Contents

1 Hyperplane

In a p-dimensional space, a hyperplane is a flat affine (meaning the subspace need not pass through the origin) subspace of dimension $p-1$. For instance, in two dimensions, a hyperplane is a flat one-dimensional subspace-in other words, a line. In three dimensions, a hyperplane is a flat two-dimensional subspace-that is, a plane. In $p > 3$ dimensions, it can be hard to visualize a hyperplane, but the notion of a $(p-1)$-dimensional flat subspace still applies.

1.1 Two-Dimensional Space

The mathematical definition of a hyperplane can be expressed by an equation. In two dimensions, a hyperplane is defined by the equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{1}$$

where

- $\langle \beta_1, \beta_2 \rangle$ is the normal vector

for parameters β_0, β_1, and β_2. **When we say that the equation above "defines" the hyperplane, we mean that any $X = (X_1, X_2)^T$ for which the equation holds is a point on the hyperplane**. Note that equation 1 is simply the equation of a line, since indeed in two dimensions a hyperplane is a line in one-dimensional subspace.

1.2 Three-Dimensional Space

In three-dimensional space, the hyperplane is a plane given the the equation:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = 0$$

where

- $\langle \beta_1, \beta_2, \beta_3 \rangle$ is the normal vector
- $\beta_0 \equiv = -\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3$ is the offset for normal vector that goes through the point $\langle X_1, X_2, X_3 \rangle$

A plane specified in this form therefore has X_1 (or x), X_2 (or y), and X_3 (or z)-intercepts at

$$X_1 = -\frac{\beta_0}{\beta_1}$$
$$X_2 = -\frac{\beta_0}{\beta_2}$$
$$X_3 = -\frac{\beta_0}{\beta_3}$$

and lies at a distance

$$D = \frac{\beta_0}{\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2}}$$

from the origin.

1.3 p-Dimensional Space

In p-dimensional setting. the following equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \tag{2}$$

defines a p-dimensional hyperplane, again in the sense that if a point $X = (X_1, X_2, \ldots, X_p)^T$ in p-dimensional space (i.e. a column vector of length p) satisfies equation 2, then X lies on the hyperplane. A hyperplane is a higher-dimensional generalization of lines and planes. The equation of a hyperplane is can be expressed in vector form as follows:

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{bmatrix} \cdot \begin{bmatrix}X_1\\ X_2\\ \vdots\\ X_p\end{bmatrix} + \beta_0 = 0$$

$$\vec{\beta} \cdot \vec{X} + \beta_0 = \vec{0}$$

where

- $\vec{\beta}$ is a vector normal to the hyperplane
- β_0 is an offset

Or, when represented by matrix multiplications:

$$\begin{bmatrix}\beta_1 & \beta_2 & \ldots & \beta_p\end{bmatrix} \begin{bmatrix}X_1\\ X_2\\ \vdots\\ X_p\end{bmatrix} + \beta_0 = 0$$

$$\vec{\beta}^T \vec{X} + \beta_0 = \vec{0}$$

where

- $\vec{\beta}^T$ is a row vector normal to the hyperplane
- β_0 is an offset

Note that hyperplanes are scale invariant, meaning we can multiply the vector equation above by any constant and preserve the equality. In particular, if we multiply by $\frac{1}{\|\vec{\beta}\|}$ (thus normalizing the equation), we get a new equation

$$\hat{\vec{\beta}} \cdot \vec{X} + \beta_0' = \vec{0}$$

where

- $\hat{\vec{\beta}} = \frac{\vec{\beta}}{||\vec{\beta}||}$ is the unit normal vector
- $\beta_0' = \frac{\beta_0}{||\vec{\beta}||}$ is the distance from the hyperplane to the origin $\vec{0}$

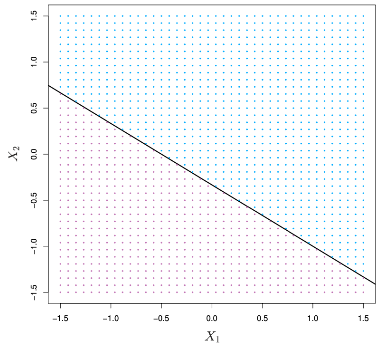
Now, suppose that X does not satisfy; rather,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p > 0$$

Then this tells us that X lies to one side of the hyperplane. On the other hand, if

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p < 0$$

then X lies on the other side of the hyperplane. **So we can think of the hyperplane as dividing p-dimensional space into two halves. One can easily determine on which side of the hyperplane a point lies by simply calculating the sign of the left hand side of equation 2**. A hyperplane in two-dimensional space is shown below:



The hyperplane $1 + 2X_1 + 3X_2 = 0$ is shown above. The blue region is the set of points for which $1 + 2X_1 + 3X_2 > 0$, and the purple region is the set of points for which $1 + 2X_1 + 3X_2 < 0$.

2 Separating Hyperplane

Suppose we have an $n \times p$ data or design matrix $\vec{X}$ that consists of n training observations in p-dimensional space,

$$\vec{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \cdots & X_{np} \end{bmatrix}$$

$$X_1 = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1p} \end{pmatrix}, \ldots, X_n = \begin{pmatrix} X_{n1} \\ \vdots \\ X_{np} \end{pmatrix}$$

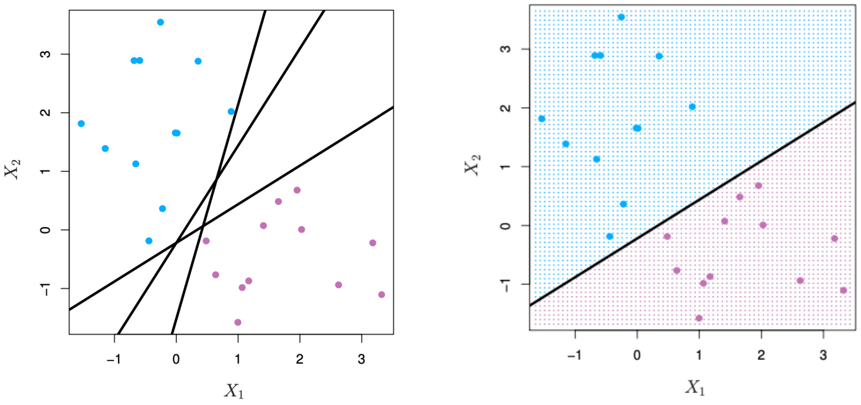
and that these observations fall into two classes; in other words,

$$Y_1, \ldots, Y_n \in \{-1, 1\}$$

where -1 represents one class and 1 the other class. On the other hand, we also have a test observation, a p-vector (vector of length p) of observed features

$$\vec{X}_i^* = \begin{pmatrix} X_{i1}^* & X_{i2}^* & \ldots & X_{ip}^* \end{pmatrix}^T$$

The goal is to develop a classifier based on the training data $\vec{X}$ that will correctly classify the test observations using their feature measurements $\vec{X}_i^*$. Suppose that it is possible to construct a hyperplane that separates the training observations perfectly according to their class labels. Examples of three such **separating hyperplanes** are shown in the left-hand panel of figure below:



In the left, there are two classes of observations, shown in blue and in purple, each of which has measurements on two variables. Three separating hyperplanes, out of many possibilities, are shown in black. In the right, the blue and purple grid indicate the decision rule made by a classifier based on this separating hyperplane: a test observation that falls in the blue portion of the grid will be assigned to the blue class, and a test observation that falls into the purple portion of the grid will be assigned to the purple class.

We can label the observations from the blue class (above hyperplane) as $Y_i = 1$ and those from the purple class (below hyperplane) as $Y_i = -1$. Then a separating hyperplane has the property that

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} > 0 \text{ if } Y_i = 1 \qquad \text{(blue class)}$$

and

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} < 0 \text{ if } Y_i = -1 \qquad \text{(purple class)}$$

Equivalently, a separating hyperplane has the property that

$$Y_i \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \right) > 0$$

for all $i = 1, \ldots, n$. This is because:

- if $Y_i = 1$:

$$\begin{aligned}(1)\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &> 0\\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &> \frac{0}{1}\\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &> 0\end{aligned}$$

- if $Y_i = -1$:

$$\begin{aligned}(-1)\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &> 0\\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &< \frac{0}{-1}\\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &< 0\end{aligned}$$

If a separating hyperplane exists, we can use it to construct a classifier— **a test observation is assigned a class depending on which side of the hyperplane it is located**. The right-hand panel of the figure above shows an example of such a classifier. That is, we classify the test observation $\vec{X}_i^*$ based on the sign of

$$f\left(\vec{X}_i^*\right) = \beta_0 + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \cdots + \beta_p X_{ip}^*$$

where

- if $f\left(\vec{X}_i^*\right)$ is positive, we assign the test observation to 1 class
- if $f\left(\vec{X}_i^*\right)$ is negative, we assign the test observation to class -1.

We can also make use of the magnitude of $f\left(\vec{X}_i^*\right)$.

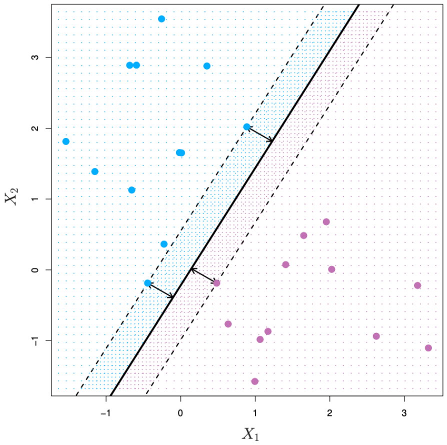
- if $f\left(\vec{X}_i^*\right)$ is far from zero, then this means that $\vec{X}_i^*$ lies far from the hyperplane, and so we can be confident about our class assignment for $\vec{X}_i^*$.
- if $f\left(\vec{X}_i^*\right)$ is close to zero, then $\vec{X}_i^*$ is located near the hyperplane, and so we are less certain about the class assignment for $\vec{X}_i^*$.

A classifier that is based on a separating hyperplane leads to a linear decision boundary.

3 The Maximal Margin Classifier

In general, if our data can be perfectly separated using a hyperplane, then there will in fact exist an infinite number of such hyperplanes. To select the optimal separating hyperplane, we use the **maximal margin**

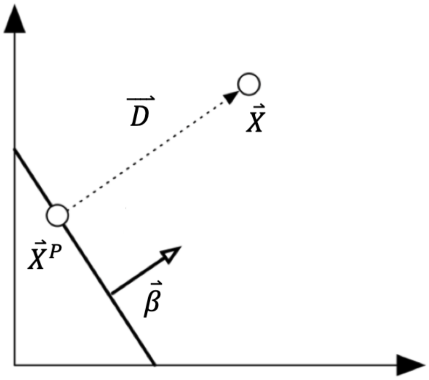
hyperplane (also known as the **optimal separating hyperplane**), which is the separating hyperplane that is farthest from the training observations. In other words, we can compute the (perpendicular) distance from each training observation to a given separating hyperplane; the smallest such distance is the minimal distance from the observations to the hyperplane, and is known as the **margin.** The maximal margin hyperplane is the **separating hyperplane for which the margin is largest**. It is the hyperplane that has the farthest minimum distance to the training observations. We can then classify a test observation based on which side of the maximal margin hyperplane it lies. A two-dimensional example is shown below:



There are two classes of observations, shown in blue and in purple. The **maximal margin hyperplane** is shown as a solid line. The margin is the distance from the solid line to either of the dashed lines. The two blue points and the purple point that lie on the dashed lines are the **support vectors**, and the distance from those points to the hyperplane is indicated by the arrows. They are known as support vectors, since they are vectors in p-dimensional space and they “support” the maximal margin hyperplane in the sense that if these points were moved slightly then the maximal margin hyperplane would move too.

3.1 Margin

Examine the figure below in two-dimensional space, in which the hyperplane is a line:



Another notation used to define a hyperplane is as follows:

Definition 3.1. A hyperplane is defined through $\vec{\beta}, \beta_0$ as a set of points such that $\mathcal{H} = \left\{\vec{X} \mid \vec{\beta}^T\vec{X} + \beta_0 = \vec{0}\right\}$.

What is the distance $\vec{D}$ of a point $\vec{X}$ to the hyperplane $\mathcal{H}$?

1. Let the margin γ be defined as the distance from the hyperplane to the closest point across both classes.
2. Consider some point $\vec{X}$ not on the hyperplane.
3. Let $\vec{D}$ be the displacement vector from the hyperplane $\mathcal{H}$ to $\vec{X}$ of minimum length (that is, a straight line from the hyperplane to the point).
4. Let $\vec{X}^P$ be the projection of $\vec{X}$ onto $\mathcal{H}$.
5. It follows then that: $\vec{X}^P = \vec{X} - \vec{D}$.
6. The displacement vector $\vec{D}$ is parallel to the normal vector $\vec{\beta}$, so $\vec{D} = \alpha\vec{\beta}$ for some constant $\alpha \in \mathbb{R}$ where
    - $\alpha\vec{\beta}$ points in the same direction as $\vec{D}$ if $\alpha > 0$ and in the opposite direction if $\alpha < 0$.
7. Since $\vec{X}^P \in \mathcal{H}$ (it is a point on the hyperplane), it implies that $\vec{X}^P$ satisfies the equation $\vec{\beta}^T\vec{X}^P + \beta_0 = \vec{0}$.
8. Therefore, plugging $\vec{X} - \vec{D}$ for $\vec{X}^P$ in the equation above:

$$\begin{aligned}\vec{\beta}^T\vec{X}^P + \beta_0 &= \vec{0}\\ \vec{\beta}^T(\vec{X} - \vec{D}) + \beta_0 &= \vec{0}\end{aligned}$$

Now, given that $\vec{D} = \alpha\vec{\beta}$, we plug in $\alpha\vec{\beta}$ for $\vec{D}$:

$$\vec{\beta}^T(\vec{X} - \alpha\vec{\beta}) + \beta_0 = \vec{0}$$

Solving for α:

$$\vec{\beta}^T(\vec{X} - \alpha\vec{\beta}) = -\beta_0$$

Matrix multiplication is distributive

$$\begin{aligned}\vec{\beta}^T\vec{X} - \vec{\beta}^T\left(\alpha\vec{\beta}\right) &= -\beta_0\\ -\alpha\vec{\beta}^T\vec{\beta} &= -\beta_0 - \vec{\beta}^T\vec{X}\\ -\alpha\vec{\beta}^T\vec{\beta} &= -\left(\beta_0 + \vec{\beta}^T\vec{X}\right)\\ \cancel{-}\alpha &= \frac{\cancel{-}\left(\beta_0 + \vec{\beta}^T\vec{X}\right)}{\vec{\beta}^T\vec{\beta}}\end{aligned}$$

$$\alpha = \frac{\beta_0 + \vec{\beta}^T\vec{X}}{\vec{\beta}^T\vec{\beta}}$$

Therefore,

$$\begin{aligned}
\vec{D} &= \alpha\vec{\beta} \\
&= \Big(\frac{\beta_0 + \vec{\beta}^T\vec{X}}{\vec{\beta}^T\vec{\beta}}\Big)\vec{\beta}
\end{aligned}$$

9. The l^p-norm of a vector is defined by:

$$\|x_i\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

The norm of a vector is related to the dot product as follows:

$$\begin{aligned}
\vec{x}\cdot\vec{x} &= \|\vec{x}\|_2\|\vec{x}\|_2\cos\theta \\
&= \|\vec{x}\|\|_2\vec{x}\|_2\cos 0 \\
&= \|\vec{x}\|_2\|\vec{x}\|_2(1) \\
&= \|\vec{x}\|_2^2
\end{aligned}$$

Therefore,

$$\|\vec{x}\|_2 = \sqrt{\vec{x}\cdot\vec{x}}$$

Now, the length or magnitude or l^2-norm of the vector $\vec{D}$ is:

$$\begin{aligned}
\|\vec{D}\|_2 &= \sqrt{\vec{D}\cdot\vec{D}} \\
&= \sqrt{\alpha\vec{\beta}\cdot\alpha\vec{\beta}} \\
&= \sqrt{\alpha^2(\vec{\beta}\cdot\vec{\beta})} \\
&= \sqrt{\alpha^2(\vec{\beta}^T\vec{\beta})} \\
&= \sqrt{\alpha^2}\sqrt{\vec{\beta}^T\vec{\beta}} \\
&= \alpha\sqrt{\vec{\beta}^T\vec{\beta}} \\
&= \Big(\frac{\beta_0 + \vec{\beta}^T\vec{X}}{\cancel{\vec{\beta}^T\vec{\beta}}}\Big)\cancel{\sqrt{\vec{\beta}^T\vec{\beta}}} \\
&= \frac{|\beta_0 + \vec{\beta}^T\vec{X}|}{\sqrt{\vec{\beta}^T\vec{\beta}}} \\
&= \frac{|\beta_0 + \vec{\beta}^T\vec{X}|}{\sqrt{\vec{\beta}\cdot\vec{\beta}}}
\end{aligned}$$

$$= \frac{|\beta_0 + \vec{\beta}^T \vec{X}|}{\|\vec{\beta}\|_2}$$

10. Finally, the margin of the hyperplane $\mathcal{H}$ with respect to D, which stands for the data set (training set), is represented by the following minimization problem:

$$\gamma(\vec{\beta}, \beta_0) = \min_{\vec{X} \in D} \frac{\left|\vec{\beta}^T \vec{X} + \beta_0\right|}{\|\vec{\beta}\|_2}$$

    We compute the distance $\vec{D}$ for all $\vec{X} \in D$ and find the smallest distance— this distance is the **margin**.

By definition, the margin and hyperplane are scale invariant. Take a constant c:

$$\gamma(c\vec{\beta}, c\beta_0) = \gamma(\vec{\beta}, \beta_0), \quad \forall\, c \neq 0$$

Note that if the hyperplane $\mathcal{H}$ is such that γ is maximized, it must lie right in the middle of the two classes. In other words, γ must be the distance to the closest point within both classes. (If not, we would be able move the hyperplane towards data points of the class that is further away and increase γ, which contradicts that γ is maximized.)

3.2 Construction of Maximal Margin Classifier

The construction of the maximal margin classifier, a linear classifier, in a binary classification setting involves an optimization problem— that is, we wish to find the **maximum margin separating hyperplane**. Unfortunately, although the optimization is simple in nature, it is often presented using different notations and in slightly different frameworks. Below, we provide two of such representations from two sources— the Introduction to Statistical Learning textbook and the Cornell Machine Learning course.

3.2.1 Introduction to Statistical Learning

The goal is to find the maximal margin hyperplane based on a set of n training observations $X_1, \ldots, X_n \in R^p$ and associated class labels $Y_1, \ldots, Y_n \in \{-1, 1\}$. The maximal margin hyperplane is the solution to the constrained optimization problem

$$\begin{gathered}\max_{\beta_0, \beta_1, \ldots, \beta_p, \gamma} \gamma \\ \text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \text{and} \\ Y_i \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) \geq \gamma \qquad \forall i = 1, \ldots, n\end{gathered}$$

In the optimization problem, γ represents the margin of the hyperplane, and the optimization problem chooses $\beta_0, \beta_1, \ldots, \beta_p$ to maximize γ.

Second constraint The second constraint is a linear constraint:

$$Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) \geq \gamma \qquad \forall i = 1, \ldots, n$$

where

- if $Y_i = 1$:

$$\begin{aligned} (1)\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &\geq \gamma \\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &\geq \frac{\gamma}{1} \\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &\geq \gamma \end{aligned}$$

- if $Y_i = -1$:

$$\begin{aligned} (-1)\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &\geq \gamma \\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &\leq \frac{\gamma}{-1} \\ \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) &\leq -\gamma \end{aligned}$$

It guarantees that each observation will be on the correct side of the hyperplane, provided that γ is positive. In fact, for each observation to be on the correct side of the hyperplane, we would simply need

$$Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) > 0$$

so this constraint actually **requires that each observation be on the correct side of the hyperplane, with some cushion, provided that γ is positive**.

First constraint The first constraint is also linear:

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

and it constraints the l^2-norm of $\vec{D}$. Recall that:

$$\|\vec{D}\|_2 = \frac{|\vec{\beta}^T \vec{X} + \beta_0|}{\|\vec{\beta}\|_2}$$

As shown earlier, the denominator of the l^2-norm of $\vec{D}$ can be expressed as a the square root of a dot product:

$$\|\vec{\beta}\|_2 = \sqrt{\vec{\beta} \cdot \vec{\beta}}$$

$$= \sqrt{\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{bmatrix} \cdot \begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{bmatrix}}$$
$$= \sqrt{\beta_1^2 + \beta_2^2 + \ldots + \beta_p^2}$$
$$= \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

As can be seen, the first constraints conveniently leaves us with:

$$\|\vec{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2 = 1}$$
$$= 1$$

Thus, the l^2-norm of $\vec{D}$ is simply:

$$\|\vec{D}\|_2 = \frac{|\vec{\beta}^T\vec{X} + \beta_0|}{1}$$
$$= |\vec{\beta}^T\vec{X} + \beta_0|$$

And the margin subsequently becomes:

$$\gamma(\vec{\beta}, \beta_0) = \min_{\vec{X}\in D} \left|\vec{\beta}^T\vec{X} + \beta_0\right|$$

3.2.2 Cornell Machine Learning Course

In a similar framework, we again formulate our search for the **maximum margin separating hyperplane** as a constrained optimization problem. The objective is to maximize the margin under the constraints that all data points must lie on the correct side of the hyperplane:

$$\underbrace{\max_{\vec{\beta},\beta_0} \gamma(\vec{\beta}, \beta_0)}_{\text{maximize margin}} \text{ such that} \underbrace{\forall\ i = 1, \ldots, n \quad Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) \geq 0}_{\text{separating hyperplane}}$$

When we plug in the definition of the margin γ (and take $\frac{1}{\|\vec{\beta}\|_2}$ outside of the internal minimization since it is not a function of $\vec{X}$), we obtain:

$$\underbrace{\max_{\vec{\beta},\beta_0} \underbrace{\frac{1}{\|\vec{\beta}\|_2} \min_{\vec{X}_i\in D} \left|\vec{\beta}^T\vec{X}_i + \beta_0\right|}_{\gamma(\vec{\beta},\beta_0)}}_{\text{maximize margin}} \text{ s.t. } \underbrace{\forall\ i = 1, \ldots, n \quad Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) \geq 0}_{\text{separating hyperplane}}$$

Recall that a hyperplane is scale invariant, so we can fix the scale of $\vec{\beta}$ (essentially shrinking or elongating the normal vector) and β_0 (scaling the offset) anyway we wish. For simplicity, we could choose to scale $\vec{\beta}$

and β_0 such that

$$\min_{\vec{X}_i \in D} \left|\vec{\beta}^T \vec{X}_i + \beta_0\right| = 1$$

We can do this since, regardless of what the margin value may be, we can always scale $\vec{\beta}$ and β_0 (by dividing them by the margin value) so that the margin equals 1. Keep in mind that we have to scale both parameters by the same scalar. The key idea is that we can get whatever **functional** margin (that is, the vector equation $\vec{\beta}^T \vec{X} + \beta_0 = \vec{0}$) we want but still have the same **geometric** margin. We can add this re-scaling as an equality constraint. Then our objective becomes:

$$\max_{\vec{\beta}, \beta_0} \frac{1}{\|\vec{\beta}\|_2} \cdot (\min_{\vec{X}_i \in D} \left|\vec{\beta}^T \vec{X}_i + \beta_0\right|) = \max_{\vec{\beta}, \beta_0} \frac{1}{\|\vec{\beta}\|_2} \cdot 1$$

Note that $\|\vec{\beta}\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$ is non-negative. Thus the reciprocal function $\frac{1}{\|\vec{\beta}\|_2}$ is a monotone decreasing function of $\|\vec{\beta}\|_2$ (proof). Minimizing $\|\vec{\beta}\|_2$ is therefore equivalent to maximizing $\frac{1}{\|\vec{\beta}\|_2}$. Note here that by "equivalent", we do not mean that the optimal objective value will be the same; instead, we mean that any optimal solution to the maximization problem will also be optimal for the minimization problem and vice versa.

$$\max_{\vec{\beta}, \beta_0} \frac{1}{\|\vec{\beta}\|_2} = \min_{\vec{\beta}, \beta_0} \|\vec{\beta}\|_2 = \min_{\vec{\beta}, \beta_0} \vec{\beta}^\top \vec{\beta}$$

Further, squaring a positive quantity is a monotone increasing transformation as $f(z) = z^2$ is a monotonically increasing function for $z \geq 0$. We already know that $\|\vec{\beta}\| \geq 0$ is non-negative. Therefore, minimizing $\|\vec{\beta}\|_2$, a formula which includes a square root $\sqrt{\sum_{j=1}^p \beta_j^2}$, is also equivalent to minimizing $\left(\|\vec{\beta}\|_2\right)^2 = \vec{\beta} \cdot \vec{\beta} = \vec{\beta}^T \vec{\beta}$. The new optimization problem becomes:

$$\begin{aligned} &\min_{\vec{\beta}, \beta_0} \vec{\beta}^\top \vec{\beta} \\ &\text{s.t.} \quad \begin{array}{l} \forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0\right) \geq 0 \\ \min_{\vec{X}_i \in D} \left|\vec{\beta}^T \vec{X}_i + \beta_0\right| = 1 \end{array} \end{aligned}$$

3.2.3 Even Simpler Formulation

We can show that (for the optimal solution), the above set of constraints (correct side and re-scaling) are equivalent to a much simpler formulation:

$$\begin{array}{l} \forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0\right) \geq 0 \\ \min_{\vec{X}_i \in D} \left|\vec{\beta}^T \vec{X}_i + \beta_0\right| = 1 \end{array} \quad \Leftrightarrow \quad \forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0\right) \geq 1$$

Proof. From left to right, we note that the first constraint

$$\forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0\right) \geq 0$$

ensures that each observation i is on the correct side of the hyperplane. The second constraint

$$\min_{\vec{X}_i \in D} \left| \vec{\beta}^T \vec{X}_i + \beta_0 \right| = 1$$

scales the hyperplane parameters so that the equation is satisfied when it equals $\vec{1}$ instead of $\vec{0}$. We can actually rewrite this second constraint to not only accomplish the re-scaling but also ensure that each observation i is on the correct side of the hyperplane. In other words, the two constraints can be combined into one constraint that accomplishes both:

$$\min_{\vec{X}_i \in D} Y_i \left| \vec{\beta}^T \vec{X}_i + \beta_0 \right| = 1$$

Notice that the internal $Y_i \left| \vec{\beta}^T \vec{X}_i + \beta_0 \right|$ determines on which side of the hyperplane each training observation X_i for $\forall X_1, \ldots, X_n$ belongs. When $Y_i = -1$, the value $f(X_i) = \vec{\beta}^T \vec{X}_i + \beta_0$ is multiple by -1 and thus it is on one side of the hyperplane. When $Y_i = 1$, the value $f(X_i) = \vec{\beta}^T \vec{X}_i + \beta_0$ is multiple by 1 and thus it is on the other side of the hyperplane. And we have re-scaled the hyperplane such that the minimum value among all i training observations is 1, i.e. $\min_{\vec{X}_i \in D} Y_i \left| \vec{\beta}^T \vec{X}_i + \beta_0 \right| = 1$. Because 1 is the minimum value, it follows that the values for the other $n - 1$ observations must be greater than 1 (with the minimum equal to 1):

$$\forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right) \geq 1$$

□

Proof. From right to left, the single constraint

$$\forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right) \geq 1$$

already satisfies the first constraint of the left side. That is, if the $Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right)$ are greater or equal to 1 than they are greater than or equal to 0:

$$\forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right) \geq 0$$

Therefore, we only need to ensure that the second constraint

$$\min_{\vec{X}_i \in D} \left| \vec{\beta}^T \vec{X}_i + \beta_0 \right| = 1$$

is satisfied. With the single constraint

$$\forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right) \geq 1$$

the minimum value could actually be any value ≥ 1. For instance, $2, 7, 20, 47, \ldots$ and this single constraint is still satisfied. However, having any of these values as the minimum will surely break the second constraint on the left side. But we are minimizing $\vec{\beta}$:

$$\min_{\vec{\beta},\beta_0} \vec{\beta}^\top \vec{\beta}$$

Therefore, for any minimum value m such that $m \geq 1$, we can always re-scale the hyperplane, by scaling the parameters $\vec{\beta}$ and β_0 by $\frac{1}{m}$ and obtain smaller objective values $\vec{\beta}^\top \vec{\beta}$. The dot product $\vec{\beta}^\top \vec{\beta}$ is smaller since $\vec{\beta}$ and β_0 are both smaller after we scale them by $\frac{1}{m}$. Thus, **in the optimum, the right side (single constraint) implies the left side (two constraints) since the margin will be pushed down towards 1**. □

Finally, the minimization problem becomes:

$$\begin{aligned} &\min_{\vec{\beta},\beta_0} \vec{\beta}^T \vec{\beta} \\ &\text{s.t.} \quad \forall\, i = 1, \ldots, n \quad Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right) \geq 1 \end{aligned}$$

This new formulation is a quadratic optimization problem. The objective is quadratic and the constraints are all linear. We can be solve it efficiently with any QCQP (Quadratically Constrained Quadratic Program) solver. It has a unique solution whenever a separating hyper plane exists. It also has a nice interpretation: Find the simplest hyperplane (where simpler means smaller $\vec{\beta}^\top \vec{\beta}$) such that all observations lie at least 1 unit away from the hyperplane on the correct side.

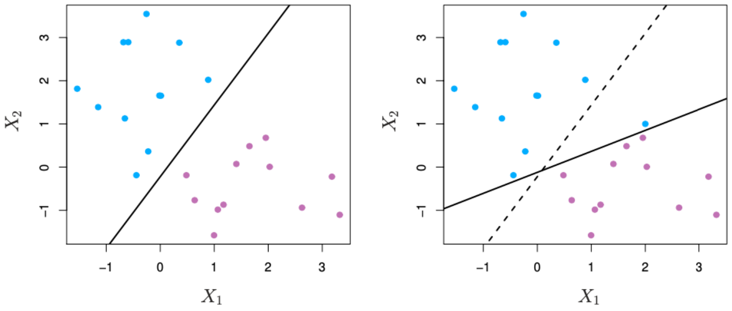
Note that for the optimal $\vec{\beta}, \beta_0$ pair, some training observations will have tight constraints, i.e.

$$Y_i \left(\vec{\beta}^T \vec{X}_i + \beta_0 \right) = 1$$

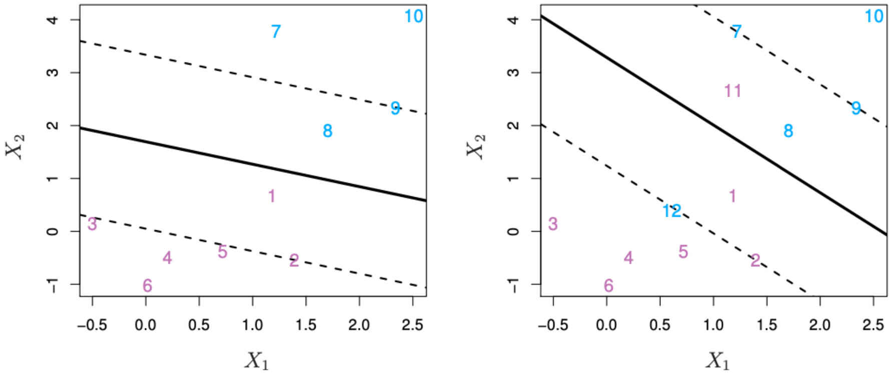
This must be the case, because if for all training observations we had a strict $>$ inequality, it would be possible to scale down both parameters $\vec{\beta}, \beta_0$ until the constraints are tight and obtained an even lower objective value.) Again, we refer to these training observations $\vec{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ as support vectors. Support vectors are special because they are the training observations that define the maximum margin of the hyperplane to the data set and they therefore determine the shape of the hyperplane. If we 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 they don't move too much, or they would turn into support vectors themselves).

4 Support Vector Classifiers

The **maximal margin classifier** is a way to perform classification, if a separating hyperplane exists. However, in many applications and when the data is low dimensional, no separating hyperplane exists, and so there is no maximal margin classifier. In these cases, the optimization problem has no solution. Even when the hyperplane does exist, the maximal margin classifier may be prone to overfitting the training set, leading to poor performance and generalization to the testing set.



A classifier based on a separating hyperplane will necessarily perfectly classify all of the training observations, but this may lead to sensitivity to individual observations. In the right panel above, an additional blue observation has been added, leading to a dramatic shift in the maximal margin hyperplane shown as a solid line. The dashed line indicates the maximal margin hyperplane that was obtained in the absence of this additional point. In either case, it could be worthwhile to misclassify a few training observations in order to do a better job in classifying the remaining observations. The **support vector classifier**, sometimes called a **soft margin classifier**, may be used when a separating hyperplane does not exist or when it is not satisfactory; **rather than seeking the largest possible margin so that every observation is not only 1) on the correct side of the hyperplane but also 2) on the correct side of the margin, we instead allow some observations to be 1) on the incorrect side of the margin, or 2) even the incorrect side of the hyperplane.** The margin is soft because it can be violated by some of the training observations.



In the **left** panel of the figure above, a support vector classifier was fit to a small data set. The hyperplane is shown as a solid line and the margins are shown as dashed lines. Purple observations: Observations 3,4,5, and 6 are on the correct side of the margin, observation 2 is on the margin, and observation 1 is on the wrong side of the margin. Blue observations: Observations 7 and 10 are on the correct side of the margin, observation 9 is on the margin, and observation 8 is on the wrong side of the margin. No observations are on the wrong side of the hyperplane. Int the **right** panel, there are two additional points, 11 and 12. These two observations are on the wrong side of the hyperplane and the wrong side of the margin. Observations on the wrong side of the hyperplane correspond to training observations that are misclassified by the support vector classifier; this is inevitable when a separating hyperplane does not exist.

4.1 Formulation of Support Vector Classifiers

4.1.1 Introduction to Statistical Learning

The support vector classifier classifies a test observation depending on which side of a hyperplane it lies. The hyperplane is chosen to correctly separate **most** of the training observations into the two classes, but may misclassify a few observations. It is the solution to the optimization problem

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\gamma}{\text{maximize}} \gamma$$

subject to

- $\sum_{j=1}^{p} \beta_j^2 = 1$
- $Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) \geq \gamma\left(1 - \epsilon_i\right)$
- $\epsilon_i \geq 0$
- $\sum_{i=1}^{n} \epsilon_i \leq C$

We choose $\beta_0, \beta_1, \ldots, \beta_p$ and $\epsilon_1, \ldots, \epsilon_n$ that maximizes the margin γ. The objective function is once again simplified to

$$\gamma(\vec{\beta}, \beta_0) = \min_{\vec{X} \in D} \left| \vec{\beta}^T \vec{X} + \beta_0 \right|$$

due to the first constraint

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

We classify a test observation $\vec{X}_i^* \in \mathbb{R}^p$ by simply determining on which side of the hyperplane it lies. That is, we again classify the test observation based on the sign of

$$f\left(\vec{X}_i^*\right) = \beta_0 + \beta_1 X_{i1}^* + \cdots + \beta_p X_{ip}^* = \vec{\beta}^T \vec{X}_i^* + \beta_0$$

The formulation of the optimization has the following new variables, parameters and constraints:

1. The $\epsilon_1, \ldots, \epsilon_n$ are **slack variables** that allow individual observations to be on the wrong side of the margin or the hyperplane. **They tell us where the i^{th} observation is located, relative to the hyperplane and relative to the margin**.

- If $\epsilon_i = 0$ then the i^{th} observation is on the correct side of the margin:

  $$\begin{aligned} Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) &\geq \gamma\left(1-0\right) \\ Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) &\geq \gamma \end{aligned}$$

  This is the same constraint as earlier, which requires that the observation be on the correct side of the hyperplane, **outside some cushion**, provided that γ is positive:

  $$Y_i = 1 \longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \geq \gamma$$

  or

  $$Y_i = -1 \longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \leq -\gamma$$

- If $0 < \epsilon_i \leq 1$ then the i^{th} observation is on the wrong side of the **margin**, and we say that the i^{th} observation has violated the margin:

  $$\begin{aligned} Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) &\geq (\gamma \times \text{ 1 minus a value between } 0 < \epsilon_i \leq 1) \\ Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) &\geq (\text{a value} < \gamma \text{ but} \geq 0) \end{aligned}$$

  For instance, if $\epsilon_i = 1$

  $$\begin{aligned} Y_i = 1 &\longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \geq \gamma\left(1-1\right) \\ Y_i = 1 &\longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \geq 0 \end{aligned}$$

  or

  $$Y_i = -1 \longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \leq 0$$

  In both these cases, the observations are still on the correct side of the hyperplane but they can now be **inside some cushion** γ or $-\gamma$. Therefore, they may be in violation of the margin.

- If $\epsilon_i > 1$ then the observation is on the wrong side of the hyperplane:

  $$\begin{aligned} Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) &\geq (\gamma \times \text{ 1 minus a value} > 1) \\ Y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) &\geq -c\gamma \end{aligned}$$

  Thus:

  $$\begin{aligned} Y_i = 1 &\longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \geq \frac{-c\gamma}{1} \\ Y_i = 1 &\longrightarrow \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}\right) \geq -c\gamma \end{aligned}$$

or

$$Y_i = -1 \longrightarrow (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \geq \frac{-c\gamma}{-1}$$

$$Y_i = -1 \longrightarrow (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \leq c\gamma$$

where c is some constant. In both of these cases, the training observation is missclassified.

2. The tuning or hyperparameter C bounds the sum of the ϵ_i 's, and so **it determines the number and severity of the violations to the margin (and to the hyperplane) that we will tolerate**. We can think of C as a budget for the amount that the margin can be violated by the n observations.

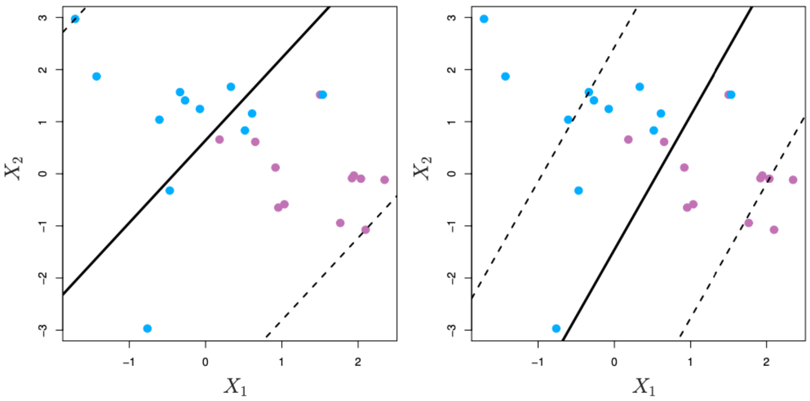
   - If $C = 0$ then there is no budget for violations to the margin, and it must be the case that $\epsilon_1 = \cdots = \epsilon_n = 0$, in which case this simply amounts to the maximal margin hyperplane optimization problem.
   - For $C > 0$, no more than C observations among n training observations can be on the wrong side of the hyperplane. This is because if an observation is on the wrong side of the hyperplane then $\epsilon_i > 1$, and the last constraint requires that $\sum_{i=1}^{n} \epsilon_i \leq C$.

   As the budget C increases, we become more tolerant of violations to the margin, and so the margin will widen. Conversely, as C decreases, we become less tolerant of violations to the margin and so the margin narrows.

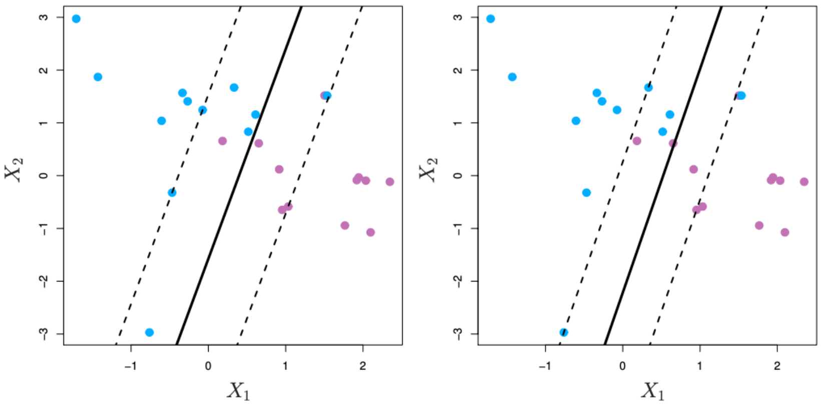
For the support vector classifier, **only observations that either lie on the margin or that violate the margin will affect the hyperplane, and hence the classifier obtained**. In other words, an observation that lies strictly on the correct side of the margin does not affect the support vector classifier. Changing the position of that observation would not change the classifier at all, provided that its position remains on the correct side of the margin. **Observations that lie directly on the margin, or on the wrong side of the margin for their class, are known as support vectors. These observations do affect the support vector classifier.**

The paramter C controls the bias-variance trade-off of the statistical learning technique. When C is small, we seek narrow margins that are rarely violated; this amounts to a classifier that is highly fit to the data, which may have low bias but high variance. On the other hand, when C is larger, the margin is wider and we allow more violations to it; this amounts to fitting the data less hard and obtaining a classifier that is potentially more biased but may have lower variance.

When the tuning or hyperparameter C is large, the margin is wide, many observations violate the margin, and so there are many support vectors. In this case, many observations are involved in determining the hyperplane. In contrast, if C is small, then there will be fewer support vectors and hence the resulting classifier will have low bias but high variance.



When C is large, then there is a high tolerance for observations being on the wrong side of the margin, and so the margin will be large. As C decreases, the tolerance for observations being on the wrong side of the margin decreases, and the margin narrows.



4.1.2 Cornell Machine Learning Course

In this formulation, we allow the constraints demonstrated earlier to be violated ever so slight with the introduction of slack variables:

$$\min_{\vec{\beta},\beta_0} \vec{\beta}^T\vec{\beta} + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t. } \forall\, i = 1,\ldots,n \quad Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) \geq 1 - \xi_i$$

$$\forall\, i = 1,\ldots,n \quad \xi_i \geq 0$$

The slack variable ξ_i (pronouced "/ksa/" or "/za/") allows the input observation $\vec{X}_i$ to be closer to the hyperplane (meaning within 1 from either side of the hyperplane) or even be on the wrong side, but there is a penalty in the objective function for such "slack". **Note that the hyperparameter C is the opposite of how it is formulated in the Introduction to Statistical Learning textbook**. **If C is very**

large, the SVM becomes very strict and tries to get all points to be on the right side of the hyperplane. If C is very small, the SVM becomes very loose and may "sacrifice" some points to obtain a simpler (i.e. lower $(\|\vec{\beta}\|_2)^2 = \vec{\beta}^T\vec{\beta}$) solution— the normal vector of the hyperplane will be smaller. To see this, note that if $(\|\vec{\beta}\|_2)^2$ is lower, then the denominator in the margin formula will be smaller:

$$\gamma(\vec{\beta}, \beta_0) = \min_{\vec{X} \in D} \frac{\left|\vec{\beta}^T\vec{X} + \beta_0\right|}{\|\vec{\beta}\|_2}$$

Therefore, the margin is greater. This is because, when C is large, the objective function will be minimized for a larger optimal value (due to the extra summation term scaled by C) compared to when C is small.

- When C is relatively large, the sum of the slack variables $\sum_{i=1}^{n} \xi_i$ becomes *costly* for the objective function to minimize; as a result, the optimal solution $\vec{\beta}, \beta_0$ will have larger values. Consequently, the denominator of the margin formula will be larger, and the classifier will have smaller or tighter margins with fewer support vectors.

- When C is relatively small, the sum of the slack variables is not as *costly* for the objective function to minimize; therefore, the solution $\vec{\beta}, \beta_0$ may have smaller values. The classifier would involve more support vectors within the cushion or lie on the wrong side of the margins since the slack variables are not penalized as much by the hyperparameter C.

- When $C = 0$, the optimization problem becomes the same problem as when a separating hyperplane exists.

Uncontrained Formulation If we examine the second constraint in the optimization problem:

$$\forall\, i = 1, \ldots, n \quad Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) \geq 1 - \xi_i$$

We can in fact solve for ξ_i as follows:

$$\xi_i \geq 1 - Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right)$$

Because the objective will always try to minimize ξ_i as much as possible (that is, we would never set ξ_i to be greater), the inequality may be changed to an equality:

$$\xi_i = 1 - Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right)$$

Now, we may determine the ξ_i of a training observation i as follows:

$$\xi_i = \begin{cases} 1 - Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) & \text{if } Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) < 1 \quad \text{(Not naturally satisfied \& a positive } \xi_i \text{ must be allowed for)} \\ 0 & \text{if } Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right) \geq 1 \quad \text{(Constraint is naturally satisfied)} \end{cases}$$

This is equivalent to the following **closed form**:

$$\xi_i = \max\left(1 - Y_i\left(\vec{\beta}^T\vec{X}_i + \beta_0\right), 0\right)$$

To see this:

- In the case where a positive ξ_i must be allowed for, $Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right) < 1$ since the observation lies within the margin or the wrong side of it; as a results, $1-Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)$ must be positive > 0. So,

$$\max\left(1-Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right),0\right)=1-Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)$$

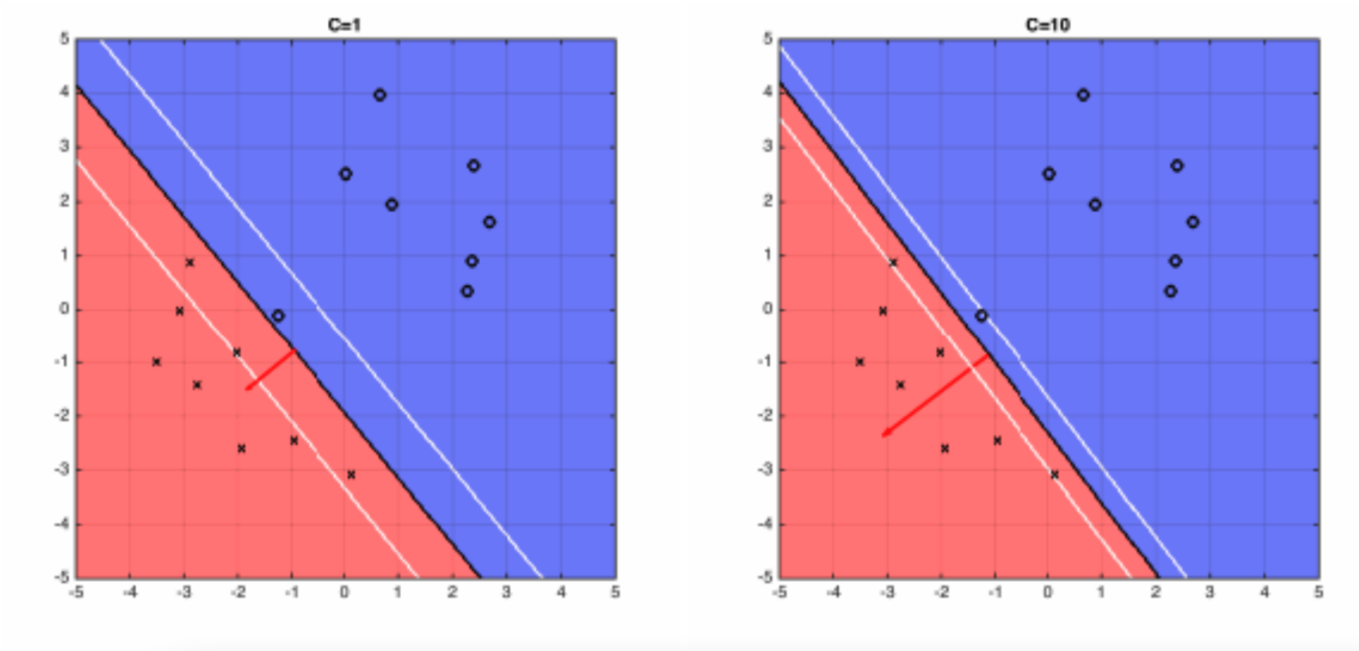
- In the case where the constraint is naturally satisfied, it must be the case that $Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)\geq 1$; that is, the observation lies on the correct side of the hyperplane and the margin. Therefore, the largest $1-Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)$ can be is 0 (that is, when $Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)=1$). If $Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)>1$ then $1-Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right)$ will be negative. Therefore,

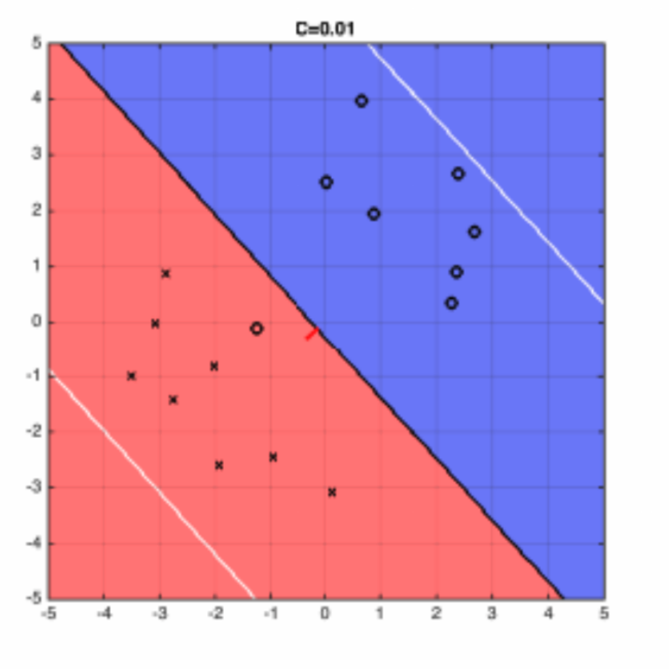
$$\max\left(1-Y_i\left(\vec{\beta}^T\vec{X}_i+\beta_0\right),0\right)=0$$

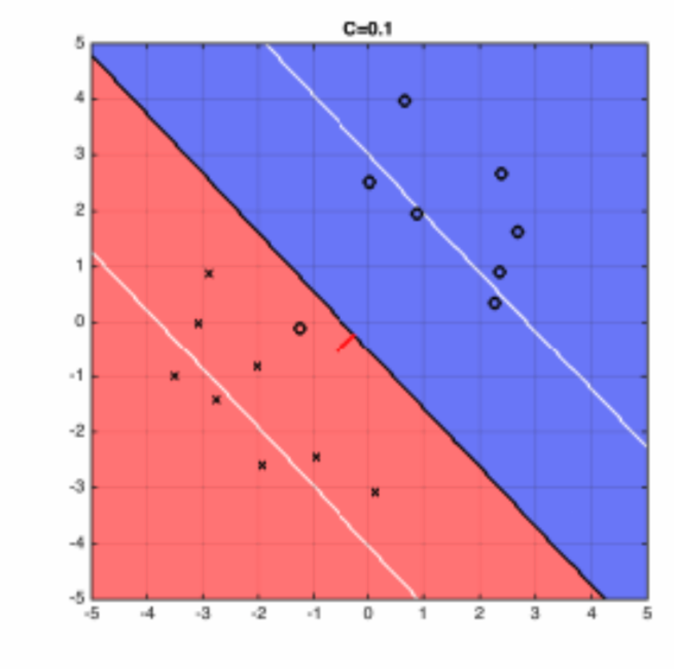
If we plug this closed form into the objective of our SVM optimization problem, we obtain the following unconstrained version as a loss function and a regularizer:

$$\min_{\vec{\beta},\beta_0}\quad\underbrace{\vec{\beta}^T\vec{\beta}}_{l_2-\text{ regularizer}}\quad+\quad C\sum_{i=1}^{n}\underbrace{\max\left[1-Y_i\left(\vec{\beta}^T\vec{X}+\beta_0\right),0\right]}_{\text{hinge-loss}}$$

This formulation allows us to optimize the SVM paramters $(\vec{\beta},\beta_0)$ just like logistic regression (e.g. through gradient descent). In the figures below, we can see a few **soft margin support vector classifiers** with different values for the hyperparameter C. As can be seen, as C gets smaller, the margin becomes softer, allowing for more violations of the margins; in other words, there are more support vectors invovled in determining the hyperplane. In addition, the normal vector $\vec{\beta}$ of the optimal hyperplane also gets smaller as C gets smaller.




C=0.01
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5


C=0.1
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5