CS 189 Introduction to Machine Learning
Spring 2022 Marvin Zhang

DIS4

1 SVMs: Step by Step

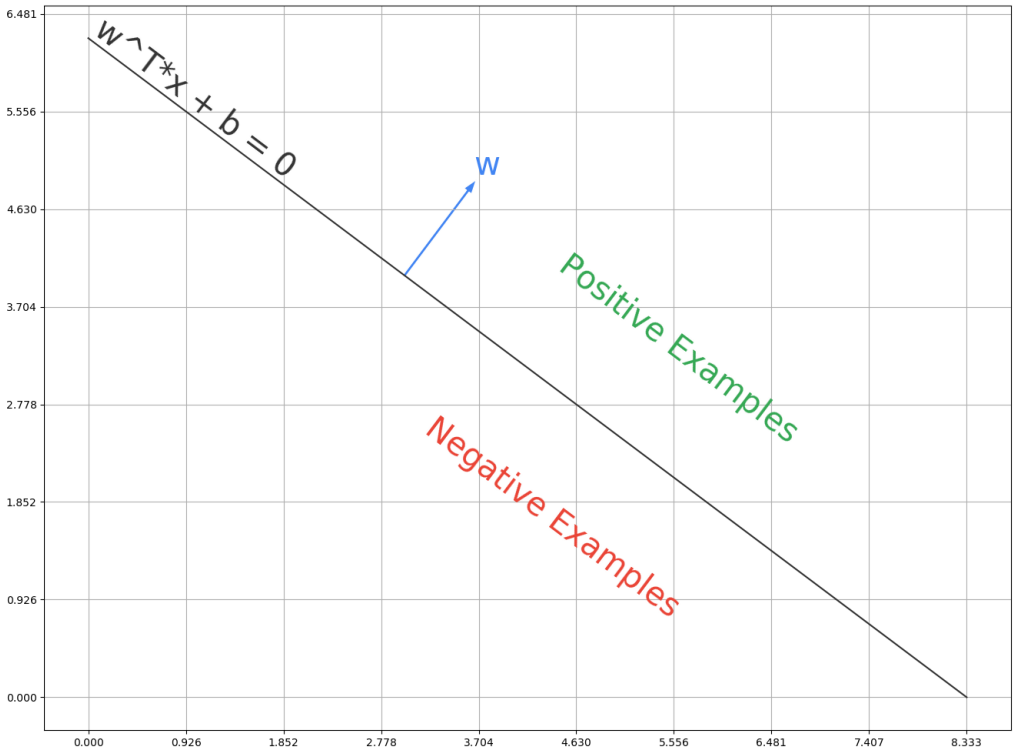
A *decision rule* (or *classifier*) is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to $+1$ or -1. The decision rule for SVMs is

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

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

(a) **Draw a figure depicting the line $\ell = \{\mathbf{u} \mid \mathbf{u}^\top \mathbf{w} + b = 0\}$ with $\mathbf{w} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ and $b = -25$.** Include in your figure the vector w, drawn relative to ℓ. Indicate in your figure the region in which data points $x \in \mathbb{R}^2$ would be classified as 1 vs -1.

Solution:



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.

Recall from homework that the *signed distance* of a point from a hyperplane defined by $H = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + b = 0\}$ is equal to:

$$\text{SignedDistance}(\mathbf{x}, H) = \frac{\mathbf{w}^\top \mathbf{x} + b}{\|\mathbf{w}\|_2}$$

If we are given c as the maximum margin width from the decision boundary to the nearest point and $\mathbf{w}$ as a *unit vector* defining the decision boundary hyperplane, then we can express the margin requirement (all points must lie on the correct side of the boundary) with the constraints:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq c, \quad \forall i \in \{1, \ldots, n\}, \tag{2}$$

(b) **What role does y_i play in Equation 2?**

Solution: y_i allows us to write a single constraint instead of two separate constraints for positive and negative examples.

(c) The maximum margin width $c > 0$ is something we're trying to find, so we can't plug it into our constraints right away. However, one thing to realize is that the constraint can be rescaled to 1 without affecting the decision rule:

$$y_i(\mathbf{w}'^\top \mathbf{x} + b') \geq 1, \quad \forall i \in \{1, \ldots, n\}. \tag{3}$$

What is $\mathbf{w}', b'$ in terms of $\mathbf{w}, b$? Why can we rescale right hand side to 1?

Solution:

$$y_i(\mathbf{w}^\top \mathbf{x} + b) \geq c$$
$$y_i\left(\frac{\mathbf{w}'^\top}{c} \mathbf{x} + \frac{b'}{c}\right) \geq 1$$
$$\implies \mathbf{w}' = \frac{1}{c}\mathbf{w} \qquad b' = \frac{1}{c}b$$

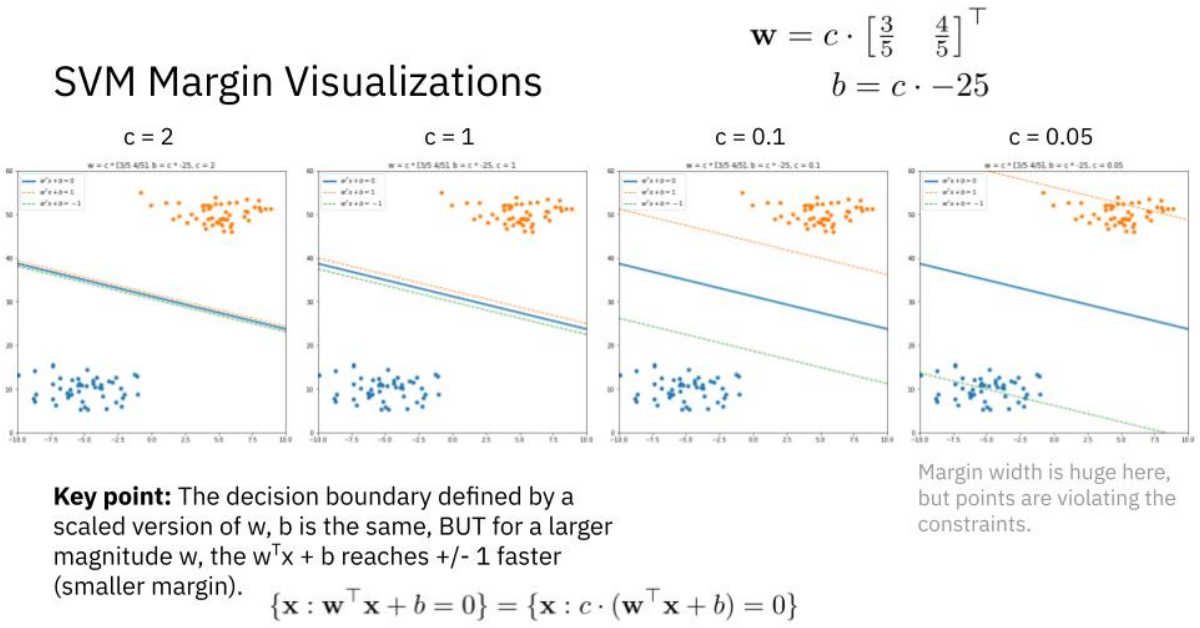
We are going to be running an optimizer to find the values of $\mathbf{w}', b'$, so we don't ever find the unit vector $\mathbf{w}$ and then doing this scaling by hand. The purpose of this question is to show that allowing the $\mathbf{w}'$ to be something other than a unit vector allows us to simplify the margin constraint while maintaining the same decision boundary.

With the new scaled-down constraint, the decision boundary, top margin, and bottom margin are:

$$\text{Decision Boundary} = \{\mathbf{x} : \mathbf{w}'^\top \mathbf{x} + b' = \frac{1}{c}\mathbf{w}^\top \mathbf{x} + b = \mathbf{w}^\top \mathbf{x} + b = 0\}$$
$$\text{Top Margin} = \{\mathbf{x} : \mathbf{w}'^\top \mathbf{x} + b' = +1\}$$
$$\text{Bottom Margin} = \{\mathbf{x} : \mathbf{w}'^\top \mathbf{x} + b' = -1\}$$

Notice that the decision boundary after scaling down the parameters doesn't actually change!

The width of the margin is defined by the distance between the top and bottom margins. This is now a function magnitude of the weight vector $\mathbf{w}'$. Intuitively, a $\mathbf{w}'$ with a large magnitude has the linear function $\mathbf{w}'^{\top}\mathbf{x} + b'$ go quickly from -1 to $+1$, which corresponds to a tighter margin. A $\mathbf{w}'$ with a small magnitude has a "flatter" function that goes from -1 to $+1$ much more slowly. This corresponds to a wide margin. See the following figure of the margins given various scalings of the same unit vector direction.



(d) **For which examples is** $y_i(\mathbf{w}^{\top}\mathbf{x}_i + b) = 1$**?** What is the geometric interpretation and significance of these examples?

Solution: The examples i where $y_i(X_i \cdot w + b) = 1$ are the examples that lie on the margins. The corresponding X_i are called the support vectors.

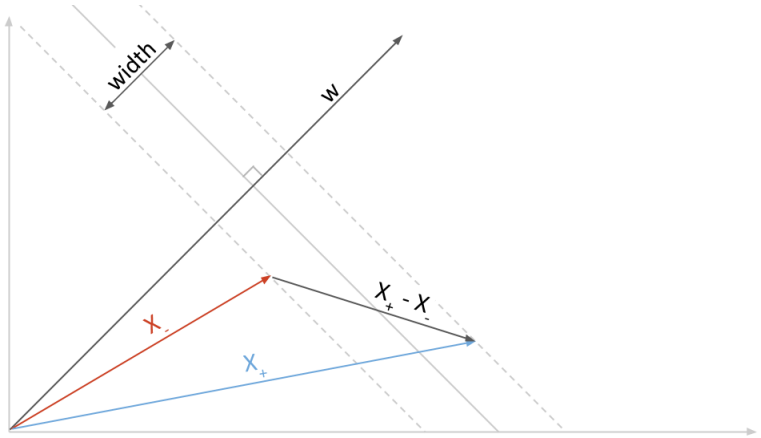


Figure 1: Diagram depicting $\mathbf{x}_+$, $\mathbf{x}_-$, $\mathbf{w}$, and the width of the margins.

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 $\mathbf{w}$ and b. We want the maximum possible margin. We'll need an objective we can optimize to obtain a maximum margin in terms of $\mathbf{w}$ and b. To obtain this objective, we rewrite Equation 3 as

$$y_i \mathbf{w}^\top \mathbf{x}_i \geq 1 - y_i b, \quad i \in \{1, \ldots, n\} \tag{4}$$

Let $\mathbf{x}_-$ and $\mathbf{x}_+$ 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 find the margin width by computing the magnitude of the projection of $\mathbf{x}_+ - \mathbf{x}_-$ onto the $\mathbf{w}$ direction. Recall the formula for the projection of one vector $\mathbf{u}$ onto another vector $\mathbf{v}$:

$$\text{proj}_\mathbf{v} \mathbf{u} = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{v}\|_2} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$$

(e) **Write out the projection** $\text{proj}_\mathbf{w}(\mathbf{x}_+ - \mathbf{x}_-) = \text{proj}_\mathbf{w} \mathbf{x}_+ - \text{proj}_\mathbf{w} \mathbf{x}_-$**.**

Draw this projection as a vector in Figure 1. What is its direction and magnitude?

Solution:

$$\begin{aligned} \text{proj}_\mathbf{w} \mathbf{x}_+ - \text{proj}_\mathbf{w} \mathbf{x}_- &= \frac{\mathbf{x}_+^\top \mathbf{w}}{\|\mathbf{w}\|_2} \frac{\mathbf{w}}{\|\mathbf{w}\|_2} - \frac{\mathbf{x}_-^\top \mathbf{w}}{\|\mathbf{w}\|_2} \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \\ &= \frac{(\mathbf{x}_+^\top \mathbf{w} - \mathbf{x}_-^\top \mathbf{w})}{\|\mathbf{w}\|_2} \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \end{aligned}$$

This is a vector pointing in the same direction as $\mathbf{w}$, with a magnitude equal to the width between the margins.

(f) **Write down Equation 4 for** $\mathbf{x}_+$ **and** $\mathbf{x}_-$**.**

Use it to substitute a value for $\mathbf{x}_+^\top \mathbf{w}$ and $\mathbf{x}_-^\top \mathbf{w}$ into the projection you found in the previous part.

Solution: These are examples on the margin, so we have equality.

For $\mathbf{x}_-$:

$$(-1)(\mathbf{x}_-^\top \mathbf{w}) = 1 - (-1) \cdot b$$
$$\implies \mathbf{x}_-^\top \mathbf{w} = -(1 + b)$$

For $\mathbf{x}_+$:

$$(+1)(\mathbf{x}_+^\top \mathbf{w}) = 1 - (+1) \cdot b$$
$$\implies \mathbf{x}_+^\top \mathbf{w} = 1 - b$$

Plugging into the solution from the previous part:

$$\frac{1 - b - (-(1 + b))}{\|\mathbf{w}\|_2} \frac{\mathbf{w}}{\|\mathbf{w}\|_2} = \frac{2}{\|\mathbf{w}\|_2} \frac{\mathbf{w}}{\|\mathbf{w}\|_2}$$

(g) **Now, take the magnitude of this projection to calculate the margin width in terms of w.**

Solution:

$$\left\| \frac{2}{\|\mathbf{w}\|_2} \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \right\|_2 = \frac{2}{\|\mathbf{w}\|_2}$$

Thus, the margin width is equal to $\frac{2}{\|\mathbf{w}\|_2}$.

(h) We want to maximize the margin width. Using the answer from the previous part, **show how the SVM objective can be written as** $\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$**.**

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.

2 SVM with custom margins

In the lecture, we covered the soft-margin SVM. The objective to be optimized over the training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ is

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{5}$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \xi_i \quad \forall i \tag{6}$$

$$\xi_i \geq 0 \quad \forall i \tag{7}$$

In this problem, we are interested in a modified version of the soft-margin SVM where we have a custom margin for each of the n data points. In the standard soft-margin SVM, we pay a penalty of ξ_i for each of the data point. In practice, we might not want to treat each training point equally – for example, we might know that some data points are more important than the others.

We formally define the following optimization problem:

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \phi_i\xi_i \tag{8}$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \xi_i \quad \forall i \tag{9}$$

$$\xi_i \geq 0 \quad \forall i \tag{10}$$

Note that the only difference is that we have a custom weighting factor $\phi_i > 0$ for each of the slack variables ξ_i in the objective function. These ϕ_i are some constants given by the prior knowledge, and thus they can be treated as known constants in the optimization problem. Intuitively, this formulation weights each of the violations (ξ_i) differently according to the prior knowledge (ϕ_i).

(a) For the standard soft-margin SVM, we have shown that the constrained optimization problem is equal to the following unconstrained optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i - b)) \tag{11}$$

What's the corresponding unconstrained optimization problem for the SVM with custom margins?

Solution: The corresponding unconstrained optimization problem is

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \phi_i \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i - b)). \tag{12}$$

We can see this as follows. Manipulating the first inequality, we have that

$$\xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i - b) \quad \forall i. \tag{13}$$

Combining this with the second inequality, we have that

$$\xi_i \geq \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i - b), 0). \tag{14}$$

Since we are minimizing and since we know that $\phi_i > 0$ for all i, we conclude that the constraint must be tight:

$$\xi_i = \max(1 - y_i(\mathbf{w}^\top \mathbf{x}_i - b), 0). \tag{15}$$

The above unconstrained problem then follows when we substitute for ξ_i.

(b) As seen in lecture, the dual form of the standard soft-margin SVM is:

$$\max_{\alpha} \quad \alpha^\top \mathbf{1} - \frac{1}{2}\alpha^\top \mathbf{Q}\alpha \tag{16}$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{17}$$

$$0 \leq \alpha_i \leq C \quad i = 1, \cdots, n \tag{18}$$

where $\mathbf{Q} = (\operatorname{diag} \mathbf{y})\mathbf{X}\mathbf{X}^T(\operatorname{diag} \mathbf{y})$.

What's the dual form of the SVM with custom margins? To start, we provide you the Lagrangian, which is given by

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i(\mathbf{w}^\top \mathbf{x}_i - b) + \sum_{i=1}^{n} (C\phi_i - \alpha_i - \beta_i)\xi_i \tag{19}$$

Solution:

The optimization we want to solve is

$$\max_{\alpha \geq 0, \beta \geq 0} \min_{\mathbf{w}, b, \xi_i} \mathcal{L}(\mathbf{w}, \mathbf{b}, \xi, \alpha, \beta).$$

We know that the KKT conditions hold. Thus, we first set the gradients with respect to $\mathbf{w}$, b, and ξ_i equal to 0 to get

$$\nabla_{\mathbf{w}} \mathcal{L} = \mathbf{w}^* - \sum_{i=1}^{n} \alpha_i^* y_i x_i = 0 \implies \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$

$$\nabla_b \mathcal{L} = \sum_{i=1}^{n} \alpha_i^* y_i = 0.$$

$$\nabla_{\xi_i} \mathcal{L} = C\phi_i - \alpha_i^* - \beta_i^* = 0 \quad i = 1, \ldots, n. \tag{20}$$

The last equality relates α to β. Since α, β are restricted to being greater than or equal to 0, the last equality also implies that $\alpha_i^* \leq C\phi_i$. Now using the equations above, we can simplify the Lagrangian to

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha^*, \beta^*) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i^* - \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{w}^\top \mathbf{x}_i.$$

Plugging in $\mathbf{w}^*$, we get

$$\begin{aligned}
\mathcal{L}(\mathbf{w}^*, b, \xi, \alpha^*, \beta^*) &= \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i^* y_i x_i\|^2 + \sum_{i=1}^{n} \alpha_i^* - \sum_{i=1}^{n} \alpha_i^* y_i (\sum_{j=1}^{n} \alpha_j^* y_j x_j)^\top \mathbf{x}_i && (21)\\
&= \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i^* y_i x_i\|^2 + \sum_{i=1}^{n} \alpha_i^* - \sum_{i=1}^{n} (\sum_{j=1}^{n} \alpha_j^* y_j x_j)^\top (\alpha_i^* y_i \mathbf{x}_i) && (22)\\
&= \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i^* y_i x_i\|^2 + \sum_{i=1}^{n} \alpha_i^* - \| \sum_{i=1}^{n} \alpha_i^* y_i x_i\|^2 && (23)\\
&= \sum_{i=1}^{n} \alpha_i^* - \frac{1}{2}\| \sum_{i=1}^{n} \alpha_i^* y_i x_i\|^2 && (24)\\
&= \mathbf{1}^\top \alpha^* - \frac{1}{2}\alpha^* \mathbf{Q} \alpha^* && (25)
\end{aligned}$$

where we let $\mathbf{Q} = (\operatorname{diag} \mathbf{y})\mathbf{X}\mathbf{X}^T(\operatorname{diag} \mathbf{y})$.

Incorporating the previous constraints for obtained via taking gradients with respect to b and ξ_i, we then get the dual problem is

$$\begin{aligned}
\max_{\alpha} \quad & \alpha^\top \mathbf{1} - \frac{1}{2}\alpha^\top \mathbf{Q} \alpha && (26)\\
s.t. \quad & \alpha^\top \mathbf{y} = 0 && (27)\\
& 0 \leq \alpha_i \leq C\phi_i \quad i = 1, \ldots, n && (28)
\end{aligned}$$