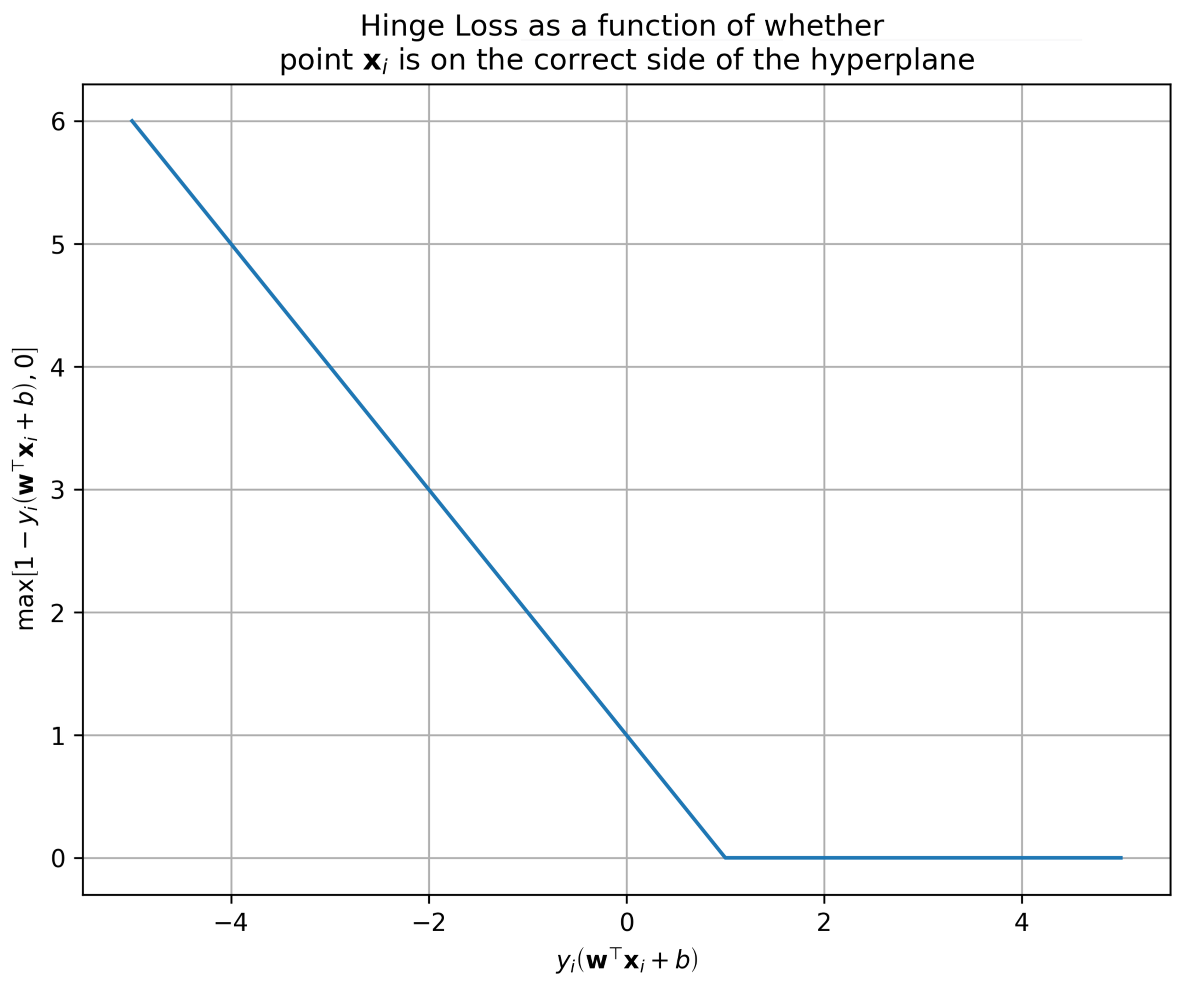
Compute Gradient of Loss Functions

Recall the optimization problem we are trying to solve:

$$\min_{\mathbf{w},b} \underbrace{\mathbf{w}^\top \mathbf{w}}_{l_2-\text{regularizer}} + \underbrace{C\sum_{i=1}^n \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]}_{\text{hinge loss}}$$

Observe that the hinge loss is not smooth (mathematically, not differentiable at $y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) = 1$) as a function of whether points are on the correct side of the hyperplane (mathematically, $y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)$). You can see this from the plot of hinge loss below as well; there is a kink in the plot at 1.



Consequently, the gradient will be undefined at 1. Although there are ways around this, one simple trick is to minimize the square hinge loss instead:

$$\min_{\mathbf{w},b} \underbrace{\mathbf{w}^\top \mathbf{w}}_{l_2-\text{regularizer}} + \underbrace{C\sum_{i=1}^n \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]^2}_{\text{squared hinge loss}}$$

Note that the squared hinge-loss doesn't quite exactly return the maximum margin hyperplane, but the solution is very close.

The Gradient of the Linear SVM Loss Function

Let the loss function be

$$\ell(\mathbf{w}, b) = \mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^n \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]^2$$

Find the gradients $\nabla_{\mathbf{w}}\ell$ and $\nabla_b \ell$ of the loss function ℓ.

The gradients will depend on whether the **max** function evaluates to $1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)$ or 0 for each data point $(\mathbf{x}_i, y_i)$. So we recommend calculating the gradient using case functions δ_i that are either equal to $1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)$ or 0. Alternatively, you can also use indicator functions $\mathbf{1}_{1-y_i(\mathbf{w}^\top \mathbf{x}_i + b)>0}$ if you know about them.

When you are done solving this problem, click the button below to check your answer.

Hide Solution

Answer:

To compute the gradients, we will use the case functions δ_i defined by

$$\delta_i = \begin{cases} 1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) & \text{if } 1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

for each $i \in \{1, \ldots, n\}$. With this notation, the gradients are:

$$\begin{aligned} \nabla_{\mathbf{w}}\ell &= \frac{\partial\, \mathbf{w}^\top \mathbf{w}}{\partial \mathbf{w}} + C\sum_{i=1}^n \frac{\partial \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]^2}{\partial \mathbf{w}} \\ &= 2\mathbf{w} + C\sum_{i=1}^n 2\delta_i \cdot \frac{\partial\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right)}{\partial \mathbf{w}} \\ &= 2\mathbf{w} + C\sum_{i=1}^n 2\delta_i \cdot (-y_i \mathbf{x}_i) \\ &= 2\mathbf{w} - C\sum_{i=1}^n 2\delta_i \cdot y_i \mathbf{x}_i \end{aligned}$$

and

$$\begin{aligned} \nabla_b \ell &= \frac{\partial\, \mathbf{w}^\top \mathbf{w}}{\partial b} + C\sum_{i=1}^n \frac{\partial \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]^2}{\partial b} \\ &= C\sum_{i=1}^n 2\delta_i \cdot \frac{\partial\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right)}{\partial b} \\ &= C\sum_{i=1}^n 2\delta_i \cdot (-y_i) \\ &= -C\sum_{i=1}^n 2\delta_i \cdot y_i \end{aligned}$$

Alternatively, you can use indicator functions defined by

$$\mathbf{1}_{1-y_i(\mathbf{w}^\top \mathbf{x}_i + b)>0} = \begin{cases} 1 & \text{if } 1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

for each $i \in \{1, \ldots, n\}$ to get the following gradients:

$$\nabla_{\mathbf{w}}\ell = 2\mathbf{w} - C\sum_{i=1}^n 2\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right) \cdot \mathbf{1}_{1-y_i(\mathbf{w}^\top \mathbf{x}_i + b)>0} \cdot y_i \mathbf{x}_i$$

$$\nabla_b \ell = -C\sum_{i=1}^n 2\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right) \cdot \mathbf{1}_{1-y_i(\mathbf{w}^\top \mathbf{x}_i + b)>0} \cdot y_i$$

Note that for each $i \in \{1, \ldots, n\}$,

$$\begin{aligned} \delta_i &= \left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right)\right) \cdot \mathbf{1}_{1-y_i(\mathbf{w}^\top \mathbf{x}_i + b)>0} \\ &= \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right] \end{aligned}$$

so the gradients are exactly the same as before. Cleaned up a bit,

$$\nabla_{\mathbf{w}}\ell = 2\mathbf{w} - 2C\sum_{i=1}^n y_i \mathbf{x}_i \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]$$

$$\nabla_b \ell = -2C\sum_{i=1}^n y_i \max\left[1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right]$$