



Course Shortcuts

> Student Lounge

> Q&A

Find Maximum Margin Classifiers

> Module Introduction: Find Maximum Margin Classifiers

Notations for Machine

Learning Find the Maximum Margin

Formalize Maximum

Classifiers

Margin Classifiers Simplify the Optimization

Problem

Explore Maximum Margin Classifiers

Optimize SVMs

Use Slack Variables in

Optimization Problems

**Explore Slack Variables** 

Visualize a Linear SVM

Removing Constraints

Problem Soft-SVM Unconstrained Formulation

from Optimization

Linear SVM Cheat Sheet

Compute Gradient of

**Loss Functions** Make SVM Output

Interpretable With Platt <u>Scaling</u>

Simplify Multi-Class SVMs **Build a Linear SVM** 

Module Wrap-up: Find Maximum Margin Classifiers

Minimize Empirical Risk

> Reduce Bias With Kernelization

> Course Resources

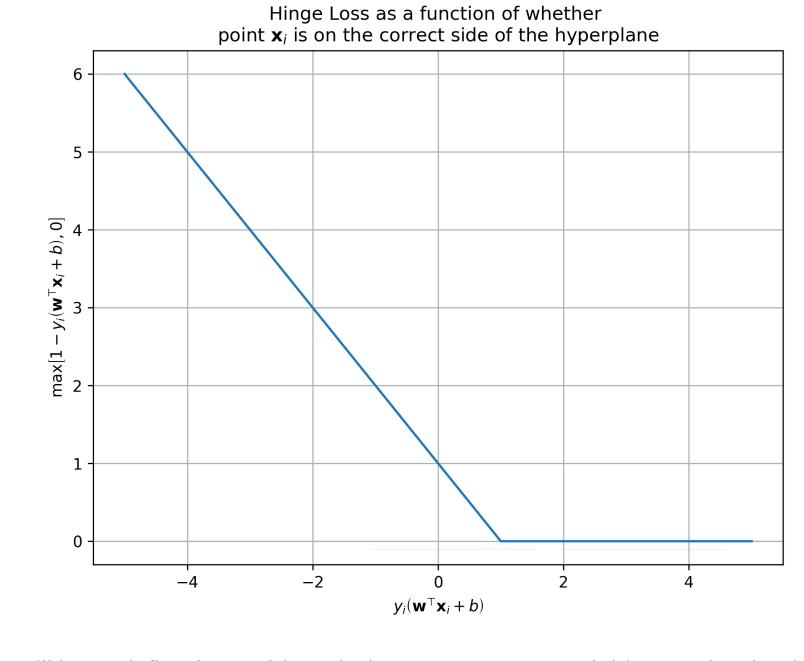
Learning with Kernel Machines >

## **Compute Gradient of Loss Functions**

Recall the optimization problem we are trying to solve:

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

Observe that the hinge loss is not smooth (mathematically, not differentiable at  $y_i$  ( $\mathbf{w}^ op \mathbf{x}_i + b$ ) = 1) as a function of whether points are on the correct side of the hyperplane (mathematically,  $y_i$  ( $\mathbf{w}^{ op}\mathbf{x}_i+b$ ). 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:

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

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

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

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

The gradients will depend on whether the  $\max$  function evaluates to  $1-y_i\left(\mathbf{w}^{ op}\mathbf{x}_i+b\right)$  or 0 for each data point  $(\mathbf{x}_i,y_i)$ . So we recommend calculating the gradient using case functions  $\delta_i$  that are either equal to  $1-y_i\left(\mathbf{w}^ op\mathbf{x}_i+b
ight)$  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  $\delta_i$  defined by

$$\delta_i = egin{cases} 1 - y_i \left( \mathbf{w}^ op \mathbf{x}_i + b 
ight) & ext{if } 1 - y_i \left( \mathbf{w}^ op \mathbf{x}_i + b 
ight) > 0 \ & ext{otherwise} \end{cases}$$

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

$$egin{aligned} 
abla_{\mathbf{w}}\ell &= rac{\partial \, \mathbf{w}^ op \mathbf{w}}{\partial \mathbf{w}} + C \sum_{i=1}^n rac{\partial \max \left[ 1 - y_i \left( \mathbf{w}^ op \mathbf{x}_i + b 
ight), 0 
ight]^2}{\partial \mathbf{w}} \ &= 2\mathbf{w} + C \sum_{i=1}^n 2\delta_i \cdot rac{\partial \left( 1 - y_i \left( \mathbf{w}^ op \mathbf{x}_i + b 
ight) 
ight)}{\partial \mathbf{w}} \ &= 2\mathbf{w} + C \sum_{i=1}^n 2\delta_i \cdot \left( -y_i \mathbf{x}_i 
ight) \ &= 2\mathbf{w} - C \sum_{i=1}^n 2\delta_i \cdot y_i \mathbf{x}_i \end{aligned}$$

and

$$egin{aligned} 
abla_b \ell &= rac{\partial \, \mathbf{w}^ op \mathbf{w}}{\partial b} + C \sum_{i=1}^n rac{\partial \max \left[ 1 - y_i \left( \mathbf{w}^ op \mathbf{x}_i + b 
ight), 0 
ight]^2}{\partial b} \ &= C \sum_{i=1}^n 2 \delta_i \cdot rac{\partial \left( 1 - y_i \left( \mathbf{w}^ op \mathbf{x}_i + b 
ight) 
ight)}{\partial b} \ &= C \sum_{i=1}^n 2 \delta_i \cdot (-y_i) \ &= -C \sum_{i=1}^n 2 \delta_i \cdot y_i \end{aligned}$$

Alternatively, you can use indicator functions defined by

$$\mathbf{1}_{1-y_i(\mathbf{w}^ op \mathbf{x}_i+b)>0} = egin{cases} 1 & ext{if } 1-y_i\left(\mathbf{w}^ op \mathbf{x}_i+b
ight)>0 \ 0 & ext{otherwise} \end{cases}$$
ing gradients:

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

$$egin{aligned} 
abla_{\mathbf{w}}\ell &= 2\mathbf{w} - C\sum_{i=1}^n 2\left(1 - y_i\left(\mathbf{w}^ op \mathbf{x}_i + b
ight)
ight) \cdot \mathbf{1}_{1 - y_i\left(\mathbf{w}^ op \mathbf{x}_i + b
ight) > 0} \cdot y_i \mathbf{x}_i \ 
abla_b\ell &= -C\sum_{i=1}^n 2\left(1 - y_i\left(\mathbf{w}^ op \mathbf{x}_i + b
ight)
ight) \cdot \mathbf{1}_{1 - y_i\left(\mathbf{w}^ op \mathbf{x}_i + b
ight) > 0} \cdot y_i \end{aligned}$$

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

◆ Previous

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

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

$$egin{aligned} 
abla_{\mathbf{w}}\ell &= 2\mathbf{w} - 2C\sum_{i=1}^n y_i\mathbf{x}_i \max\left[1 - y_i\left(\mathbf{w}^ op\mathbf{x}_i + b
ight), 0
ight] \ 
abla_b\ell &= -2C\sum_{i=1}^n y_i \max\left[1 - y_i\left(\mathbf{w}^ op\mathbf{x}_i + b
ight), 0
ight] \end{aligned}$$