, pp. 1–3, 2000.