Support Vector Machines

Calvin Chi

August 25, 2020

Contents

1 Introduction

The support vector machine (SVM) learning algorithm finds a separating hyperplane between two classes from a dataset $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ of m samples, where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \{-1, +1\}$. This is depicted in Figure 1.

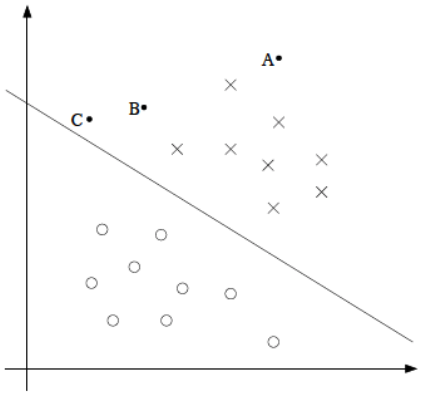

Figure 1: Fitting a separating hyperplane between data points from class "x" and data points from class "o".

In fitting this hyperplane, or decision boundary, the SVM strikes a balance between finding a large margin boundary versus a boundary that is not overly sensitive to outliers. The following note on SVM is based on Andrew Ng's machine learning course at Stanford and Laurent El Ghaoui's convex optimization course at UC Berkeley [1].

2 Hard-margin SVM

2.1 Primal problem

Assuming the two classes are linearly separable, the hard-margin SVM fits a separating hyperplane $w^\top x + b = 0$ by finding parameters $w \in \mathbb{R}^n, b \in \mathbb{R}$. The hard-margin SVM does not just find any separating hyperplane, but a hyperplane with maximum distance, or margin, to the closest data point. Define $\gamma^{(i)}$ to be the distance between sample $x^{(i)} \in \mathbb{R}^n$ and the hyperplane. A maximal margin decision boundary achieves maximal $\gamma = \min_{i=1,\ldots,m} |\gamma^{(i)}|$, which is equivalent to stating $|\gamma^{(i)}| \geq \gamma$ for all $i = 1, \ldots, m$.

We now describe the relationship between the margin $\gamma^{(i)}$ and learnable parameters w, b so that we can specify an optimization problem. Imagine a line segment from $x^{(i)}$ to a decision boundary such that the line segment is perpendicular to the boundary. Let x' be the end point of the line segment that is on the decision boundary. Since $w/||w||_2$ is a unit vector that is perpendicular to the decision boundary, we can describe x' as $x' = x^{(i)} - \gamma^{(i)} \frac{w}{||w||_2}$. Since x' is on the decision boundary, it satisfies $w^\top x' + b = 0$, and this implies

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

Depending on which side of the boundary $x^{(i)}$ is on, $\gamma^{(i)}$ can be either positive or negative. Since $y^{(i)} \in \{-1, +1\}$ and we assumed the two classes are linearly separable, we can assume without loss of generality that data points satisfying $w^\top x^{(i)} + b > 0$ have label $y^{(i)} = +1$ and data points satisfying $w^\top x^{(i)} + b < 0$ have label $y^{(i)} = -1$. Then we can re-express $\gamma^{(i)}$ as

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

The hard-margin SVM problem is to find w, b to maximize γ while satisfying $\gamma^{(i)} \geq \gamma$ for $i = 1, \ldots, m$. If we additionally impose the constraint $||w||_2 = 1$, then the hard-margin optimization problem is stated as

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

However, the constraint that $||w||_2 = 1$ makes the problem nonconvex because the set of feasible values of w is not convex[1]. To address this problem, we define $\hat{\gamma}^{(i)} = \gamma^{(i)} ||w||_2$, so that the constraint $\gamma^{(i)} \geq \gamma$ can be re-expressed as $y^{(i)}(w^\top x^{(i)} + b) \geq \hat{\gamma}^{(i)}$ by multiplying both sides by $||w||_2$. Additionally, the objective can be expressed as $\gamma = \hat{\gamma}/||w||_2$. The transformed optimization problem becomes

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

It turns out that $\hat{\gamma}$ can be set to any value via scaling (i.e. $c\hat{\gamma}$ for $c \in \mathbb{R}^+$) 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 apply the scaling such that $c\hat{\gamma} = 1$ and set $w := cw$ and $b := cb$. This scaling does not change the optimization problem because the scaling amounts to multiplying both sides of the inequality

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

constraints by c, and multiplying the numerator and denominator of the objective by c. After scaling, the problem transforms into

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

The problem is equivalent to

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

We refer to the above optimization problem 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 maximal margin in SVM. Define the margin of a decision boundary to be the two lines on either side of the decision boundary, that are both parallel to the boundary and passes through the points closest to the boundary. The margin idea is illustrated in Figure 2.

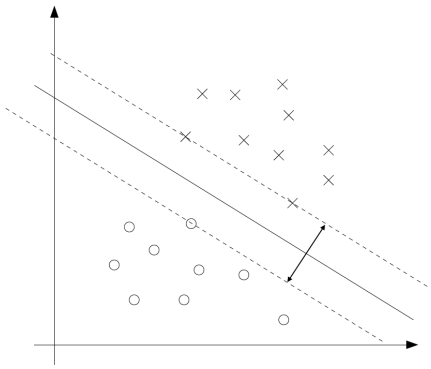

Figure 2: Margin of a SVM, with the width of the margin indicated by the two-sided arrow.

The margin points satisfy the inequality constraint with equality, either with $w^\top x^{(i)} + b = 1$ or $w^\top x^{(i)} + b = -1$. Points satisfying $w^\top x^{(i)} + b = \pm 1$ are said to be the support vectors of a SVM.

Recall that the choice of ± 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 $(w/c)^\top x^{(i)} + b/c = -1$.

Let points x_0, x_1 be points on opposite sides of the SVM decision boundary, satisfying $w^\top x_1 + b = 1$ and $w^\top x_0 + b = -1$ respectively. Then the width of the margin d can be found as the projection of $x_1 - x_0$ 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 $\frac{1}{2}||w||_2^2$ maximizes the width d of the margin.

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 map the current feature space to a new feature space where classification may become easier. This is because once the parameters of the dual problem are found, prediction involves the dot product between a test sample with 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 directly apply weak Slater's condition to assert that $p^* = d^*$.

We 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 and b,

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

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

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

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

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

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

With w^*, α^*, b^* solved, prediction with a new test sample x involves a simple dot product between x and points in the training dataset.

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

The number of dot products required is only equal to the number of support vectors because of the Karush-Kuhn-Tucker (KKT) conditions. To see this, the fact that $p^* = d^*$ implies that the KKT conditions are satisfied. Let $g(w, b) = 1 - y^{(i)}(w^\top x^{(i)} + b)$ correspond to the inequality constraint in our primal problem, then one of the KKT conditions is that $\alpha_i^* g_i(w^*, b^*) = 0$ for all $i = 1, \ldots, m$. If $g_i(w^*, b^*) < 1$, then this necessarily implies $\alpha_i^* = 0$, so we can avoid computing the dot product between the non-support vector training points with the test point. On the other hand, if $\alpha_i^* > 0$, then this necessarily implies $g_i(w^*, b^*) = 0 \Rightarrow y^{(i)}(w^\top x^{(i)} + b) = 1$. In other words, $\alpha_i > 0$ only corresponds to support vectors.

Since both the dual problem and the test prediction only involves the inner product between feature vectors x, this leads to the natural application of the kernel trick, which allows learning in a new high dimensional feature space without explicitly computing the new feature vectors. See the SVM note of Andrew Ng's machine learning course for details on the kernel trick [1].

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 3.

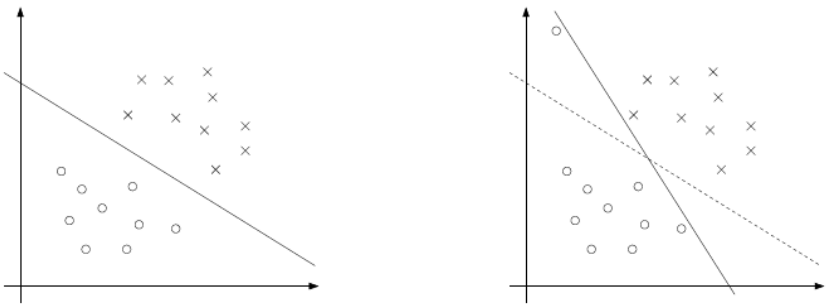

Figure 3: Fitting a hard-margin SVM with outliers.

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

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 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

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

[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 $1 + s^{(i)} y^{(i)} = 1 - s^{(i)}$.

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

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.

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

The zero-one loss is graphically depicted in Figure 4.

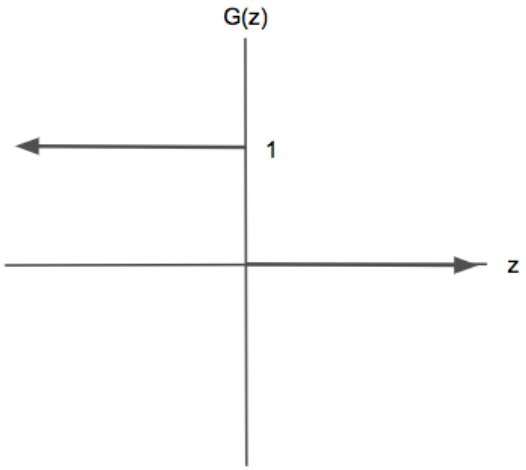

Figure 4: 1-0 loss.

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

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

The zero-one loss for m samples is

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

However, this loss treats all misclassified samples the same, regardless of how far away they are from the hyperplane. Additionally, the loss function $\mathcal{L}(w, b)$ 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 5.

[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 $\lambda \in [0, 1]$.

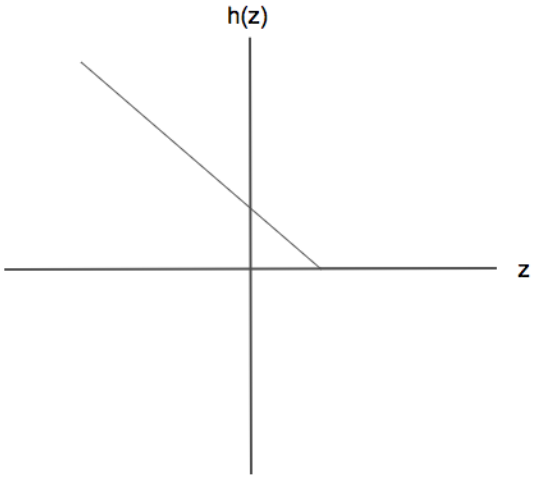

Figure 5: Hinge loss.

Thus, the hinge loss for m samples is

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

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

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

which is equivalent to the primal optimization objective 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 $s^{(i)}$ 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.