Explore Maximum Margin Classifiers

Support Vector Machines (SVMs) search for the best hyperplane that separates both the classes. They define the "best hyperplane" as the one that has the maximum margin to the closest points in both classes. These hyperplanes are thus called "maximum margin separating hyperplanes". Below we will formulate the search for the maximum margin separating hyperplane as a constrained optimization problem.

The objective is to maximize the margin of the points closest to the hyperplane, under the constraints that all data points must lie on the correct side of the hyperplane:

$$\underbrace{\max_{\mathbf{w},b} \gamma(\mathbf{w},b)}_{\textbf{maximize margin}}$$
$$\textbf{such that } \underbrace{\forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 0}_{\textbf{separating hyperplane}}$$

Recall that the margin γ is the distance from the hyperplane to the closest point(s) in set D, defined as:

$$\gamma(\mathbf{w},b) = \frac{1}{||\mathbf{w}||_2} \min_{\mathbf{x}_i \in D} \left|\mathbf{w}^\top \mathbf{x}_i + b\right|$$

We can plug in this definition to obtain:

$$\underbrace{\underbrace{\max_{\mathbf{w},b} \frac{1}{||\mathbf{w}||_2} \min_{\mathbf{x}_i \in D} \left|\mathbf{w}^\top \mathbf{x}_i + b\right|}_{\gamma(\mathbf{w},b)}}_{\textbf{maximize margin}}$$
$$\textbf{such that } \underbrace{\forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 0}_{\textbf{separating hyperplanes}}$$

Simplifying the Objective Further

We can simplify the object and constraints further. Observe that hyperplanes are scale invariant. That is, we can scale the defining parameters $\mathbf{w}$ and b with any scalar $\tau \neq 0$ to get a hyperplane $(\tau\mathbf{w})^\top \mathbf{x} + (\tau b) = 0$ that would specify the same set of points as $\mathbf{w}^\top \mathbf{x} + b = 0$ would. Why is that? The set of points $\mathbf{x}$ that make $\mathbf{w}^\top \mathbf{x} + b = 0$ also make $(\tau\mathbf{w})^\top \mathbf{x} + (\tau b) = \tau\left(\mathbf{w}^\top \mathbf{x} + b\right) = 0$ and vice versa. Therefore, we can exploit this scale invariance and assume that our maximum margin hyperplane satisfies $\min_{\mathbf{x}_i \in D} \left|\mathbf{w}^\top \mathbf{x}_i + b\right| = 1$.

We can add this re-scaling as a constraint. Then our objective simplifies to: $\max_{\mathbf{w},b} \frac{1}{||\mathbf{w}||_2} \cdot 1 = \min_{\mathbf{w},b} ||\mathbf{w}||_2 = \min_{\mathbf{w},b} \mathbf{w}^\top \mathbf{w}$. Here we applied the fact that the $\mathbf{w}$ that

maximizes $||\mathbf{w}||_2$ also maximizes $\mathbf{w}^\top \mathbf{w}$.

The new optimization problem is:

$$\min_{\mathbf{w},b} \mathbf{w}^\top \mathbf{w}$$
$$\text{such that } \forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 0$$
$$\min_i \left|\mathbf{w}^\top \mathbf{x}_i + b\right| = 1$$

Simplifying the Constraints Further

These set of constraints are still hard to deal with, but luckily we can show that, for the optimal solution, they are equivalent to a much simpler set of constraints given below. We recommend that you understand the proof in the following subsection as well.

$$\min_{\mathbf{w},b} \mathbf{w}^\top \mathbf{w}$$
$$\text{such that } \forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1$$

This type of formulation is called a quadratic optimization problem. The objective is *quadratic* and the constraints are all *linear*. Fortunately, there are existing off-the-shelf programming libraries that can solve such formulations efficiently. The problem has a unique solution whenever a separating hyperplane exists. It also has a nice interpretation:

Find the simplest hyperplane (where simpler equals smaller $\mathbf{w}^\top \mathbf{w}$) such that all points lie at least 1 unit away from the hyperplane on the correct side.

Why are the Two Sets of Constraints Equivalent?

Let's look at how the complex and simple formulations are equivalent by proving that they are equal in both directions (one implies the other).

Complex	Simple
$\min_{\mathbf{w},b} \mathbf{w}^\top \mathbf{w}$ $\text{such that } \forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 0$ $\min_i \left\|\mathbf{w}^\top \mathbf{x}_i + b\right\| = 1$	$\min_{\mathbf{w},b} \mathbf{w}^\top \mathbf{w}$ $\text{such that } \forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1$

Complex Implies Simple

We show that the optimal solution of the complex formulation satisfies all constraints of the simpler formulation.

From the second constraint $\min_i \left|\mathbf{w}^\top \mathbf{x}_i + b\right| = 1$, we know that $\forall i, \left|\mathbf{w}^\top \mathbf{x}_i + b\right| \geq 1$. Since $y_i \in \{+1, -1\}$, we can multiply $\mathbf{w}^\top \mathbf{x}_i + b$ by y_i without changing the inequality. Thus, you

get $\forall i, \left|y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right| \geq 1$. The first constraint also tells us that $\forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 0$. Together, we get that the simpler constraints are met, i.e., $\forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1$.

Simple Implies Complex

We show that the optimal solution of the simpler formulation satisfies all constraints of the complex formulation.

Firstly, it is trivial that the first constraints in the complex formulation are met as $\forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1$ implies $\forall i, y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 0$. It also implies that $\forall i, \left|y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right| \geq 1$. Since $y_i \in \{+1, -1\}$, removing y_i from the absolute value will not change the inequality. Thus,$\forall i, \left|\mathbf{w}^\top \mathbf{x}_i + b\right| \geq 1$. This shows that the second constraint is also met, i.e., $\mathbf{min}_i \left|\mathbf{w}^\top \mathbf{x}_i + b\right| = 1$.

Support Vectors

For the optimal $(\mathbf{w}, b)$ pair that defines the optimal hyperplane, some training points will satisfy the constraint with *equality,* i.e. $y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) = 1$. Convince yourself that there will always be such training points, because if for all training points you had a strict $>$ inequality, it would be possible to scale down parameters $(\mathbf{w}, b)$ and minimize $\mathbf{w}^\top \mathbf{w}$ (without misclassifying any points) until we reach equality. These training points are called **support vectors**.

Support vectors are special because they are the training points that define the maximum margin of the hyperplane to the dataset. Therefore, they determine the shape of the hyperplane. If you were to move one of them and retrain the SVM, the resulting hyperplane would change. The opposite is the case for non-support vectors (provided you don't move them so much that they become support vectors themselves).