CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan DIS1

In this discussion, we'll develop some intuition for the Support Vector Machine (SVM) optimization problem,

$$\min_{w,\alpha} \ |w|^2 \text{ subject to } y_i(X_i \cdot w + \alpha) \geq 1, \ \ \forall i \in \{1, \ldots, m\}.$$

1 SVM: Decision Rule

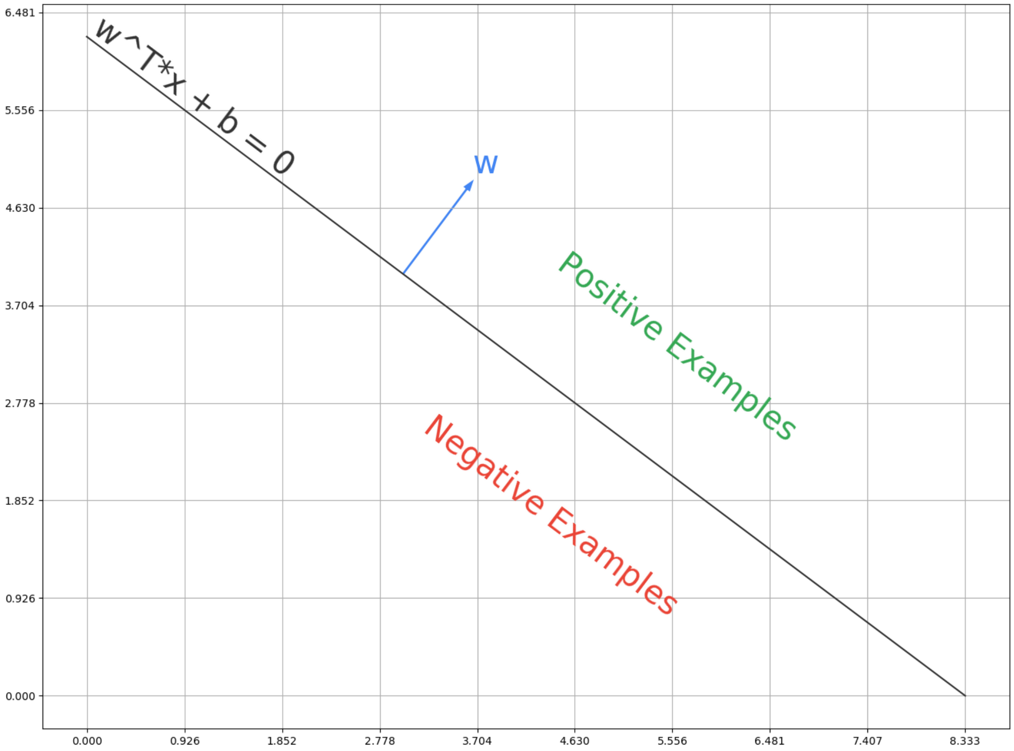
A *decision rule* (or *classifier*) is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to $+1$ ("in class") or -1 ("not in class"). The decision rule for linear SVMs is

$$r(x) = \begin{cases} +1 & \text{if } w \cdot x + \alpha \geq 0, \\ -1 & \text{otherwise,} \end{cases} \tag{1}$$

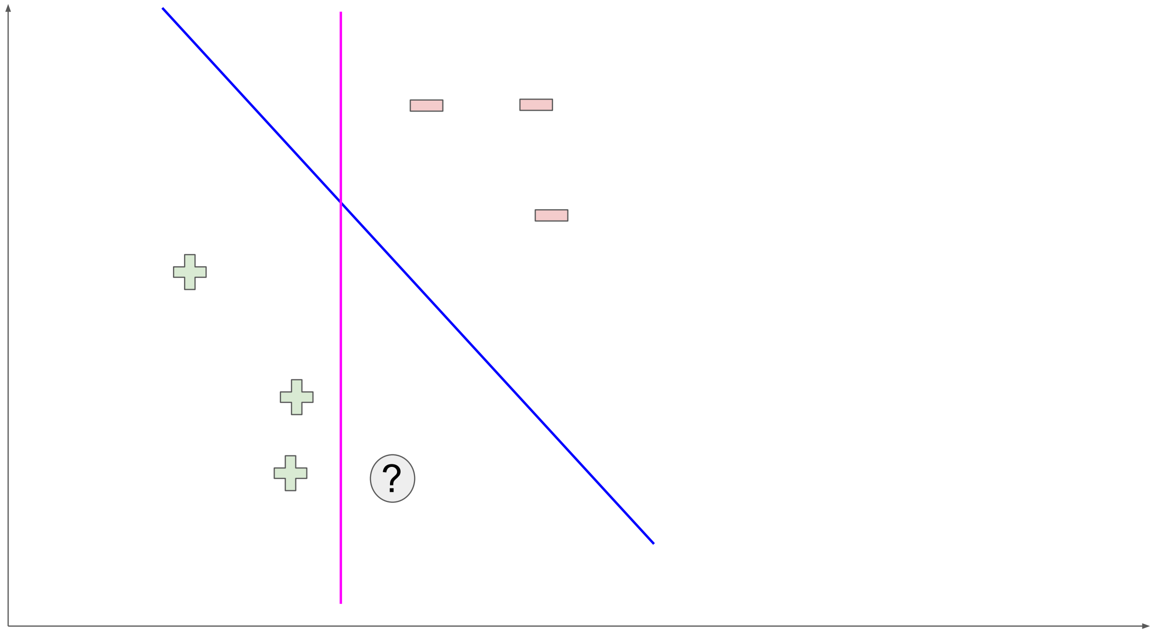
where $w \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ are the weights (parameters) of the SVM.

(a) Draw a figure depicting the line $\ell = \{u \mid u \cdot w + \alpha = 0\}$ with $w = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ and $\alpha = -25$. Include in your figure the vector w, drawn relative to ℓ.

(b) ℓ can be thought of as the decision boundary for a binary classification problem. Indicate in your figure the region in which data points $x \in \mathbb{R}^2$ would be classified as 1. Do the same for data points that would be classified as -1. **Solution:**



NOTE TO INSTRUCTOR: After students work on this problem, draw the following diagram on the white board and discuss: Is the blue line a better decision boundary than the purple line? Why? Both would score correctly classify the labelled data. What class is the unlabelled point more likely to be? Introduce the notion of margin graphically, and talk about the margin sizes of both lines. Which margin is bigger? Develop the intuition that a bigger margin is better.



2 SVM: Constraints

We train SVMs by maximizing the distance of the decision boundary from both positive (1) and negative (−1) examples. The gap between the decision boundary and the closest positive and negative examples is called the margin. We can express the margin requirement by imposing the constraints

$$y_i(X_i \cdot w + \alpha) \geq c, \quad \forall i \in \{1, \ldots, m\}, \tag{2}$$

where c is taken to be the maximum margin.

(a) What role does y_i play in Equation 3? **Solution:** y_i allows us to write a single constraint instead of two separate constraints for positive and negative examples.

(b) The margin $c > 0$ can be rescaled to 1 without affecting the decision rule:

$$y_i(X_i \cdot w + \alpha) \geq 1, \quad \forall i \in \{1, \ldots, m\}. \tag{3}$$

Why can we rescale the margin to 1? Hint: Consider the decision rule $c(u \cdot w + \alpha) \geq 0$. What role does c play in classifying the point u?

Solution: We can set c to any value > 0. The output of the decision rule $(+1, -1)$ would remain the same. Now consider the constraint in Equation 3. Rescaling w and α here changes the position of the decision boundary, and the distance of the margins from the decision boundary without affecting the decision rule. Thus, we are free to rescale w and α so that the margin is 1.

(c) For which examples i is $y_i(X_i \cdot w + \alpha) = 1$? What is the geometric interpretation and significance of these examples? **Solution:** The examples i where $y_i(X_i \cdot w + \alpha) = 1$ are the examples that lie on the margins. The corresponding X_i are called the support vectors.

3 SVM: Objective

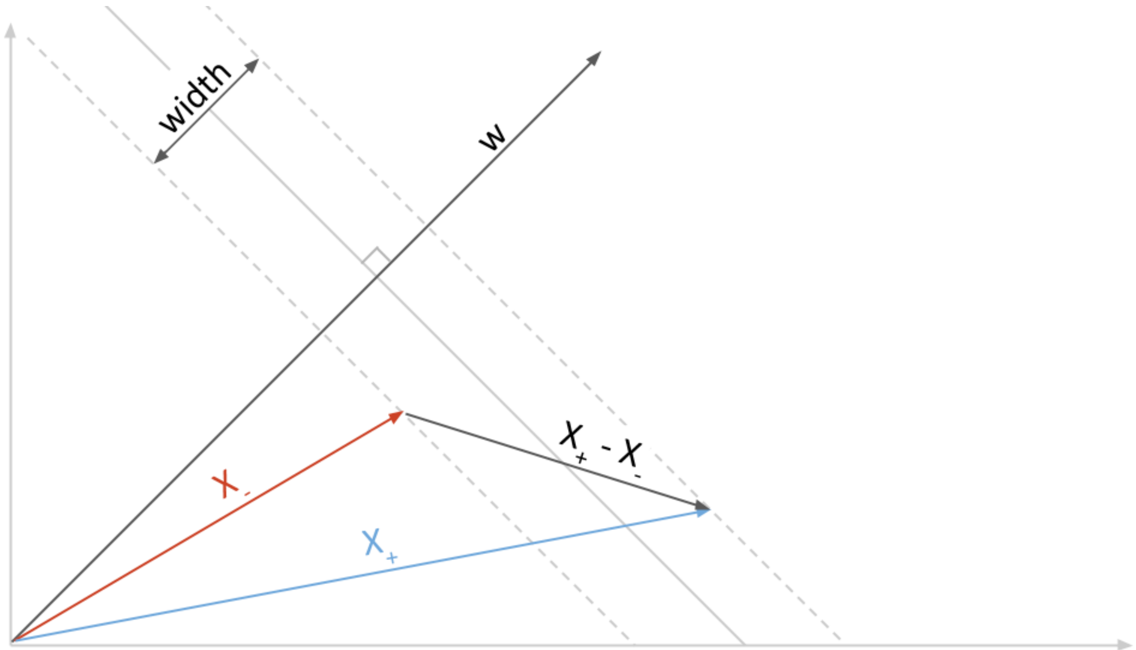


Figure 1: Diagram depicting X_+, X_-, w, and the width of the margins.

The constraints we obtained in the previous problem restrict the possible decision boundaries to those which separate the data with some margin that depends on w and b. We want the maximum possible margin. We'll need an objective we can optimize to obtain a maximum margin in terms of w and b. To obtain this objective, we rewrite Equation 3 as

$$y_i X_i \cdot w \geq 1 - y_i \alpha, \quad i = 1, ..., m. \tag{4}$$

Let X_- and X_+ be negative and positive examples **on the margins**, as depicted in Figure 1. The **width** is the distance from the negative margin to the decision boundary plus the distance from the decision boundary to the positive margin, as shown in Figure 1. We can compute the width in terms of w as follows.

(a) Write down Equation 4 for X_-. Divide through by $|w|$ to obtain a scalar projection of X_- onto $\frac{w}{|w|}$. Do the same for X_+. **Solution:** These are examples on the margin, so we have equality: $\frac{w \cdot X_-}{|w|} = -\frac{1+\alpha}{|w|}$ and $\frac{w \cdot X_+}{|w|} = \frac{1-\alpha}{|w|}$.

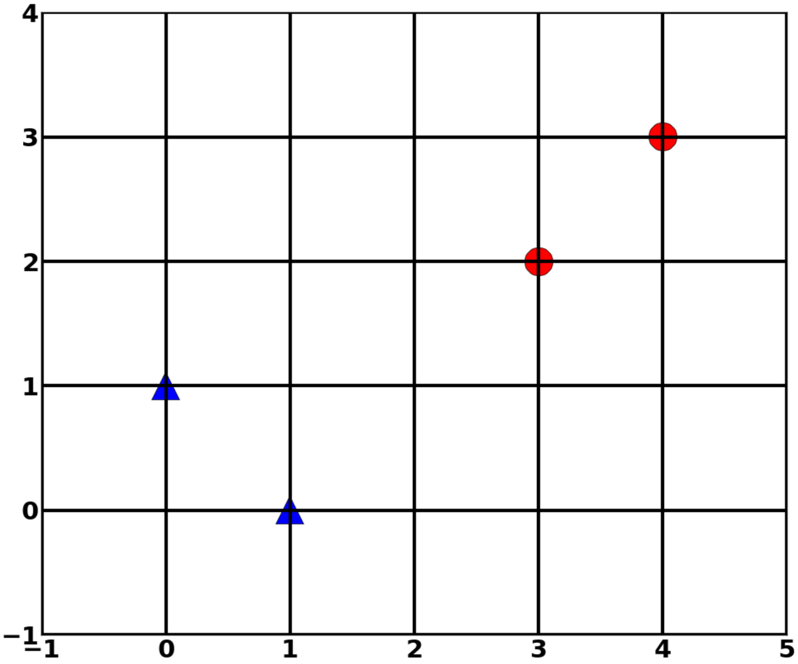
(b) You now have two vectors pointing in the same direction, both on the margins. Compute the width using these two vectors to obtain $\frac{2}{|w|}$. **Solution:** Subtracting yields $\frac{1-\alpha}{|w|} + \frac{1+\alpha}{|w|} = \frac{2}{|w|}$.

(c) Explain in words why we want to maximize $\frac{2}{|w|}$. **Solution:** This objective is the width in terms of w, so we are in fact maximizing the margin.

(d) Show that $\max_{w,b} \frac{2}{|w|}$ can be rewritten as $\min_{w,b} \frac{1}{2}|w|^2$. **Solution:** Since $\frac{2}{|w|} \geq 0$, $\max_{w,b} \frac{2}{|w|} = \min_{w,b} \frac{|w|}{2}$. Squaring simplifies the objective without changing the problem.

4 SVM: Hyperplane Exercise

(a) You are presented with the following set of data (triangle = +1, circle = -1):



Find the equation (by hand) of the hyperplane $w^T x + b = 0$ that would be used by an SVM classifier.

Solution:

Solving linearly separable classification problem is equivalent to finding the points of the two convex hulls which are closest to each other and the maximum margin hyperplane bisects and is normal to the line segment joining these two closest points.

In our case the convex hulls are the line segment joining the negative points and the line segment joining positive points, so the two points that are closest are $(3, 2)$ and $(1, 0)$. From the above we get that the equation of the hyperplane will pass through point (2, 1), with a slope of -1. The equation of this line is $x_1 + x_2 = 3$.

From the fact that $\mathbf{w}^\top \mathbf{x} + b = 0$ we know that $w_1 = w_2$. So we only need two more equations to solve for w_1, w_2 and b.

We also know that for points on the margins, the hyperlanes are $\mathbf{w}^\top \mathbf{x} + b = \pm 1$. We know that $(1, 0)$ is on the ”positive” hyperplane (as well as $(0, 1)$ in this case) and $(3, 2)$ is on the ”negative” hyperplane, we get the following system of equations:

$$\begin{cases} 1w_1 + 0w_2 + b = 1 \\ 3w_1 + 2w_2 + b = -1 \\ w_1 = w_2 \end{cases}$$

Solving this system of equations, we get $\mathbf{w} = [-\frac{1}{2}, -\frac{1}{2}]^\top$ and $b = \frac{3}{2}$.