Notes for Machine Learning – Brett Bernstein

Recitation 2: Geometric Derivation of SVMs

Intro Question

1. You have been given a data set (x_i, y_i) for $i = 1, \ldots, n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Assume $w \in \mathbb{R}^n$ and $a \in \mathbb{R}$.

   (a) Suppose $y_i(w^T x_i + a) > 0$ for all i. Use a picture to explain what this means when $d = 2$.

   (b) Fix $M > 0$. Suppose $y_i(w^T x_i + a) \geq M$ for all i. Use a picture to explain what this means when $d = 2$.

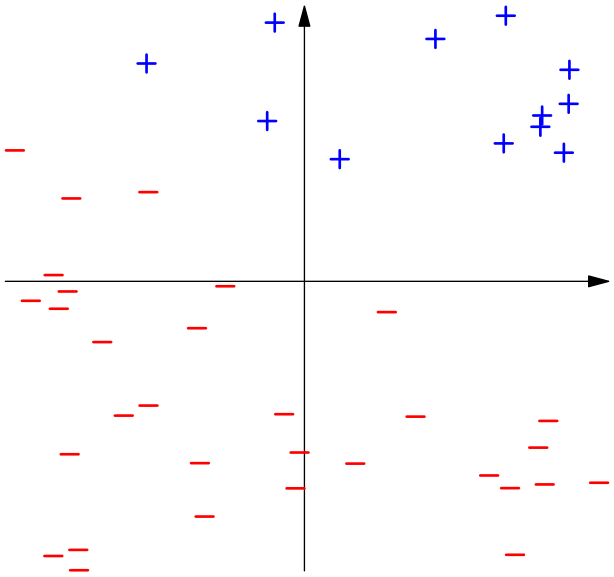

Figure 1: Data set with $x_i \in \mathbb{R}^2$ and $y_i \in \{+1, -1\}$

Support Vector Machines

Review of Geometry

If $v, w \in \mathbb{R}^d$ then the component of v in the direction w is given by $\dfrac{w^T v}{\|w\|_2}$. This can also be thought of as the signed length of v when orthogonally projected onto the line through the vector w.

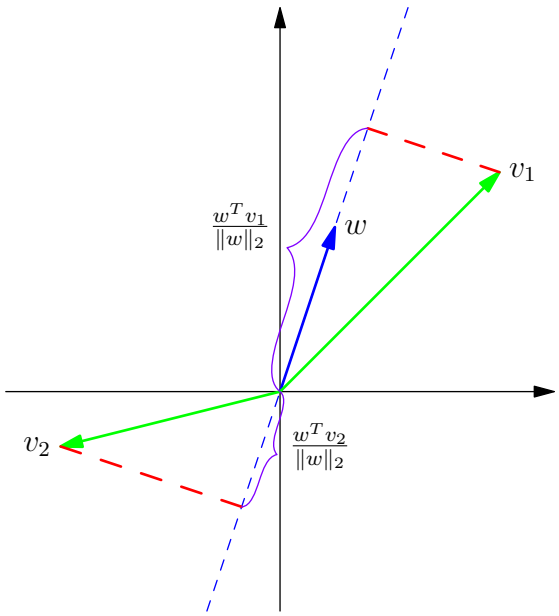

Figure 2: Component of v_1, v_2 in the direction w.

Assuming $w \neq 0$ we can use this to interpret the set

$$S = \{x \in \mathbb{R}^d \mid w^T x = b\}.$$

Noting that $w^T x = b \iff \frac{w^T x}{\|w\|_2} = \frac{b}{\|w\|_2}$ we see that S contains all vectors whose component in the direction w is $\frac{b}{\|w\|_2}$. Using linear algebra we can see this is the hyperplane orthogonal to the vector w that passes through the point $\frac{bw}{\|w\|_2^2}$. Note also that there are infinitely many pairs (w, b) that give the same hyperplane. If $c \neq 0$ then

$$\{x \in \mathbb{R}^d \mid w^T x = b\} \quad \text{and} \quad \{x \in \mathbb{R}^d \mid (cw)^T x = (cb)\}$$

result in the same hyperplanes.

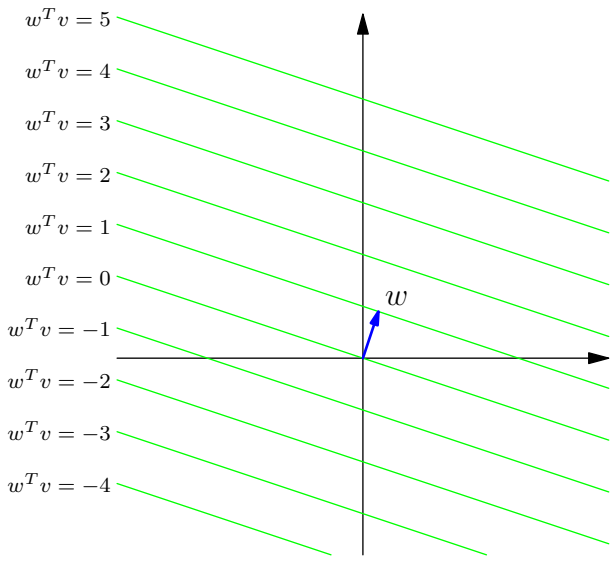

Figure 3: Level Surfaces of $f(v) = w^T v$ with $\|w\|_2 = 1$

Given a hyperplane $\{v \mid w^T v = b\}$, we can distinguish points $x \in \mathbb{R}^d$ depending on whether $w^T x - b$ is zero, positive, or negative, or in other words, whether x is on the hyperplane, on the side w is pointing at, or on the side $-w$ is pointing at.

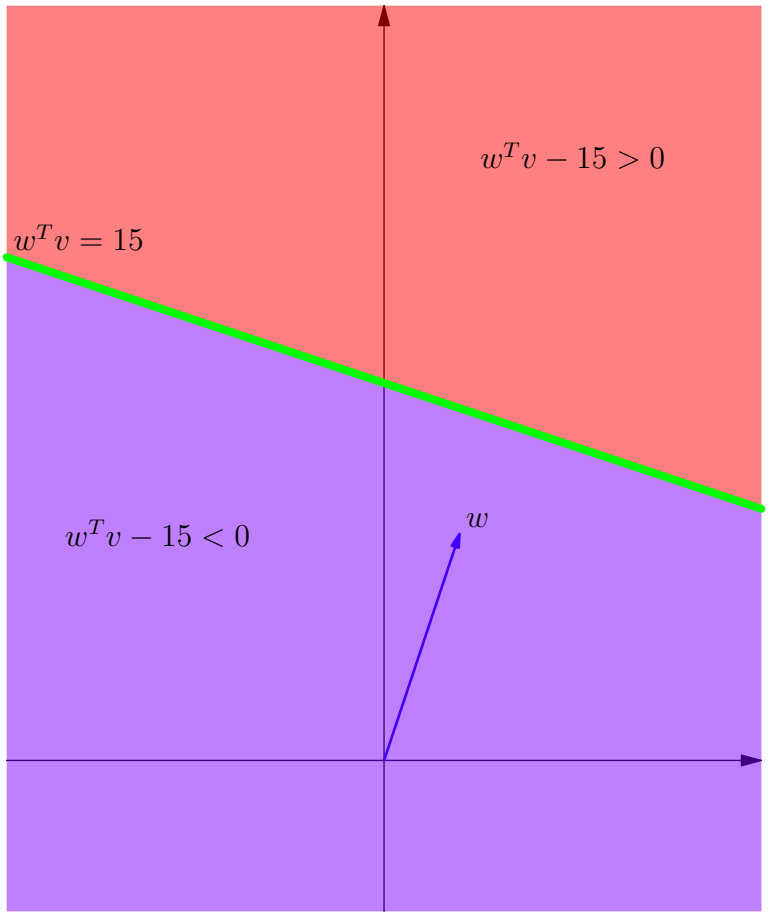

Figure 4: Sides of the Hyperplane $w^T v = 15$

If we have a vector $x \in \mathbb{R}^d$ and a hyperplane $H = \{v \mid w^T v = b\}$ we can measure the distance from x to H by

$$d(x, H) = \left| \frac{w^T x - b}{\|w\|_2} \right|.$$

Without the absolute values we get the *signed distance*: a positive distance if $w^T x > b$ and a negative distance if $w^T x < b$. To see why this formula is correct, note that we are computing

$$\frac{w^T x}{\|w\|_2} - \frac{w^T v}{\|w\|_2},$$

where v is any vector in the hyperplane $\{v \mid w^T v = b\}$.

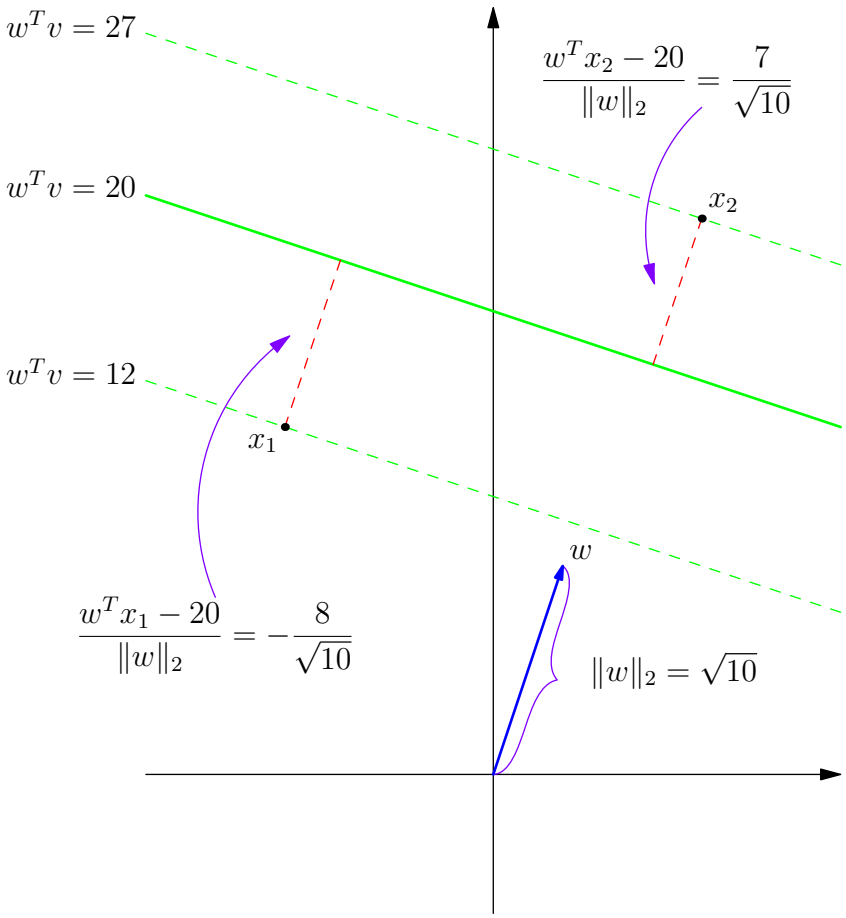

Figure 5: Signed Distance from x_1, x_2 to Hyperplane $w^T v = 20$

Hard Margin SVM

Returning to the initial question, suppose we have the data set (x_i, y_i) for $i = 1, \ldots, n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.

Definition 1 (Linearly Separable)**.** We say (x_i, y_i) for $i = 1, \ldots, n$ is *linearly separable* if there is a $w \in \mathbb{R}^d$ and $a \in \mathbb{R}$ such that $y_i(w^T x_i + a) > 0$ for all i. The set $\{v \in \mathbb{R}^d \mid w^T v + a = 0\}$ is called a *separating hyperplane*.

Let's examine what this definition says. If $y_i = +1$ then we require that $w^T x_i > -a$ and if $y_i = -1$ we require that $w^T x_i < -a$. Thus linearly separable means that we can separate all of the $+1$'s from the -1's using the hyperplane $\{v \mid w^T v = -a\}$. For the rest of this section, we assume our data is linearly separable.

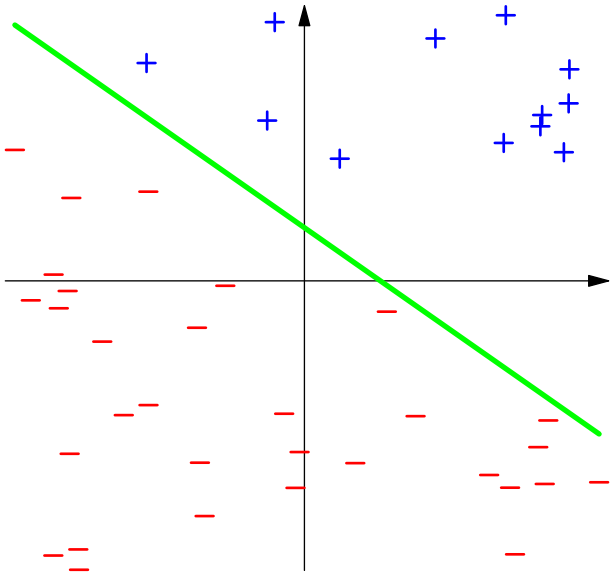

Figure 6: Linearly Separable Data

If we can find the w, a corresponding to a hyperplane that separates the data, we then have a decision function for classifiying elements of $\mathcal{X}$: $f(x) = \text{sgn}(w^T x + a)$. Before we look for such a hyperplane, we must address another issue. If the data is linearly separable, then there are infinitely many choices of separating hyperplanes.

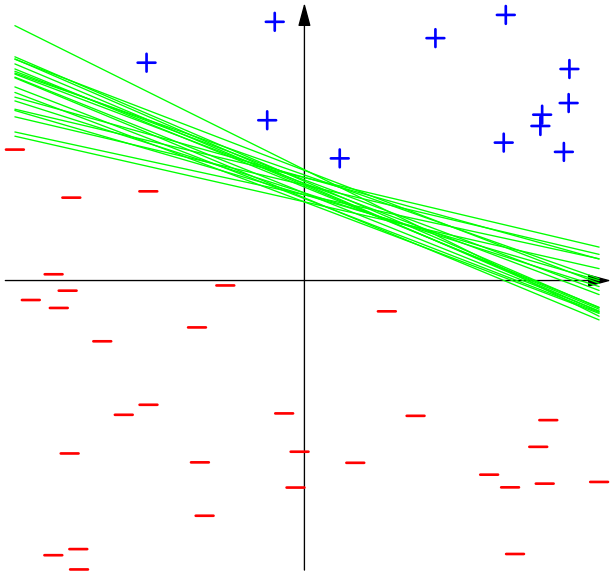

Figure 7: Many Separating Hyperplanes Exist

We will choose the hyperplane that maximizes a quantity called the *margin*.

Definition 2 (Margin)**.** Let H be a hyperplane that separates the data (x_i, y_i) for $i =$

$1, \ldots, n$. The margin of this hyperplane is

$$\min_i d(x_i, H),$$

the distance from the hyperplane to the closest data point.

Fix $w \in \mathbb{R}^d$ and $a \in \mathbb{R}$ with $y_i(w^T x_i + a) > 0$ for all i. Then we saw earlier that

$$d(x_i, H) = \left| \frac{w^T x_i + a}{\|w\|_2} \right| = \frac{y_i(w^T x_i + a)}{\|w\|_2}.$$

This gives us the following optimization problem:

$$\text{maximize}_{w,a} \quad \min_i \frac{y_i(w^T x_i + a)}{\|w\|_2}.$$

We can rewrite this in a more standard form:

$$\begin{aligned} &\text{maximize}_{w,a,M} && M \\ &\text{subject to} && \frac{y_i(w^T x_i + a)}{\|w\|_2} \geq M \quad \text{for all } i. \end{aligned}$$

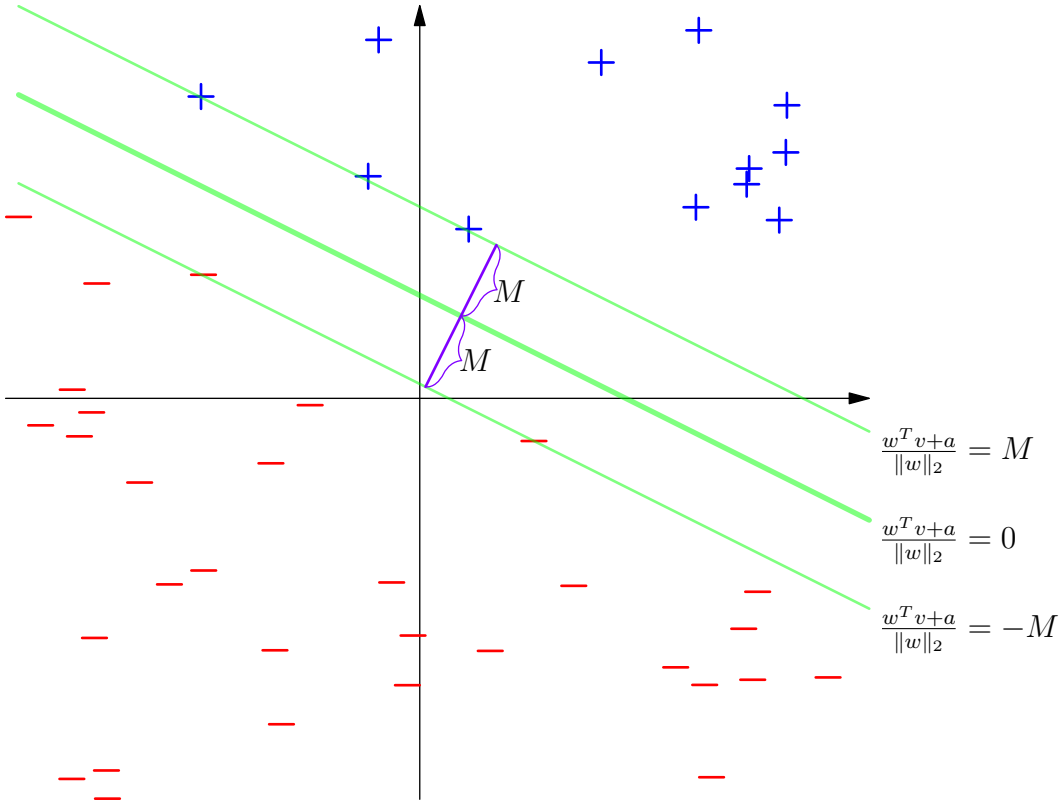

Figure 8: Maximum Margin Separating Hyperplane

Note above how the margin is achieved on both sides of the optimal hyperplane. This must be the case, as otherwise we could slightly move the hyperplane and obtain a better solution. The expression $y_i(w^T x_i + a)/\|w\|_2$ allows us to choose any positive value for $\|w\|_2$ by changing a accordingly (e.g., we can replace $w \to 2w$ and $a \to 2a$ and get the same value for all (x_i, y_i)). Thus we can fix $\|w\|_2 = 1/M$ and obtain

$$\begin{array}{ll} \text{maximize}_{w,a} & 1/\|w\|_2 \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 \quad \text{for all } i. \end{array}$$

To find the optimal w, a we can instead solve the minimization problem

$$\begin{array}{ll} \text{minimize}_{w,a} & \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 \quad \text{for all } i. \end{array}$$

This is a quadratic program that can be solved quickly on fairly large datasets.

Soft Margin SVM

The methods developed thus far require linearly separable data. To remove this restriction, we will allow vectors to violate the margin requirements, but at a penalty. More precisely, we replace our previous SVM formulation

$$\begin{array}{ll} \text{minimize}_{w,a} & \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 \quad \text{for all } i \end{array}$$

with

$$\begin{array}{ll} \text{minimize}_{w,a} & \|w\|_2^2 + \frac{C}{n}\sum_{i=1}^n \xi_i \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 - \xi_i \quad \text{for all } i \\ & \xi_i \geq 0 \quad \text{for all } i. \end{array}$$

This is the standard formulation of a support vector machine, and is equivalent to the statement from the lecture. When $\xi_i > 0$ the corresponding x_i violates the margin condition. The constant C controls how much we penalize violations. Rewriting the condition as

$$\frac{y_i(w^T x_i + a)}{\|w\|_2} \geq \frac{1 - \xi_i}{\|w\|_2}$$

shows that ξ_i measures the size of the violation in multiples of the margin. For example, $\xi_i = 1$ means x_i lies on the decision hyperplane $w^T v + a = 0$, and $\xi_i = 3$ means x_i lies 2 margin widths past the decision hyperplane.

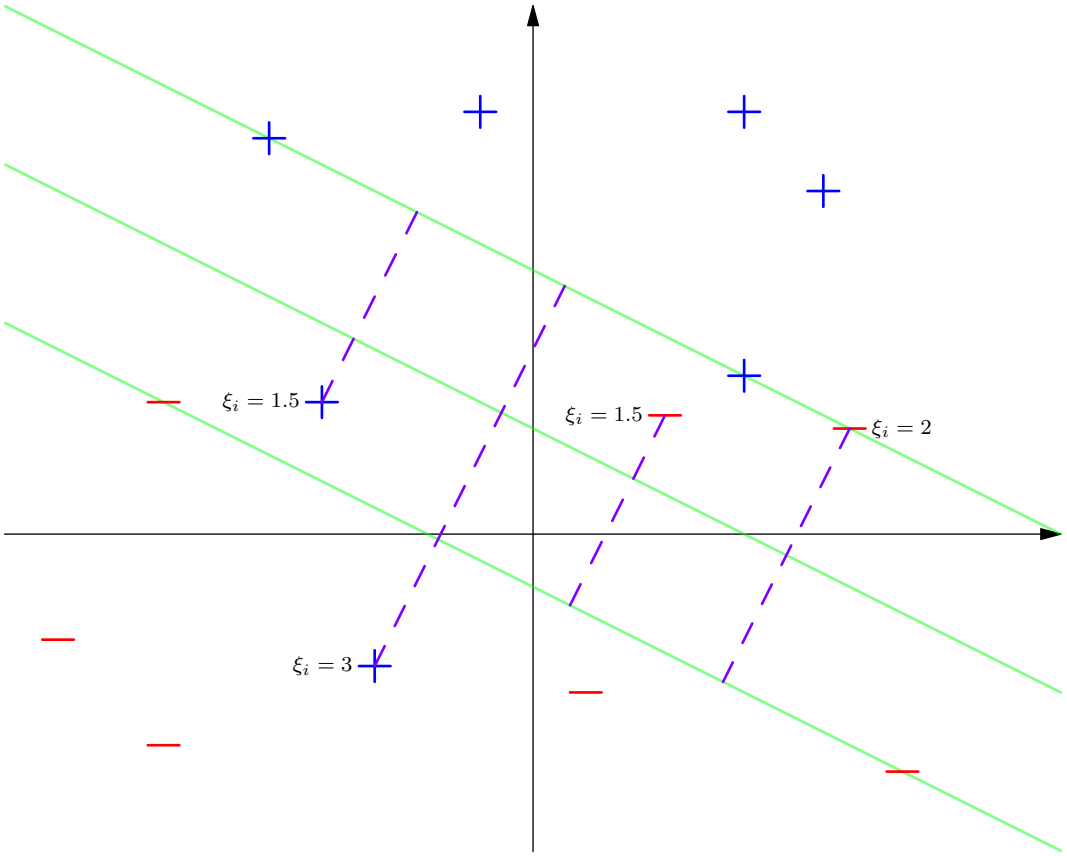

Figure 9: Soft Margin SVM (unlabeled points have $\xi_i = 0$)

Recall from the treatment in class, that the minimizer w will be a linear combination of some of the x_i, called support vectors. More precisely, the support vectors will be some subset of the x_i that either lie on the margin boundary ($y_i(w^T x_i + a) = 1$) or violate the margin boundary ($y_i(w^T x_i + a) < 1$, $\xi_i > 0$).