CS 189/289A Introduction to Machine Learning
Fall 2021 Jennifer Listgarten and Jitendra Malik

DIS8

1 SVMs: Step by Step

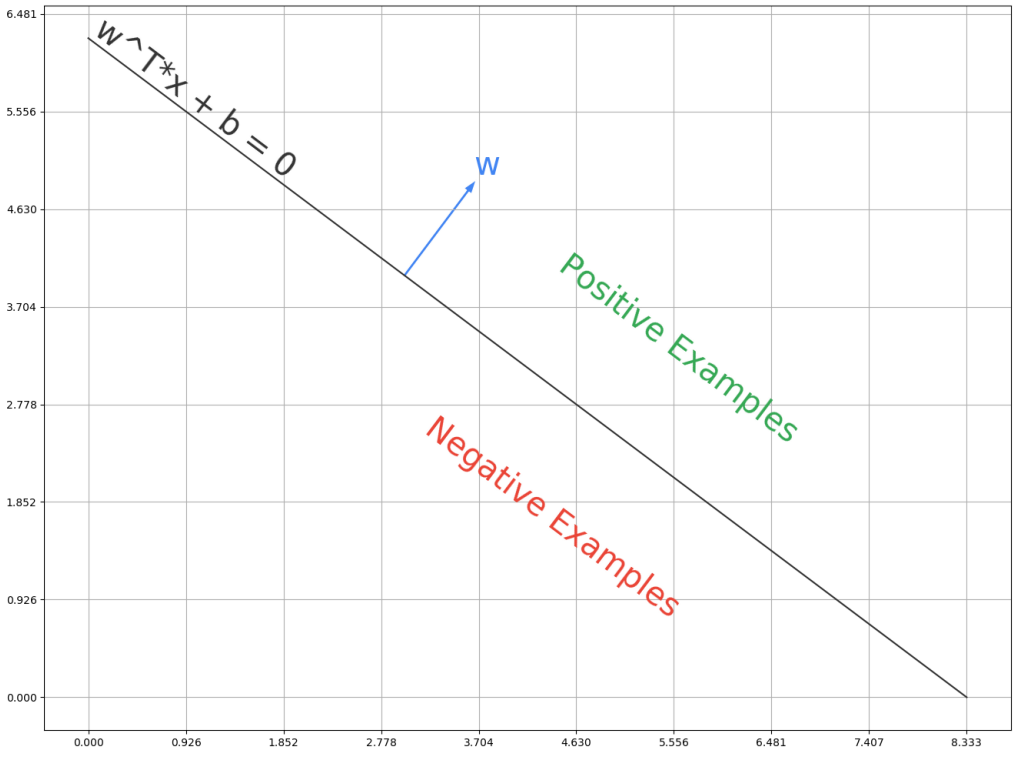
A *decision rule* (or *classifier*) is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to $+1$ or -1. The decision rule for SVMs is

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

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. 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. 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.

(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 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?

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.

(d) 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.

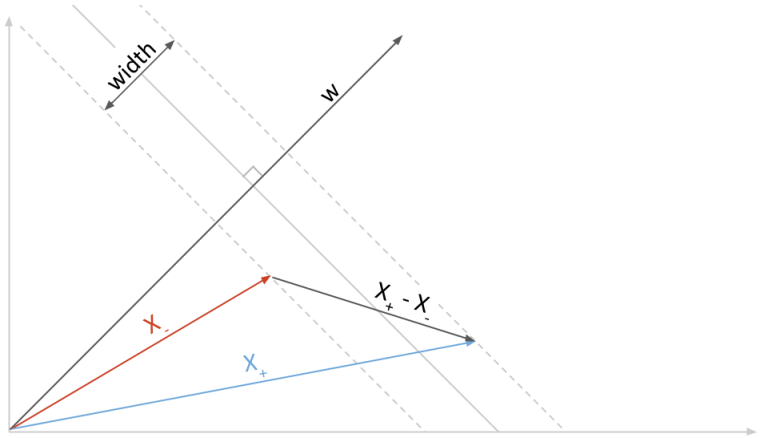


Figure 1: Diagram depicting X_+, X_-, w, and the width of the margins.

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.

(e) 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 $\frac{w \cdot X_+}{|w|} = \frac{1-\alpha}{|w|}$.

(f) You now have the distance between the decision boundary to each of the margins. What is the width?

Solution: Subtracting yields $\frac{1-\alpha}{|w|} + \frac{1+\alpha}{|w|} = \frac{2}{|w|}$.

(g) We want to maximize the width. Using the answer from the previous part, show how this can be written 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.

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

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