Support Vector Machines

Machine Learning

THE
UNIVERSITY
OF UTAH

Big picture

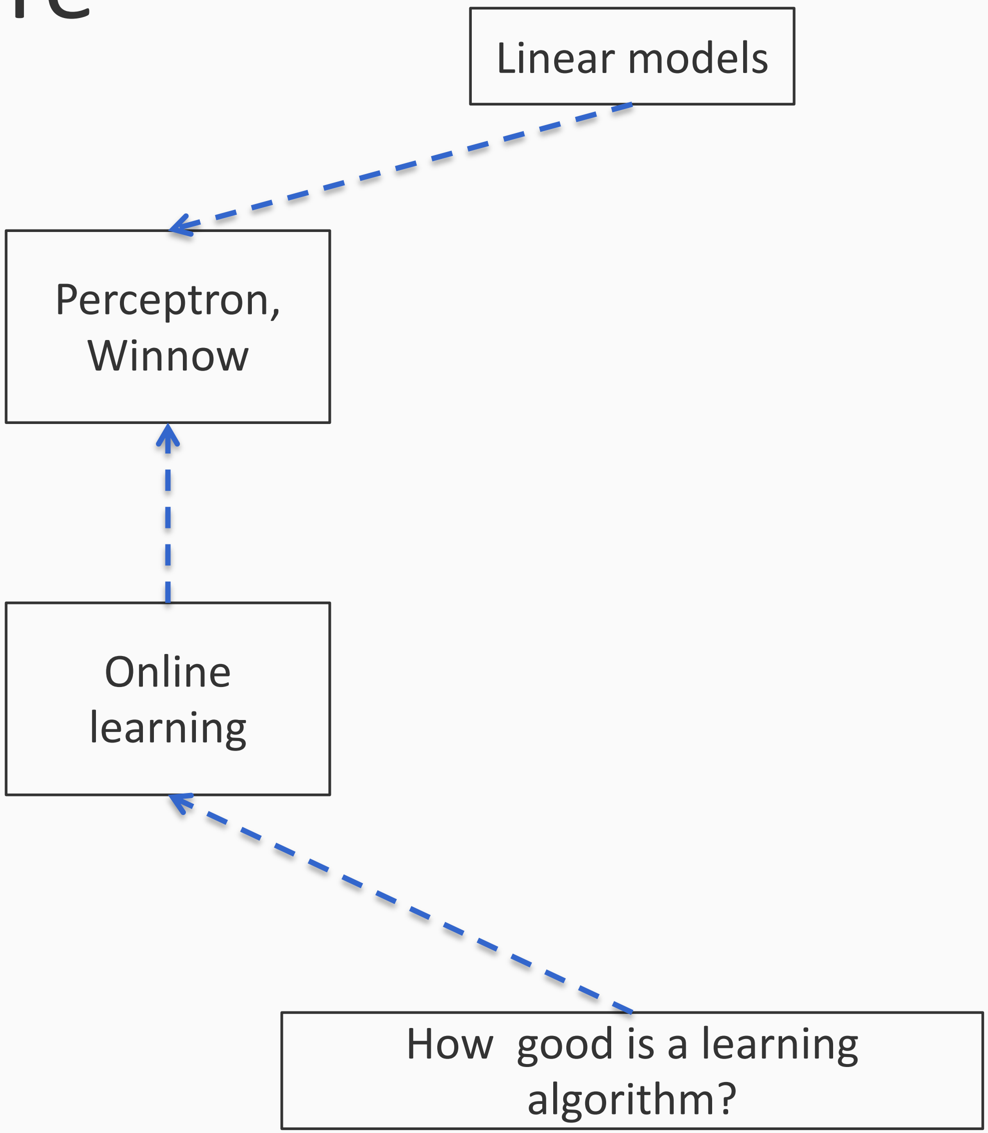
Linear models

Big picture

Linear models

How good is a learning algorithm?

Big picture



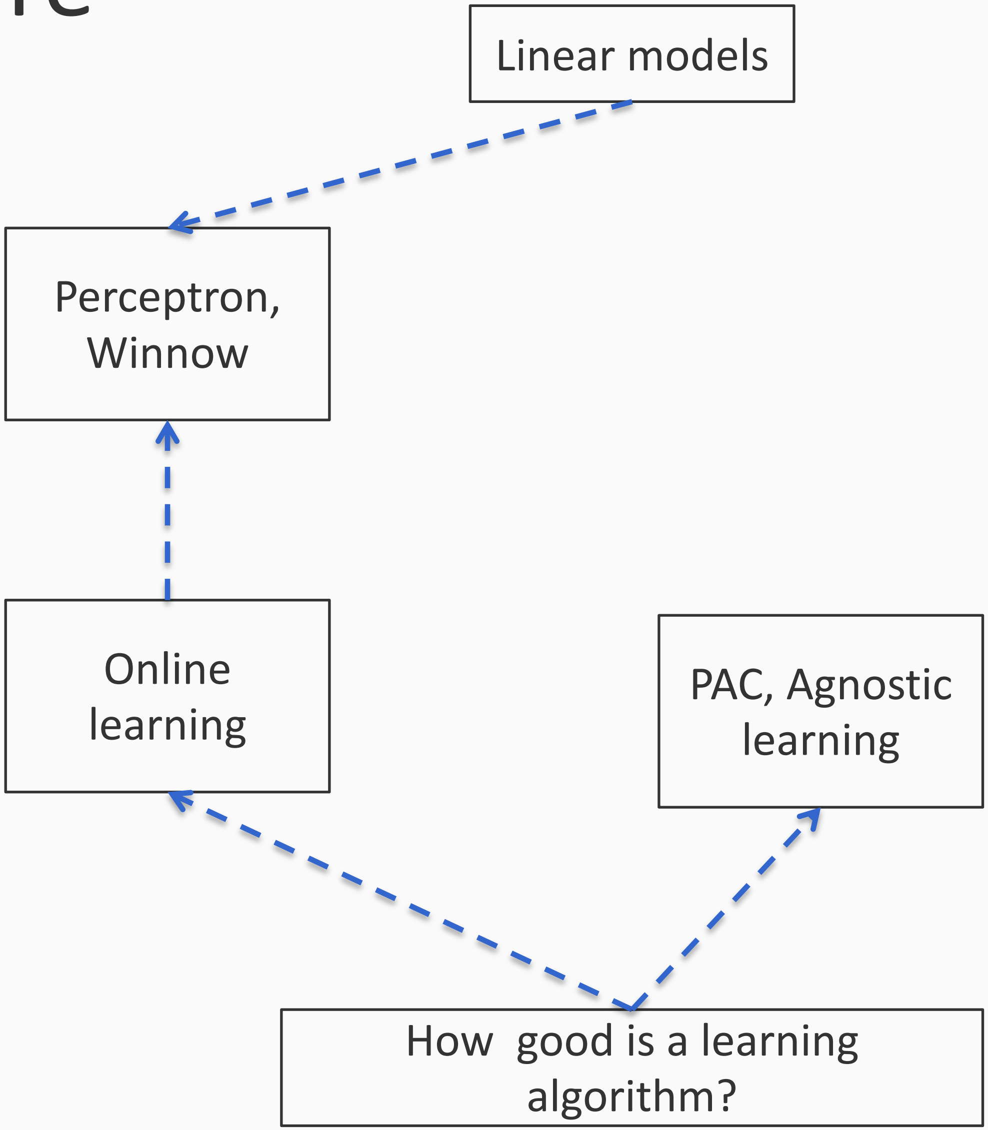
Linear models

Perceptron, Winnow

Online learning

How good is a learning algorithm?

Big picture



Linear models

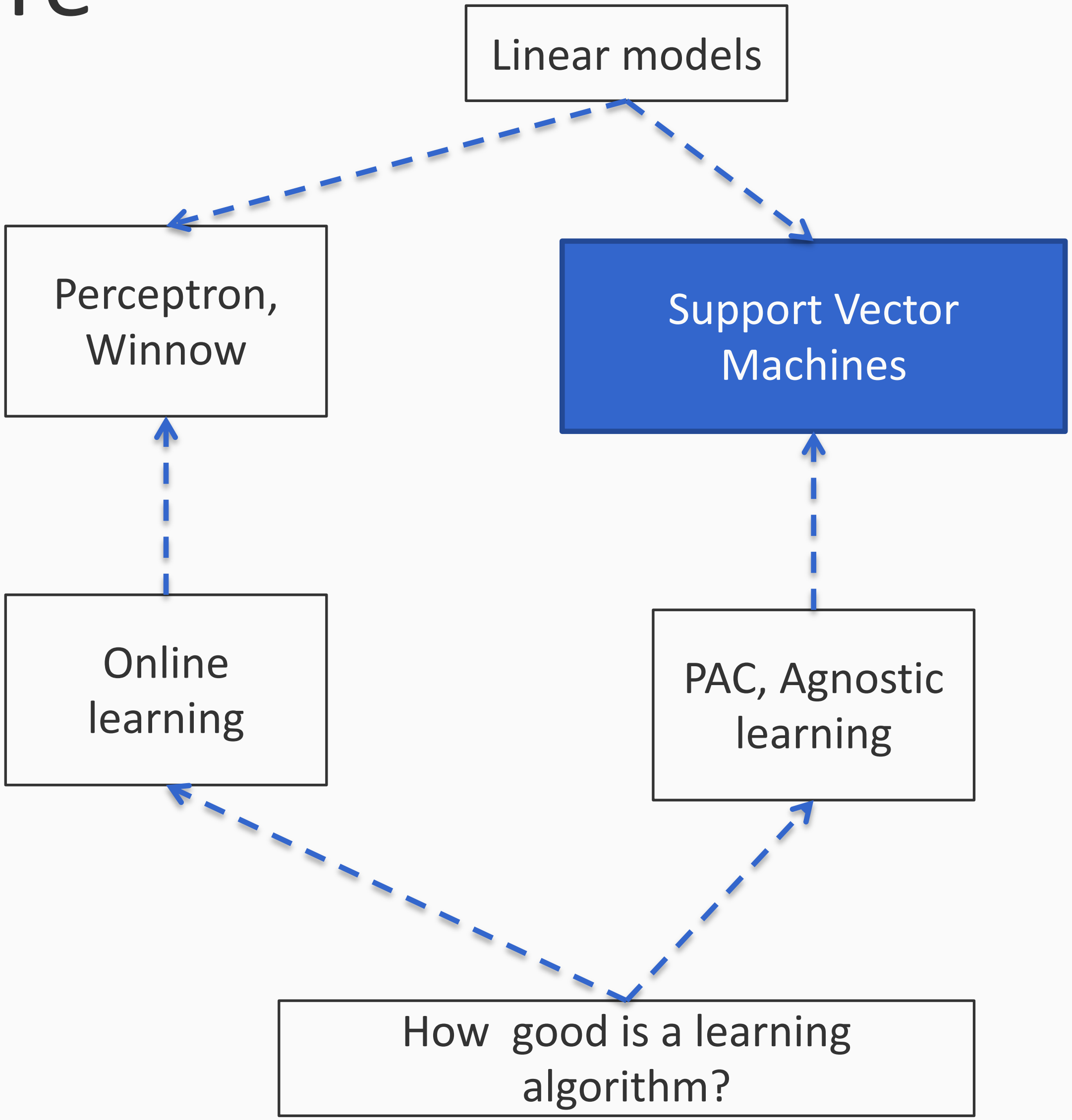
Perceptron, Winnow

Online learning

PAC, Agnostic learning

How good is a learning algorithm?

Big picture



Linear models

Perceptron, Winnow

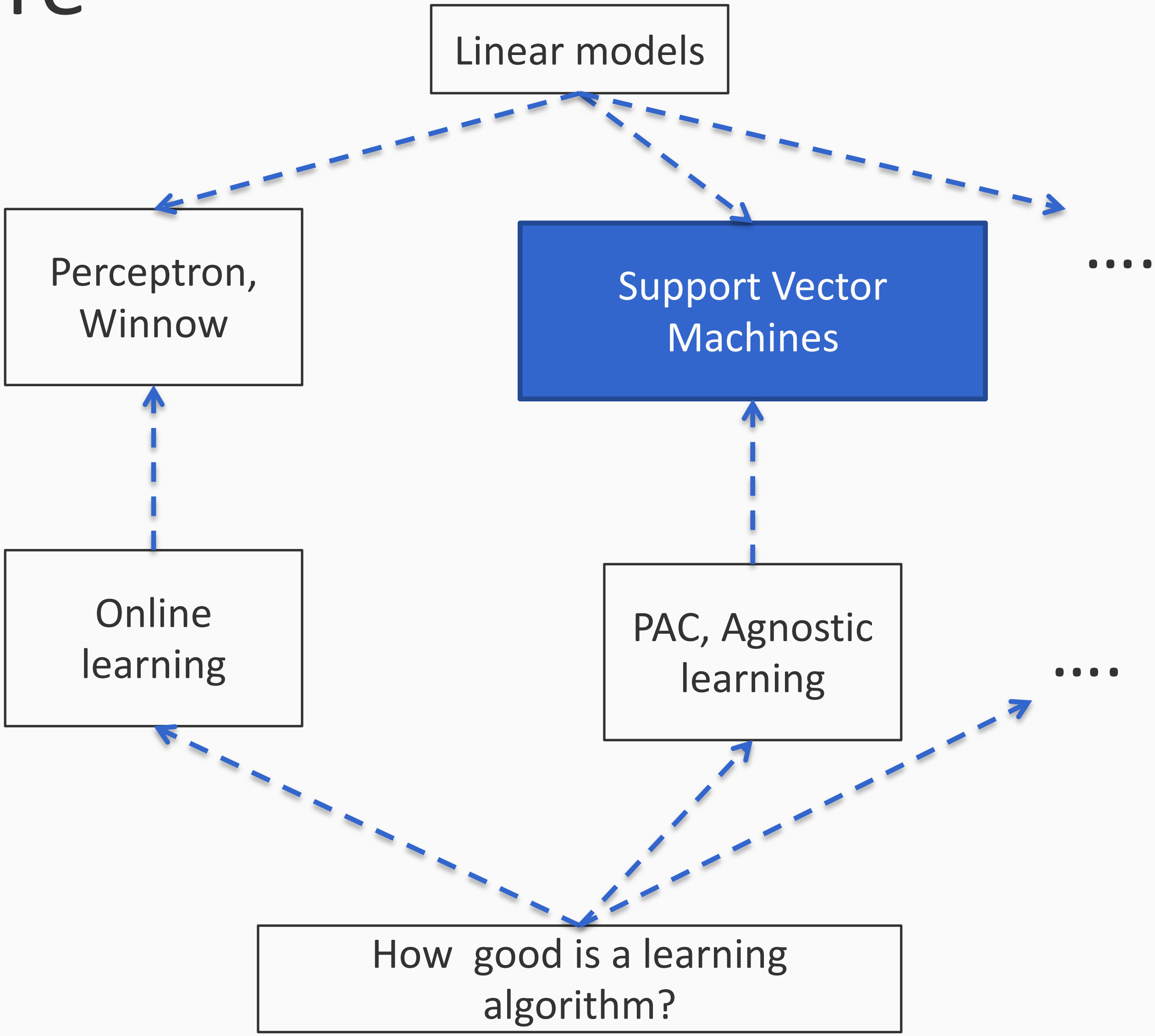
Support Vector Machines

Online learning

PAC, Agnostic learning

How good is a learning algorithm?

Big picture



Linear models

Perceptron, Winnow

Support Vector Machines

....

Online learning

PAC, Agnostic learning

....

How good is a learning algorithm?

This lecture: Support vector machines

- Training by maximizing margin
- The SVM objective
- Solving the SVM optimization problem
- Support vectors, duals and kernels

This lecture: Support vector machines

- Training by maximizing margin
- The SVM objective
- Solving the SVM optimization problem
- Support vectors, duals and kernels

VC dimensions and linear classifiers

What we know so far

1. If we have m examples, then with probability 1 - δ, the true error of a hypothesis h with training error $err_S(h)$ is bounded by

$$err_D(h) \leq err_S(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

Generalization error

Training error

A function of VC dimension.
Low VC dimension gives tighter bound

VC dimensions and linear classifiers

What we know so far

1. If we have m examples, then with probability 1 - δ, the true error of a hypothesis h with training error $err_S(h)$ is bounded by

$$\underbrace{err_D(h)}_{\text{Generalization error}} \leq \underbrace{err_S(h)}_{\text{Training error}} + \boxed{\begin{array}{c}\text{A function of VC dimension.}\\ \textit{Low VC dimension gives tighter bound}\end{array}}$$

VC dimensions and linear classifiers

What we know so far

1. If we have m examples, then with probability 1 - δ, the true error of a hypothesis h with training error $err_S(h)$ is bounded by

$$\underset{\substack{\uparrow \\ \text{Generalization error}}}{err_D(h)} \le \underset{\substack{\uparrow \\ \text{Training error}}}{err_S(h)} + \boxed{\text{A function of VC dimension.}}$$

A function of VC dimension.

Low VC dimension gives tighter bound

2. VC dimension of a linear classifier in d dimensions $= \ d \ + \ 1$

VC dimensions and linear classifiers

What we know so far

1. If we have m examples, then with probability 1 - δ, the true error of a hypothesis h with training error $err_S(h)$ is bounded by

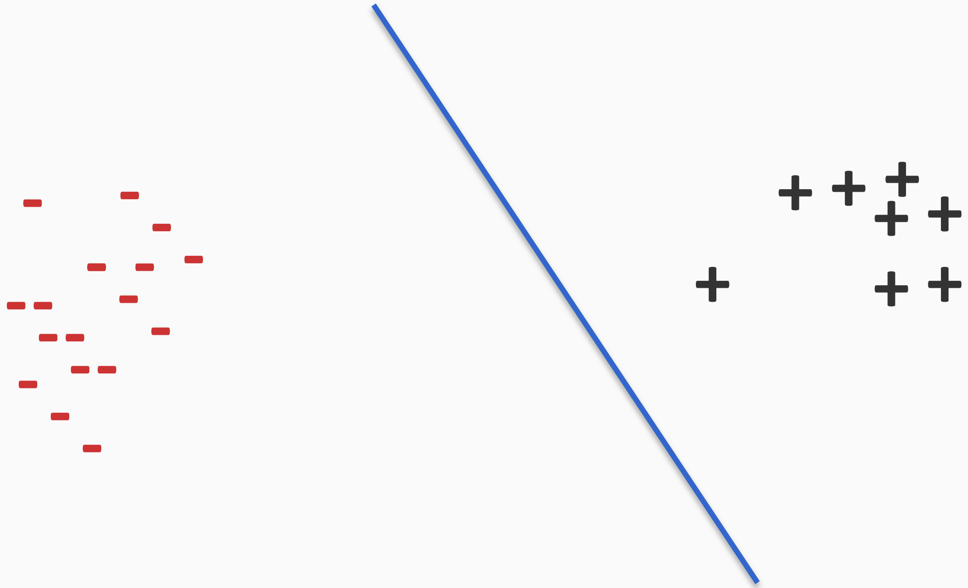
$$\underset{\substack{\uparrow \\ \text{Generalization error}}}{err_D(h)} \le \underset{\substack{\uparrow \\ \text{Training error}}}{err_S(h)} + \boxed{\begin{array}{c}\text{A function of VC dimension.} \\ \textit{Low VC dimension gives tighter bound}\end{array}}$$

2. VC dimension of a linear classifier in d dimensions $= \; d \; + \; 1$

But are all linear classifiers the same?

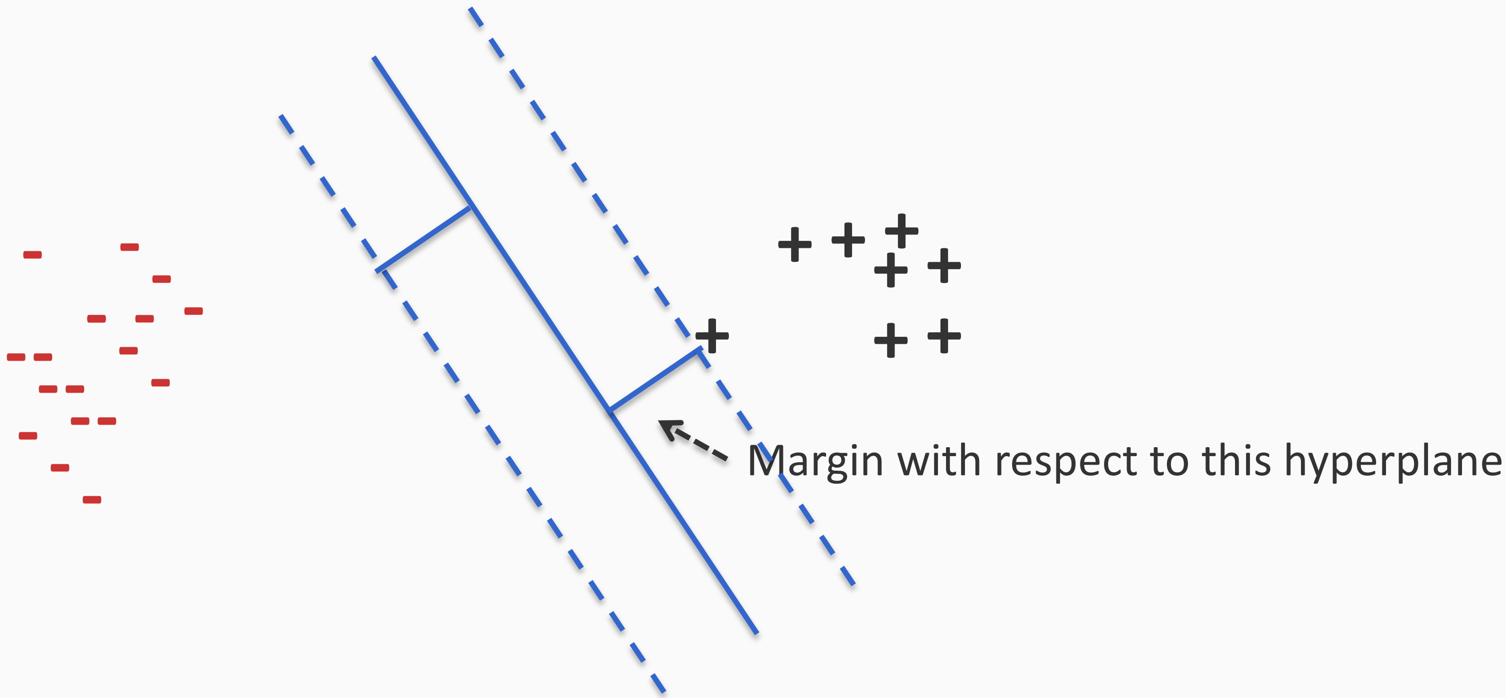
Recall: Margin

The margin of a hyperplane for a dataset is the distance between the hyperplane and the data point nearest to it.

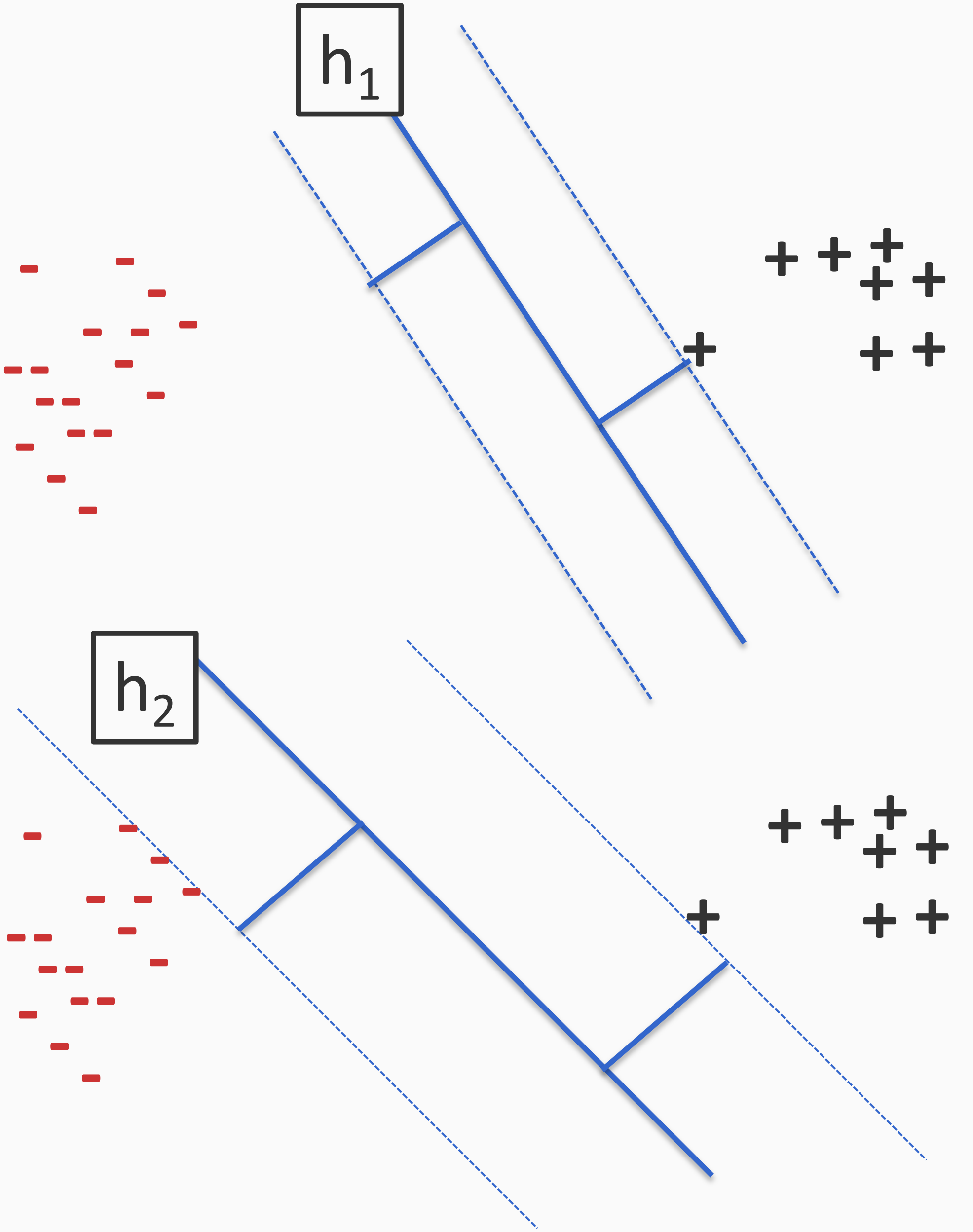


Recall: Margin

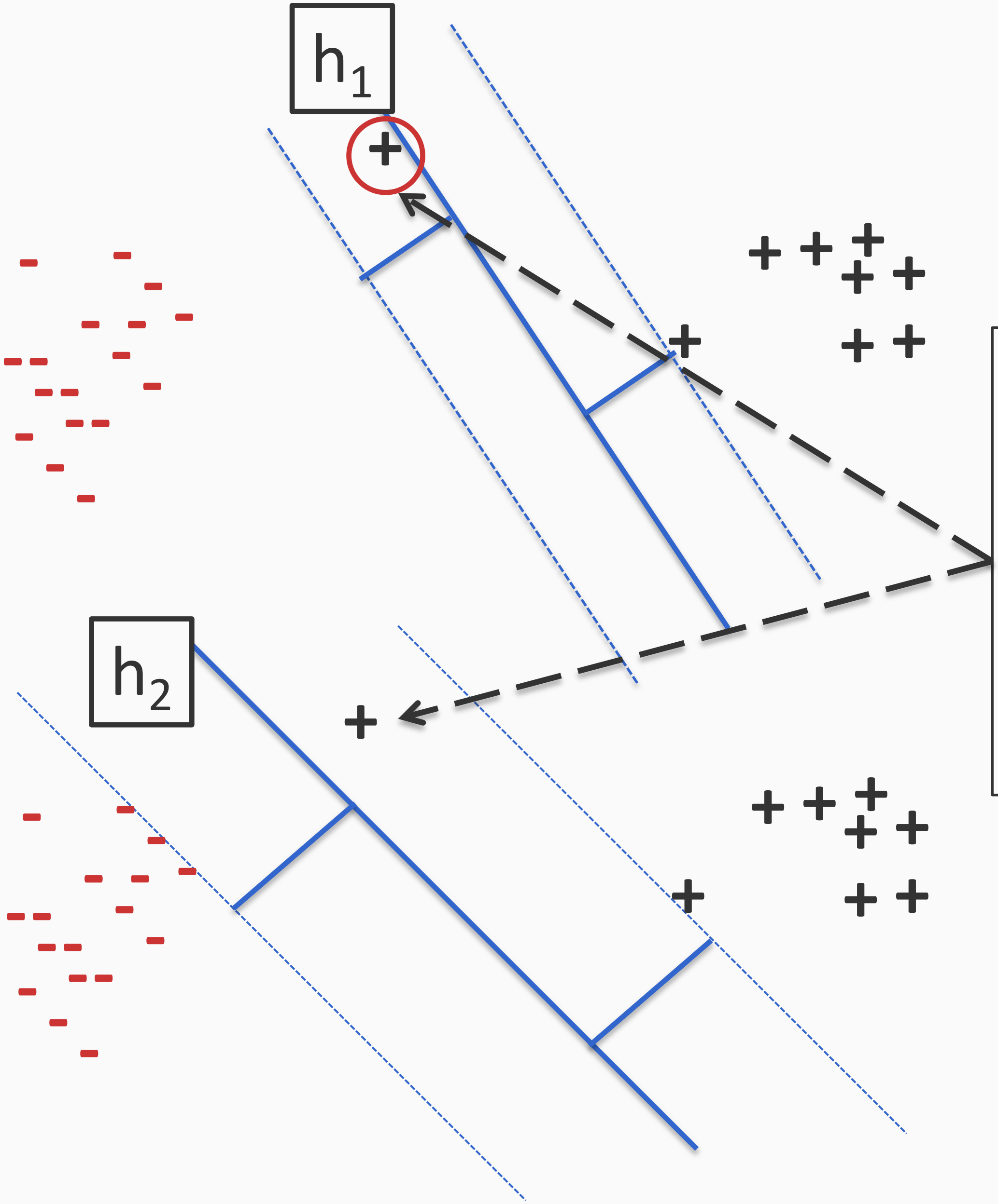
The margin of a hyperplane for a dataset is the distance between the hyperplane and the data point nearest to it.



Which line is a better choice? Why?



Which line is a better choice? Why?



A new example, not from the training set might be misclassified if the margin is smaller

Data dependent VC dimension

Intuitively, larger margins are better

Consider linear separators with margins γ_1 and γ_2

- H_1 = linear separators that have a margin γ_1
- H_2 = linear separators that have a margin γ_2
- And $\gamma_1 > \gamma_2$

Claim: The entire set of functions H_1 is “better”

Data dependent VC dimension

Theorem (Vapnik):

- Let H be the set of linear classifiers that separate the training set by a margin at least γ
- Then

$$VC(H) \leq \min\left(\frac{R^2}{\gamma^2}, d\right) + 1$$

- R is the radius of the smallest sphere containing the data

Data dependent VC dimension

Theorem (Vapnik):

- Let H be the set of linear classifiers that separate the training set by a margin at least γ
- Then

$$VC(H) \leq \min\left(\frac{R^2}{\gamma^2}, d\right) + 1$$

- R is the radius of the smallest sphere containing the data

Larger margin ⇒ Lower VC dimension

Lower VC dimension ⇒ Better generalization bound

Learning strategy

Find the linear separator that maximizes the margin

This lecture: Support vector machines

- Training by maximizing margin
- The SVM objective
- Solving the SVM optimization problem
- Support vectors, duals and kernels

Support Vector Machines

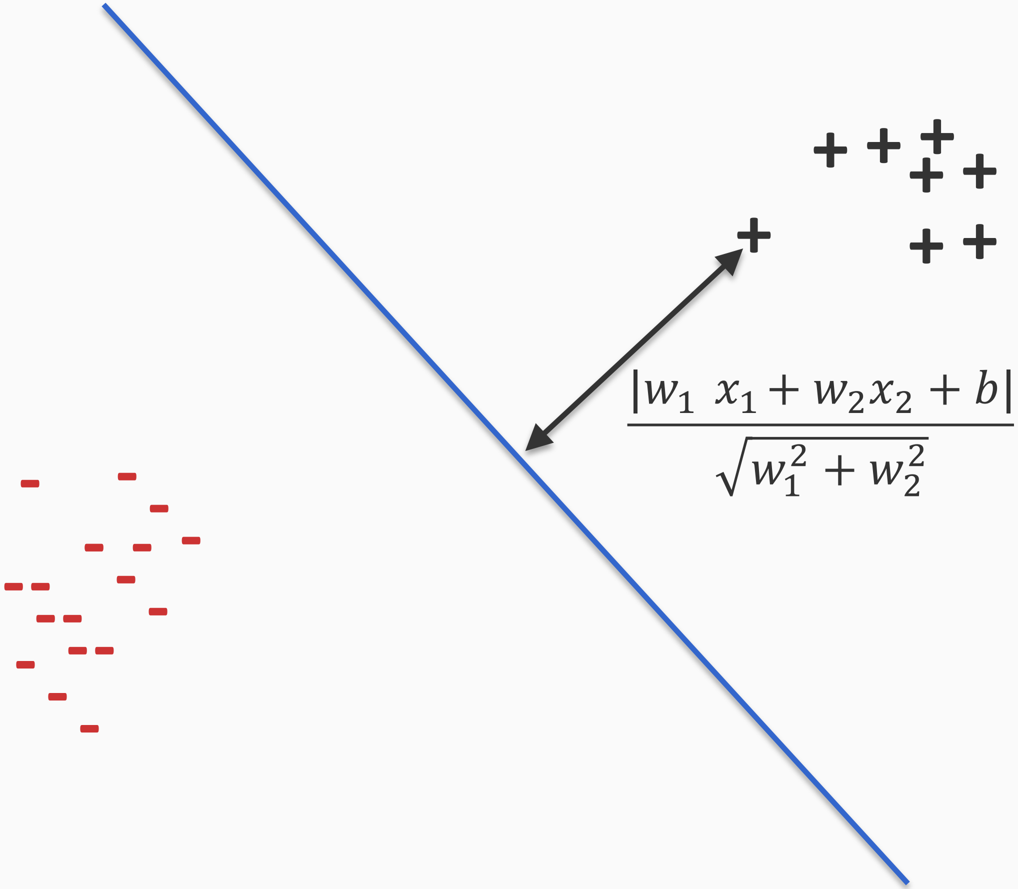
So far

- Lower VC dimension → Better generalization
- Vapnik: For linear separators, the VC dimension depends inversely on the margin
  - That is, larger margin → better generalization

Recall: The geometry of a linear classifier

Prediction = $\text{sgn}(b + w_1 x_1 + w_2 x_2)$

$b + w_1 x_1 + w_2 x_2 = 0$

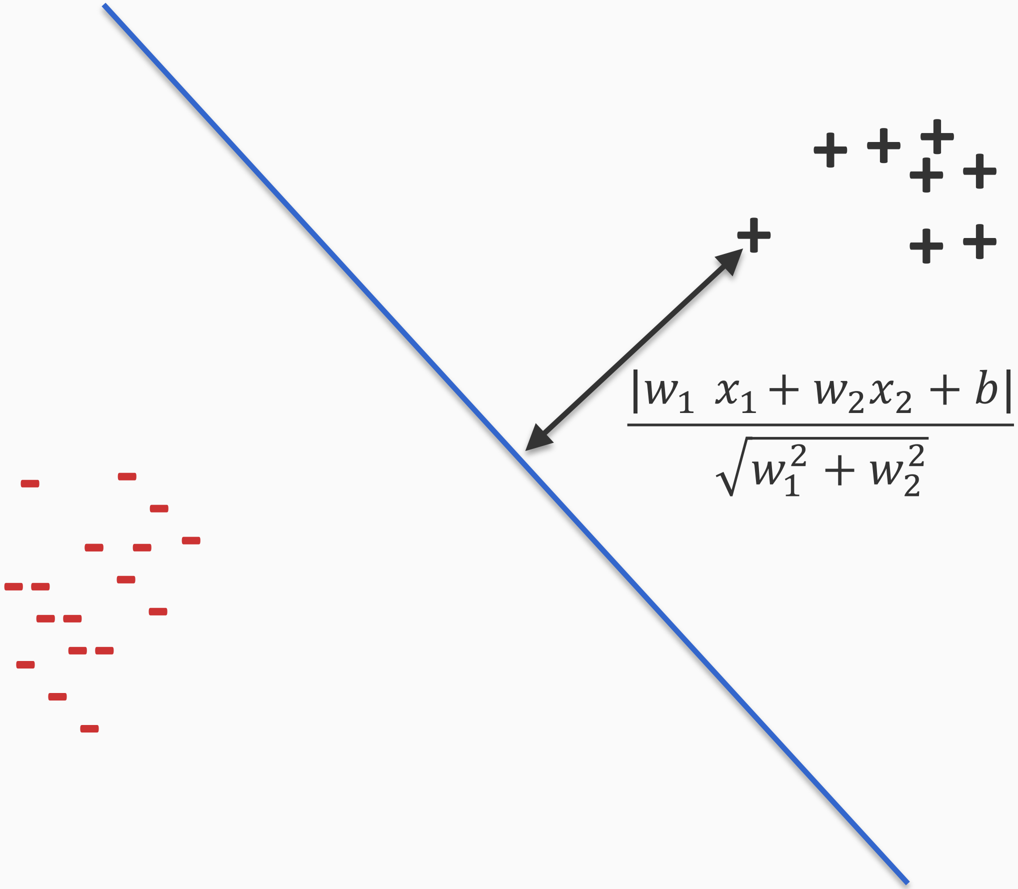


$$\frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}} = \frac{y(w_1 x_1 + w_2 x_2 + b)}{||\mathbf{w}||}$$

Recall: The geometry of a linear classifier

Prediction = $\mathrm{sgn}(b\ + w_1\, x_1\ +\ w_2 x_2)$

$b\ + w_1\, x_1\ +\ w_2 x_2 = 0$

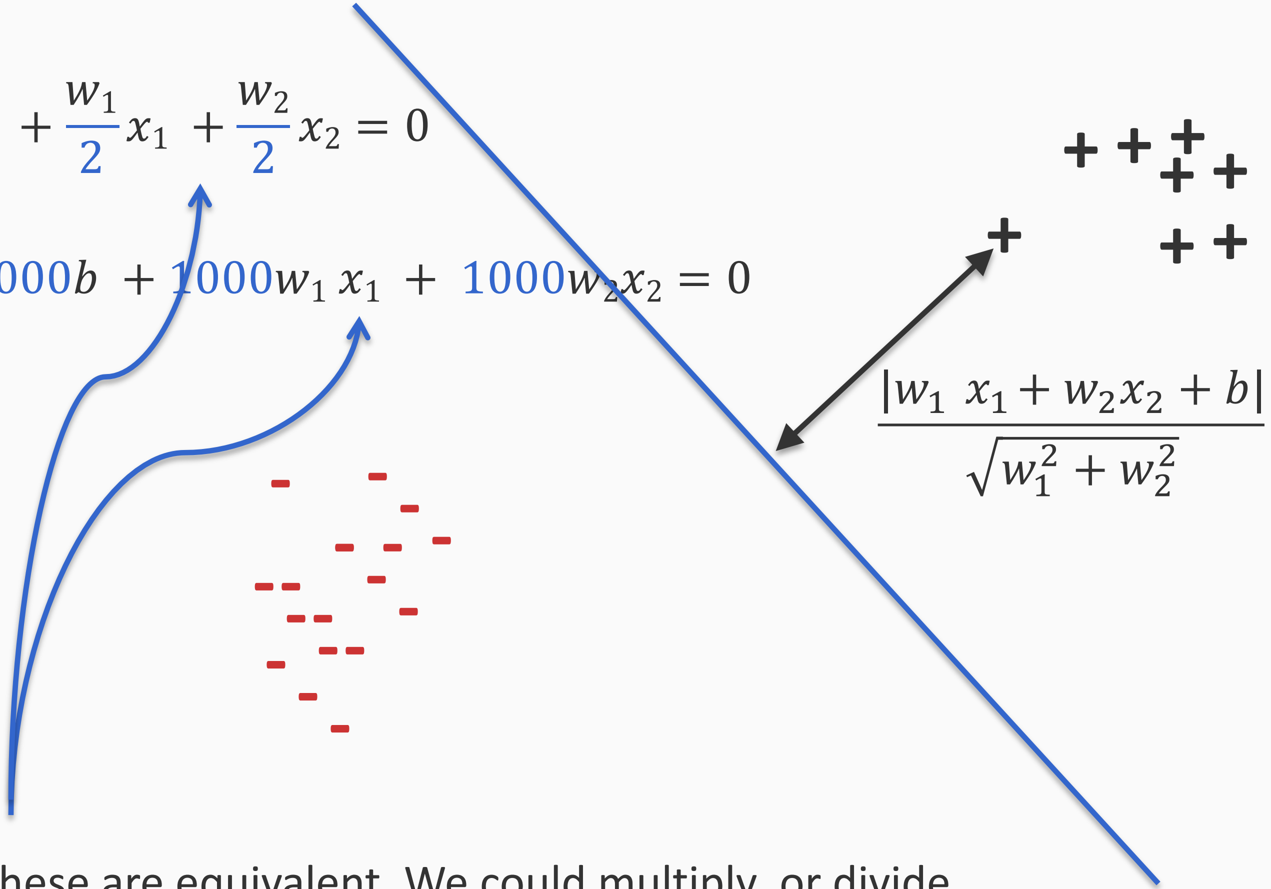


We only care about the sign, not the magnitude

$$\frac{|w_1\ x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}} = \frac{y(w_1 x_1 + w_2 x_2 + b)}{||\mathbf{w}||}$$

Recall: The geometry of a linear classifier

Prediction = $\mathrm{sgn}(b + w_1 x_1 + w_2 x_2)$



$$b + w_1 x_1 + w_2 x_2 = 0$$

$$\frac{b}{2} + \frac{w_1}{2} x_1 + \frac{w_2}{2} x_2 = 0$$

$$1000b + 1000 w_1 x_1 + 1000 w_2 x_2 = 0$$

We only care about the sign, not the magnitude

$$\frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}} = \frac{y(w_1 x_1 + w_2 x_2 + b)}{||\mathbf{w}||}$$

These are equivalent. We could multiply or divide the coefficients by any positive number and the sign of the prediction will not change

Maximizing margin

- Margin of a hyperplane = distance of the closest point from the hyperplane

$$\gamma_{\mathbf{w},b} = \min_i \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{||\mathbf{w}||}$$

Maximizing margin

- Margin of a hyperplane = distance of the closest point from the hyperplane

$$\gamma_{\mathbf{w},b} = \min_i \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{||\mathbf{w}||}$$

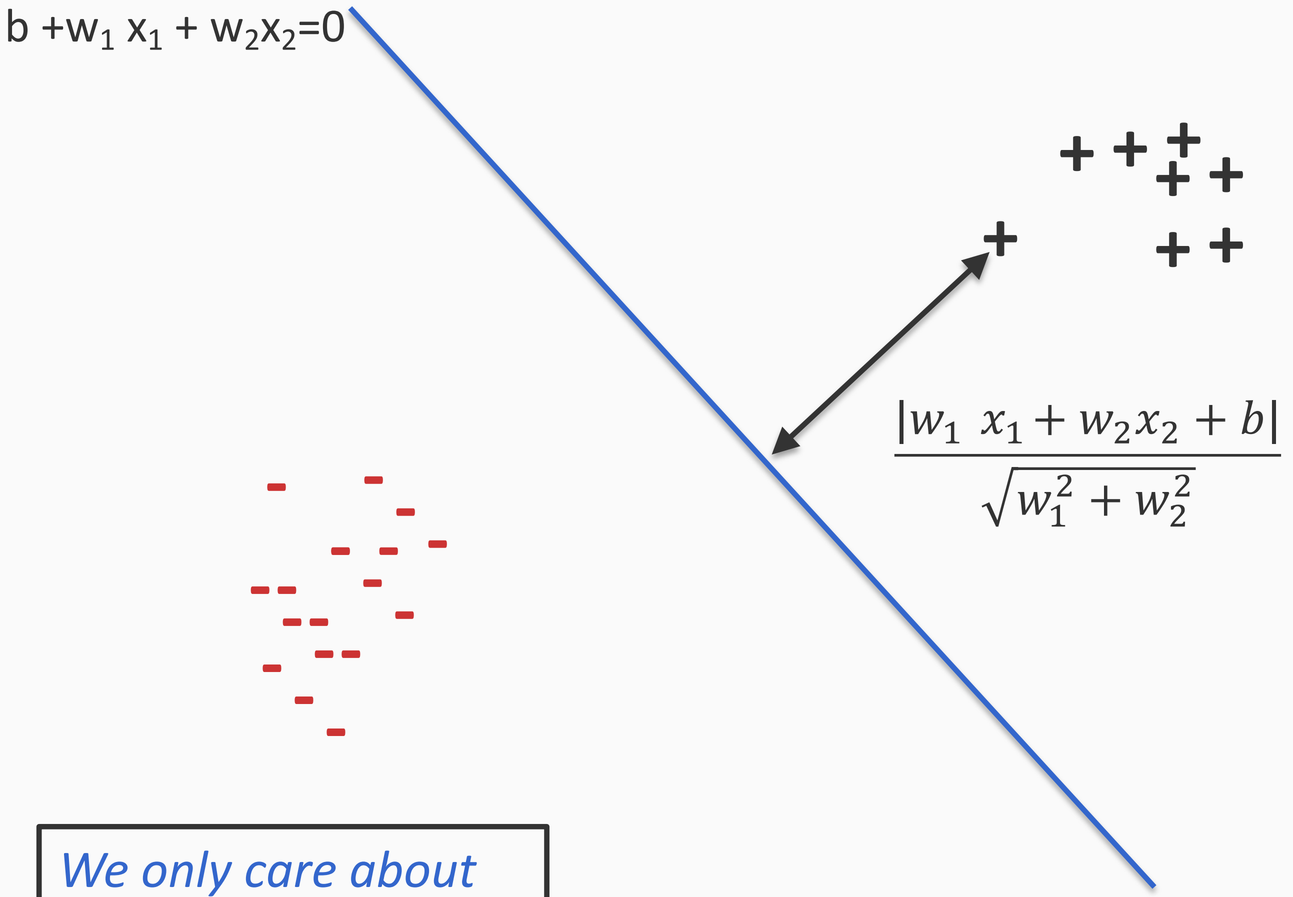
- We want to maximize this margin: $\max_{\mathbf{w},b} \gamma_{\mathbf{w},b}$

Sometimes this is called the *geometric margin*

The numerator alone is called the *functional margin*

Recall: The geometry of a linear classifier

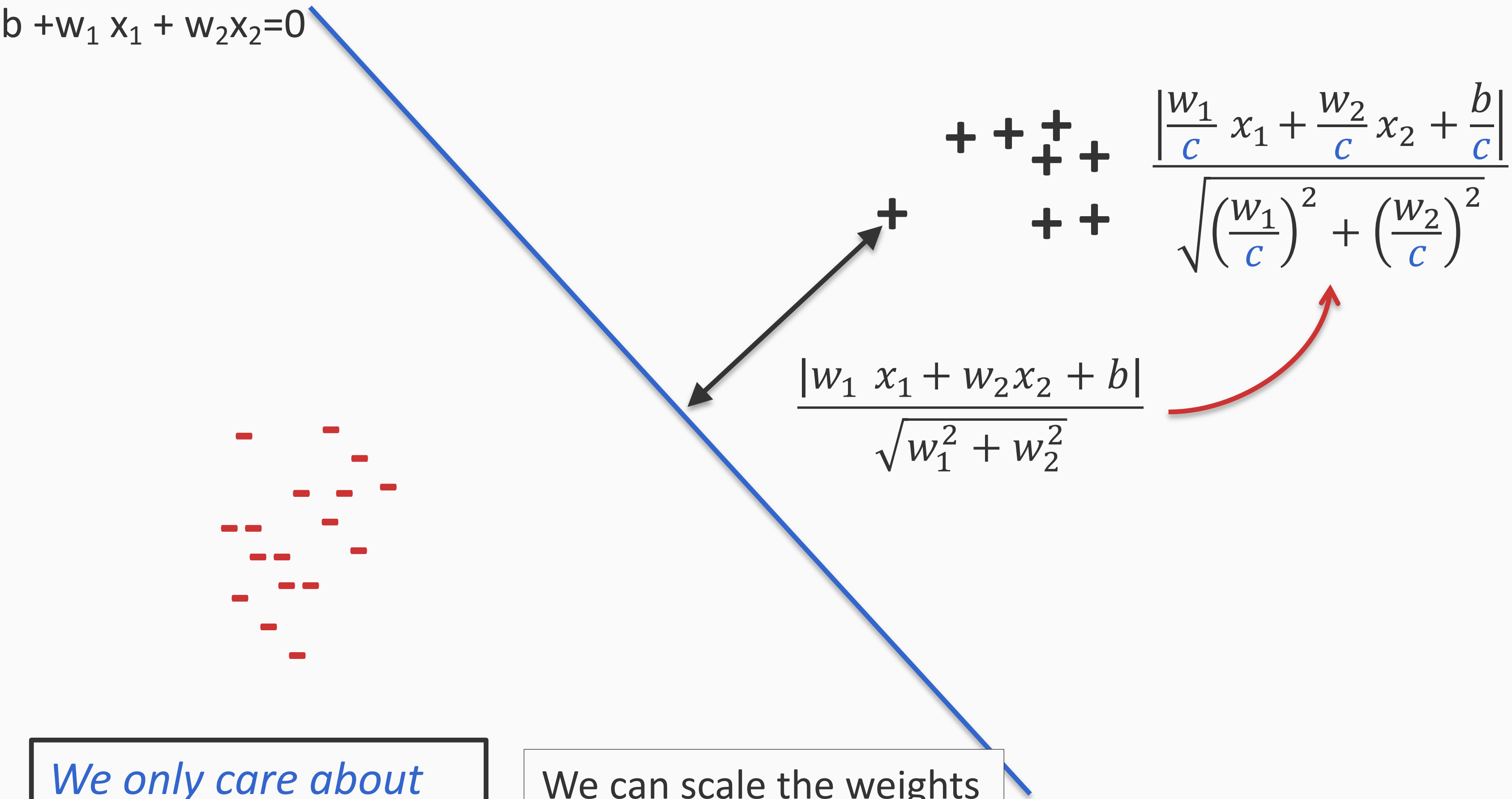
Prediction = $\text{sgn}(b + w_1 x_1 + w_2 x_2)$



$$\frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}$$

We only care about the sign, not the magnitude

Towards maximizing the margin



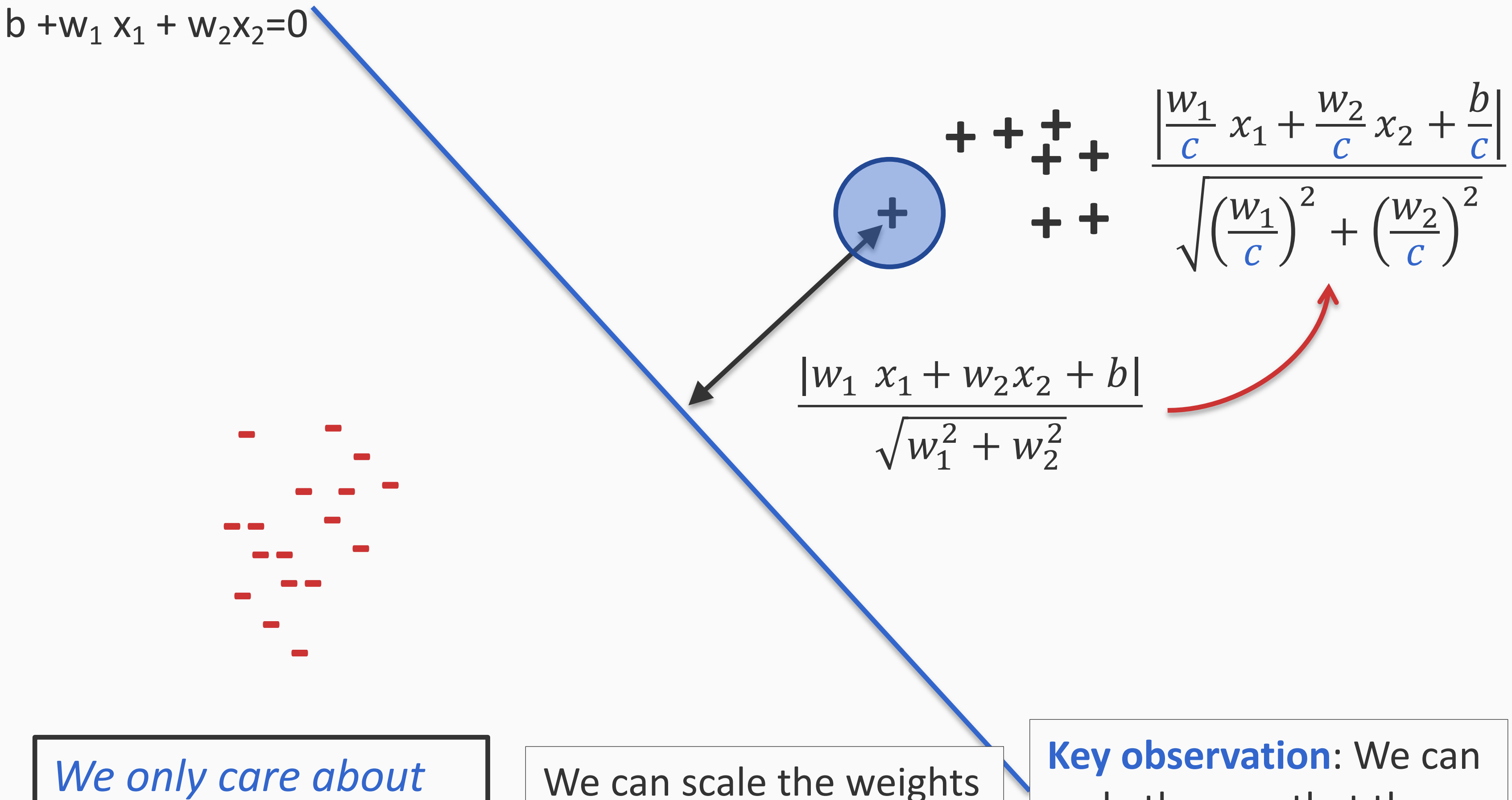
$$\frac{|w_1\ x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}$$

$$\frac{\left|\frac{w_1}{c}\ x_1 + \frac{w_2}{c}\ x_2 + \frac{b}{c}\right|}{\sqrt{\left(\frac{w_1}{c}\right)^2 + \left(\frac{w_2}{c}\right)^2}}$$

We only care about the sign, not the magnitude

We can scale the weights to make the optimization easier

Towards maximizing the margin



$$\frac{\left|\frac{w_1}{c}\,x_1+\frac{w_2}{c}\,x_2+\frac{b}{c}\right|}{\sqrt{\left(\frac{w_1}{c}\right)^2+\left(\frac{w_2}{c}\right)^2}}$$

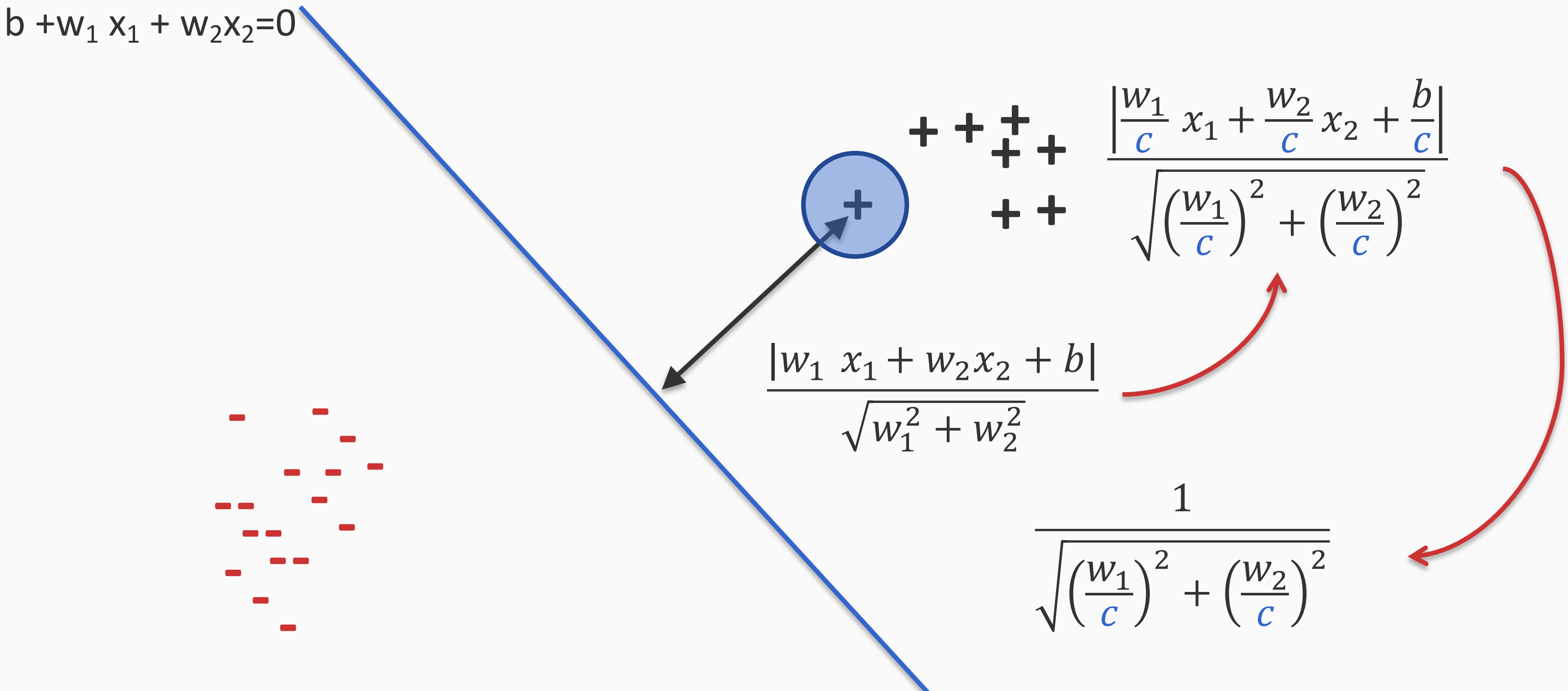
$$\frac{|w_1\ x_1+w_2x_2+b|}{\sqrt{w_1^2+w_2^2}}$$

We only care about the sign, not the magnitude

We can scale the weights to make the optimization easier

Key observation: We can scale the c so that the numerator is 1 for points that define the margin.

Towards maximizing the margin



$b + w_1 x_1 + w_2 x_2 = 0$

$$\frac{\left|\frac{w_1}{c} x_1 + \frac{w_2}{c} x_2 + \frac{b}{c}\right|}{\sqrt{\left(\frac{w_1}{c}\right)^2 + \left(\frac{w_2}{c}\right)^2}}$$

$$\frac{|w_1 \ x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}$$

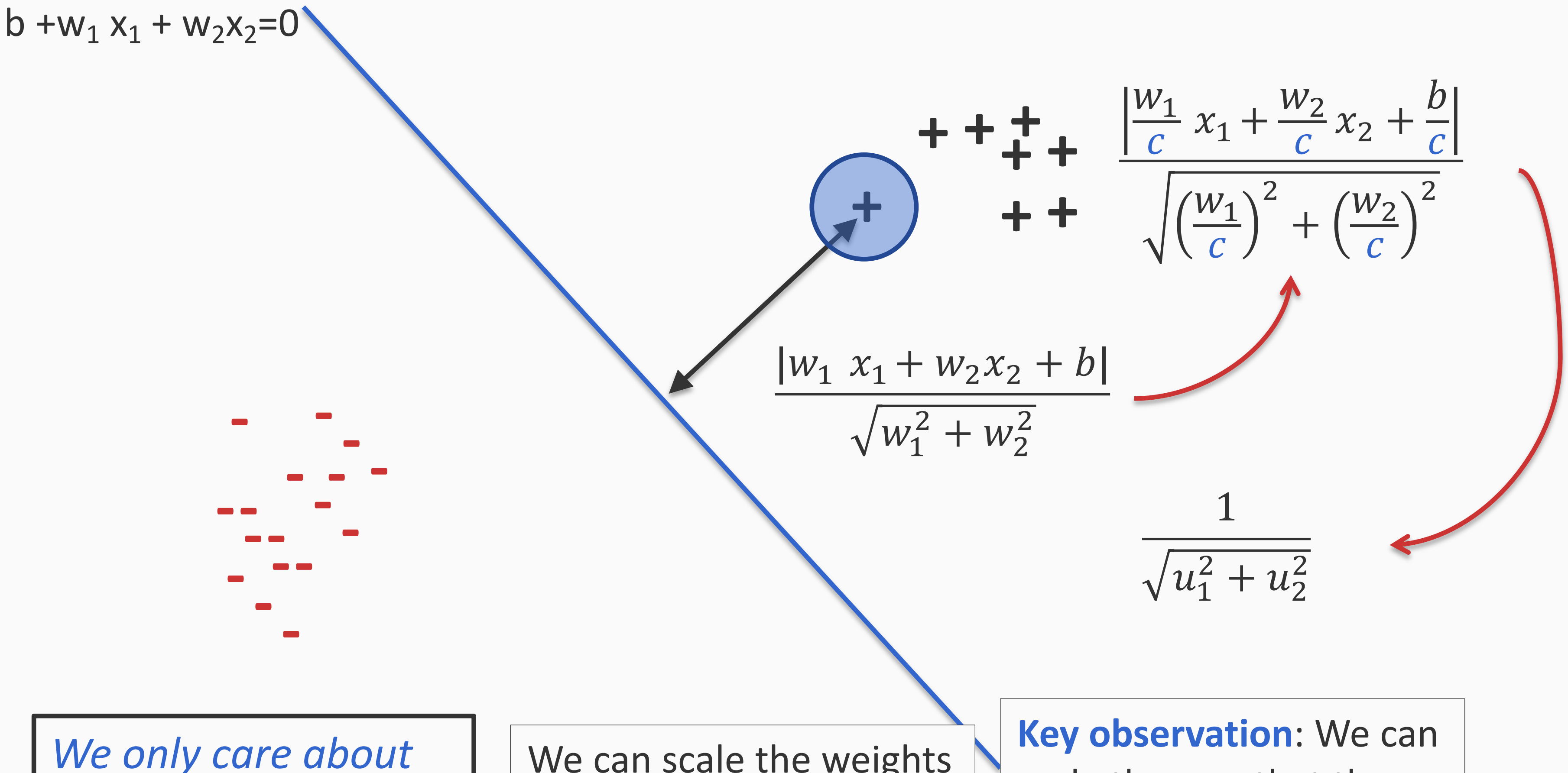
$$\frac{1}{\sqrt{\left(\frac{w_1}{c}\right)^2 + \left(\frac{w_2}{c}\right)^2}}$$

We only care about the sign, not the magnitude

We can scale the weights to make the optimization easier

Key observation: We can scale the c so that the numerator is 1 for points that define the margin.

Towards maximizing the margin



$b + w_1 x_1 + w_2 x_2 = 0$

$$\frac{\left|\frac{w_1}{c} x_1 + \frac{w_2}{c} x_2 + \frac{b}{c}\right|}{\sqrt{\left(\frac{w_1}{c}\right)^2 + \left(\frac{w_2}{c}\right)^2}}$$

$$\frac{|w_1 \ x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}$$

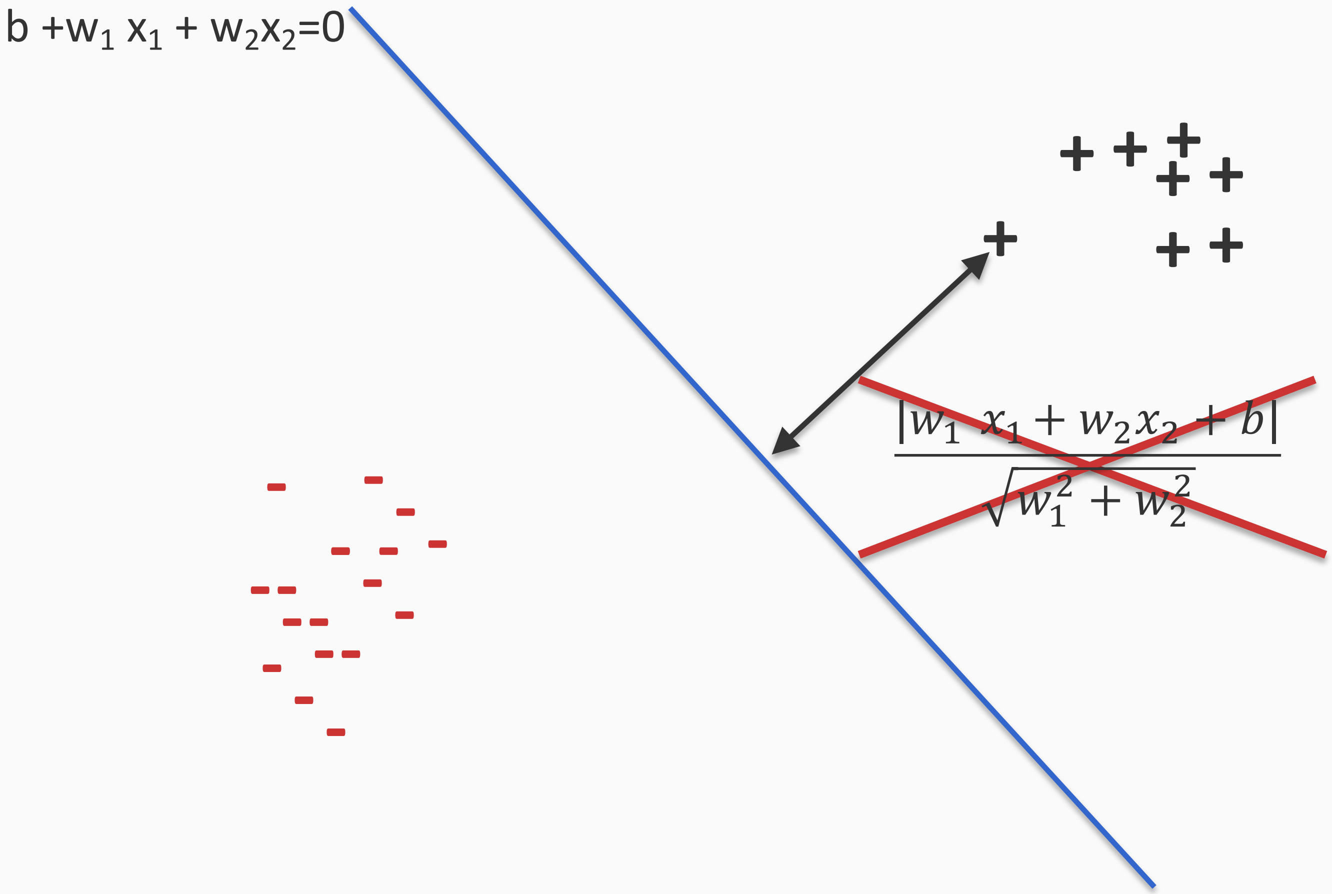
$$\frac{1}{\sqrt{u_1^2 + u_2^2}}$$

We only care about the sign, not the magnitude

We can scale the weights to make the optimization easier

Key observation: We can scale the c so that the numerator is 1 for points that define the margin.

Towards maximizing the margin



$$\xcancel{\frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}}$$

$$\frac{1}{\sqrt{w_1^2 + w_2^2}}$$

Maximizing margin

- Margin of a hyperplane = distance of the closest point from the hyperplane

$$\gamma_{\mathbf{w},b} = \min_i \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{||\mathbf{w}||}$$

- We want to maximize this margin: $\max_{\mathbf{w},b} \gamma_{\mathbf{w},b}$
- We only care about the sign of $\mathbf{w}$ and b in the end and not the magnitude
  - Set the absolute score (functional margin) of the closest point to be 1 and allow $\mathbf{w}$ to adjust itself

$\max_{\mathbf{w}} \gamma$ is equivalent to $\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}$ in this setting

$$\gamma = \min_{\mathbf{x}_i, y_i} \frac{y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)}{||\mathbf{w}||}$$

Max-margin classifiers

Learning problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{s.t.} \forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 \end{aligned}$$

$$\gamma = \min_{\mathbf{x}_i, y_i} \frac{y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)}{||\mathbf{w}||}$$

Max-margin classifiers

Learning problem:

Mimimizing gives us $\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}$

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{s.t.} \forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 \end{aligned}$$

$$\gamma = \min_{\mathbf{x}_i, y_i} \frac{y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)}{||\mathbf{w}||}$$

Max-margin classifiers

Learning problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{s.t.}\forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 \end{aligned}$$

Mimimizing gives us $\max_{\mathbf{w}} \frac{1}{||\mathbf{w}||}$

This condition is true for every example, specifically, for the example closest to the separator

Max-margin classifiers

$$\gamma = \min_{\mathbf{x}_i, y_i} \frac{y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)}{||\mathbf{w}||}$$

Learning problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{s.t.} \forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 \end{aligned}$$

Mimimizing gives us $\max_{\mathbf{w}} \frac{1}{\|\mathbf{w}\|}$

This condition is true for every example, specifically, for the example closest to the separator

- This is called the "hard" Support Vector Machine

We will look at how to solve this optimization problem later

Support Vector Machines

So far

- Lower VC dimension → Better generalization
- Vapnik: For linear separators, the VC dimension depends inversely on the margin
  - That is, larger margin → better generalization
- For the separable case:
  - Among all linear classifiers that separate the data, find the one that maximizes the margin
  - Maximizing the margin by minimizing $\boldsymbol{w}^T\boldsymbol{w}$ if for all examples $y\boldsymbol{w}^T\boldsymbol{x} \geq 1$

What if the data is not separable?

Hard SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin

$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T\mathbf{x}_i \geq 1$$

Every example has an functional margin of at least 1

What if the data is not separable?

Hard SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Maximize margin

Every example has an functional margin of at least 1

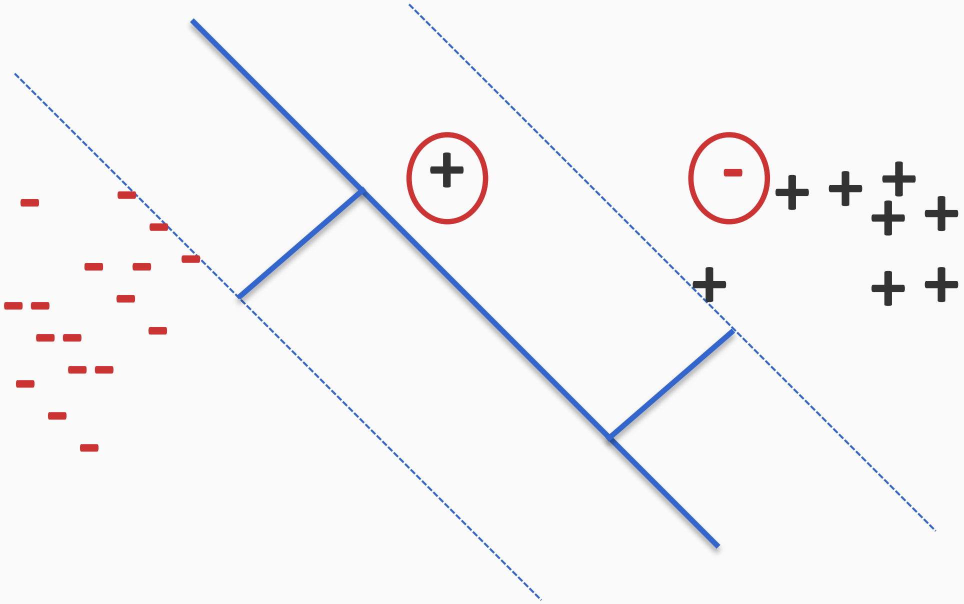
- This is a constrained optimization problem
- If the data is not linearly separable, there is no **w** that will classify the data
- Infeasible problem, no solution!

Dealing with non-separable data

Key idea: Allow some examples to “break into the margin”

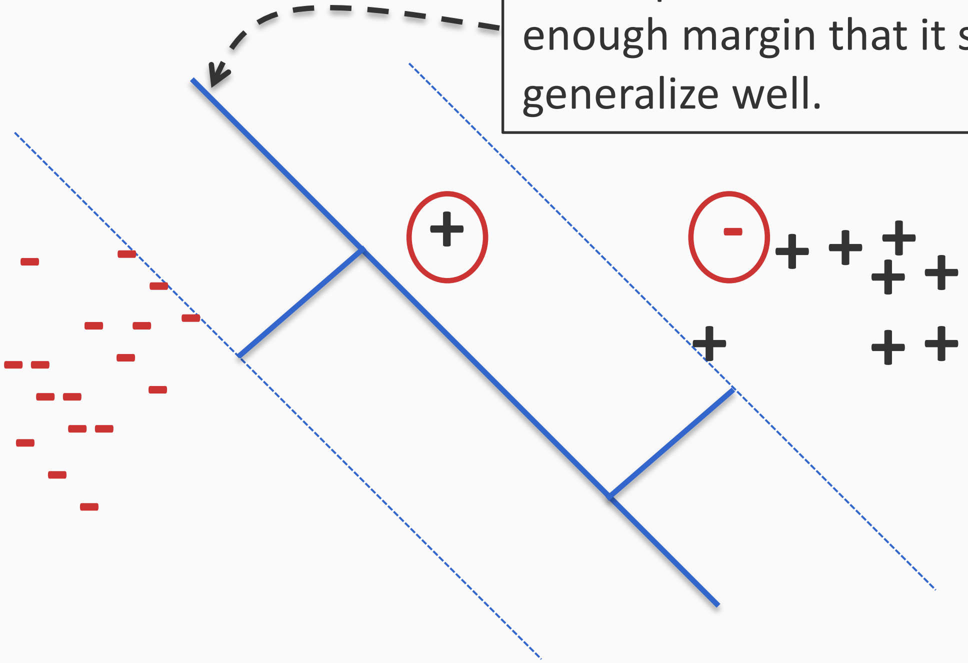
Dealing with non-separable data

Key idea: Allow some examples to "break into the margin"



Dealing with non-separable data

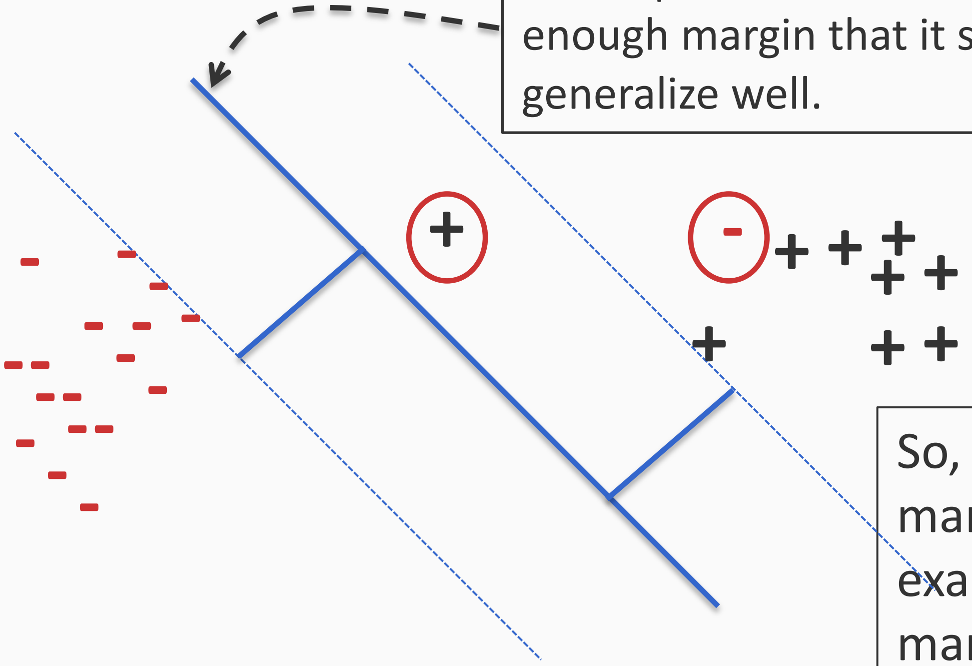
Key idea: Allow some examples to “break into the margin”



This separator has a large enough margin that it should generalize well.

Dealing with non-separable data

Key idea: Allow some examples to "break into the margin"



This separator has a large enough margin that it should generalize well.

So, when computing margin, ignore the examples that make the margin smaller or the data inseparable.

Soft SVM

- Hard SVM:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin

$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Every example has an functional margin of at least 1

Soft SVM

- Hard SVM:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin

$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Every example has an functional margin of at least 1

- Introduce one *slack variable* ξ_i per example
  - And require $y_i w^T x_i \geq 1 - \xi_i$ and $\xi_i \geq 0$

Soft SVM

- Hard SVM:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin

$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Every example has an functional margin of at least 1

- Introduce one *slack variable* ξ_i per example
  - And require $y_i w^T x_i \geq 1 - \xi_i$ and $\xi_i \geq 0$

Intuition: The slack variable allows examples to “break” into the margin

If the slack value is zero, then the example is either on or outside the margin

Soft SVM

- Hard SVM:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin

$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Every example has an functional margin of at least 1

- Introduce one *slack variable* ξ_i per example
  - And require $y_i w^T x_i \geq 1 - \xi_i$ and $\xi_i \geq 0$
- New optimization problem for learning

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i \\ \text{s.t.} \quad \forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 - \xi_i \\ \forall i, \quad & \xi_i \geq 0. \end{aligned}$$

Soft SVM

- Hard SVM:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Maximize margin

$$\text{s.t.} \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1$$

Every example has an functional margin of at least 1

- Introduce one *slack variable* ξ_i per example
  - And require $y_i w^T x_i \geq 1 - \xi_i$ and $\xi_i \geq 0$
- New optimization problem for learning

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i \\ \text{s.t.} \quad \forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 - \xi_i \\ \forall i, \quad & \xi_i \geq 0. \end{aligned}$$

Soft SVM

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i \\ \text{s.t.} \quad & \forall i, & y_i\mathbf{w}^T\mathbf{x}_i \geq 1 - \xi_i \\ & \forall i, & \xi_i \geq 0. \end{aligned}$$

Soft SVM

Maximize margin

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \boxed{\frac{1}{2}\mathbf{w}^T\mathbf{w}} + C\sum_i \xi_i \\ \text{s.t.} \quad \forall i, \quad & y_i\mathbf{w}^T\mathbf{x}_i \geq 1 - \xi_i \\ \forall i, \quad & \xi_i \geq 0. \end{aligned}$$

Soft SVM

Maximize margin

$$\min_{\mathbf{w},\xi} \quad \boxed{\frac{1}{2}\mathbf{w}^T\mathbf{w}} + C\boxed{\sum_i \xi_i}$$

Minimize total slack (i.e allow as few examples as possible to violate the margin)

$$\begin{aligned} \text{s.t.} \quad & \forall i, \quad y_i \mathbf{w}^T \mathbf{x}_i \geq 1 - \xi_i \\ & \forall i, \quad \xi_i \geq 0. \end{aligned}$$

Soft SVM

Maximize margin

Tradeoff between
the two terms

$$
\begin{aligned}
\min_{\mathbf{w},\xi} \quad & \boxed{\frac{1}{2}\mathbf{w}^T\mathbf{w}} + \boxed{C}\boxed{\sum_i \xi_i} \\
\text{s.t.} \quad & \forall i, \quad y_i\mathbf{w}^T\mathbf{x}_i \geq 1 - \xi_i \\
& \forall i, \quad \xi_i \geq 0.
\end{aligned}
$$

Minimize total slack (i.e allow
as few examples as possible
to violate the margin)

Support Vector Machines

So far

- Lower VC dimension → Better generalization

- Vapnik: For linear separators, the VC dimension depends inversely on the margin
  - That is, larger margin → better generalization

- For the separable case:
  - Among all linear classifiers that separate the data, find the one that maximizes the margin
  - Maximizing the margin by minimizing $\mathbf{w}^T\mathbf{w}$ if for all examples $y\mathbf{w}^T\mathbf{x} \geq 1$

- General case:
  - Introduce slack variables – one ξ_i for each example
  - Slack variables allow the margin constraint above to be violated

Support Vector Machines

Maximize margin

Tradeoff between
the two terms

$$\min_{\mathbf{w},\xi} \quad \boxed{\frac{1}{2}\mathbf{w}^T\mathbf{w}} + \boxed{C}\boxed{\sum_i \xi_i}$$

Minimize total slack (i.e allow
as few examples as possible
to violate the margin)

$$\begin{aligned} \text{s.t.} \quad & \forall i, \quad y_i\mathbf{w}^T\mathbf{x}_i \geq 1 - \xi_i \\ & \forall i, \quad \xi_i \geq 0. \end{aligned}$$

Eliminate the slack variables to rewrite this

Support Vector Machines

Maximize margin

Tradeoff between
the two terms

$$
\begin{aligned}
\min_{\mathbf{w},\xi} \quad & \boxed{\frac{1}{2}\mathbf{w}^T\mathbf{w}} + \boxed{C}\boxed{\sum_i \xi_i} \\
\text{s.t.} \quad \forall i, \quad & y_i \mathbf{w}^T \mathbf{x}_i \geq 1 - \xi_i \\
\forall i, \quad & \xi_i \geq 0.
\end{aligned}
$$

Minimize total slack (i.e allow
as few examples as possible
to violate the margin)

Eliminate the slack variables to rewrite this

$$
\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \max(0, 1 - y_i \mathbf{w}^T \mathbf{x}_i)
$$

This form is more interpretable

Maximizing margin and minimizing loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}$$

Maximize margin

Penalty for the prediction

Maximizing margin and minimizing loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}$$

Maximize margin

Penalty for the prediction

We can consider three cases

Maximizing margin and minimizing loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}$$

Maximize margin

Penalty for the prediction

We can consider three cases

- Example is correctly classified and is *outside* the margin: penalty = 0

Maximizing margin and minimizing loss

$$\underbrace{\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}}_{\text{Maximize margin}} + C\sum_i \underbrace{\boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}}_{\text{Penalty for the prediction}}$$

We can consider three cases

- Example is correctly classified and is *outside* the margin: penalty = 0
- Example is incorrectly classified: penalty = $1 - y_i\mathbf{w}^T\mathbf{x}_i$

Maximizing margin and minimizing loss

$$\underbrace{\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}}_{\text{Maximize margin}} + C\sum_i \underbrace{\boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}}_{\text{Penalty for the prediction}}$$

We can consider three cases

- Example is correctly classified and is ***outside*** the margin: penalty = 0
- Example is incorrectly classified: penalty = $1 - y_i\mathbf{w}^T\mathbf{x}_i$
- Example is correctly classified but within the margin: penalty =$1 - y_i\mathbf{w}^T\mathbf{x}_i$

Maximizing margin and minimizing loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}$$

Maximize margin

Penalty for the prediction

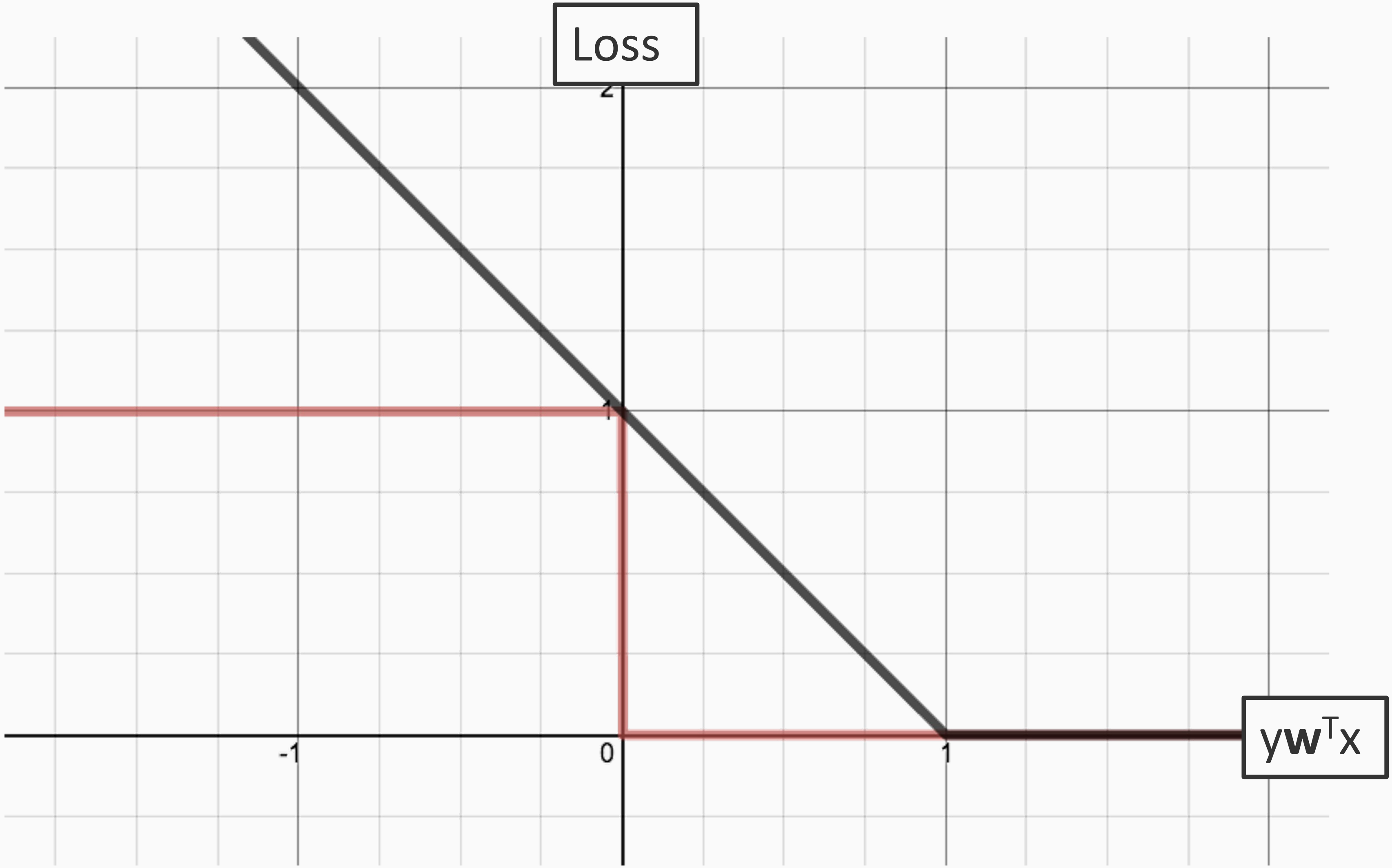
We can consider three cases

- Example is correctly classified and is *outside* the margin: penalty = 0
- Example is incorrectly classified: penalty = $1 - y_i\mathbf{w}^T\mathbf{x}_i$
- Example is correctly classified but within the margin: penalty =$1 - y_i\mathbf{w}^T\mathbf{x}_i$

This is the hinge loss function

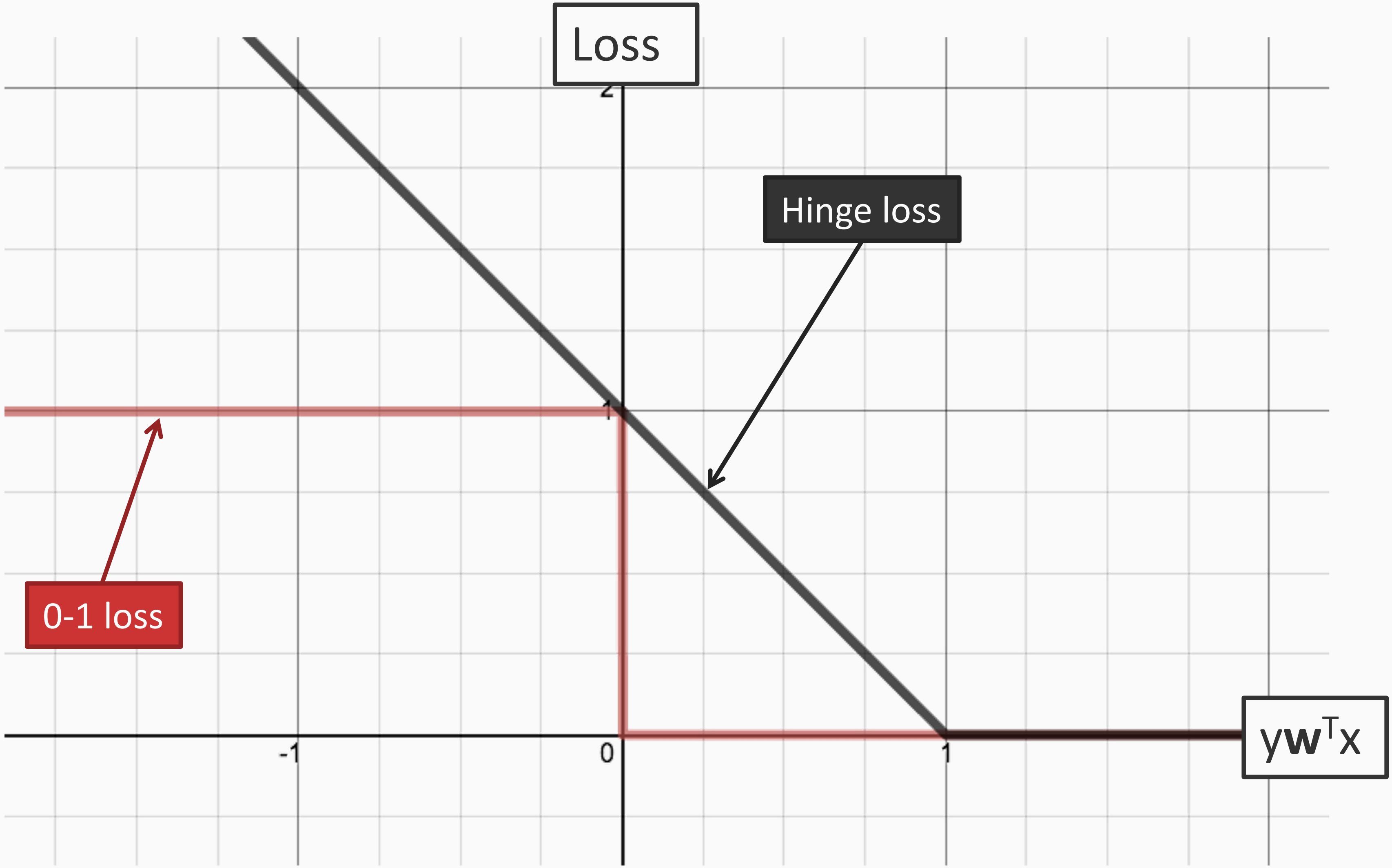
$$L_{Hinge}(y, \mathbf{x}, \mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

The Hinge Loss



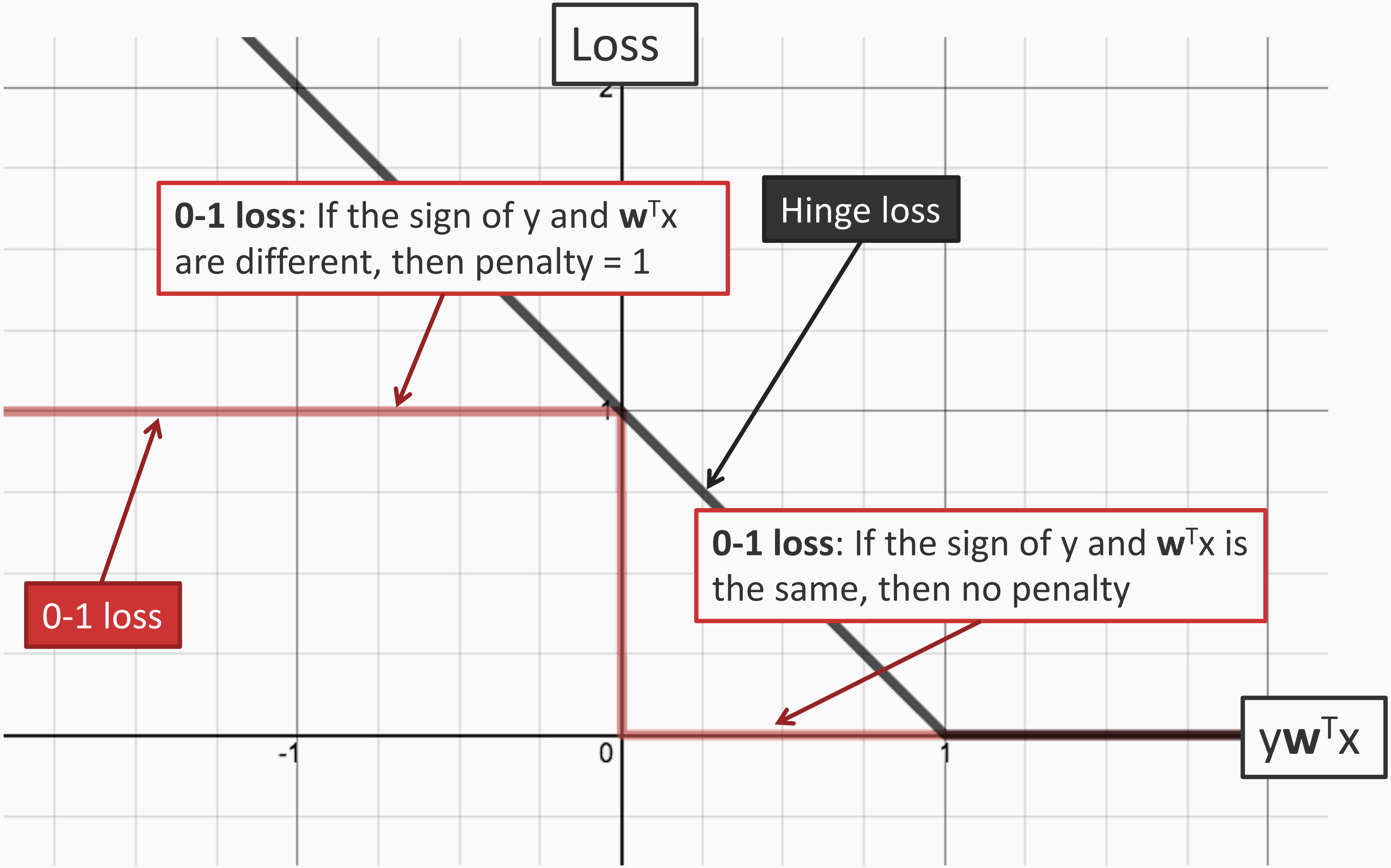
$$L_{Hinge}(y, \mathbf{x}, \mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

The Hinge Loss



The Hinge Loss

$$L_{Hinge}(y, \mathbf{x}, \mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

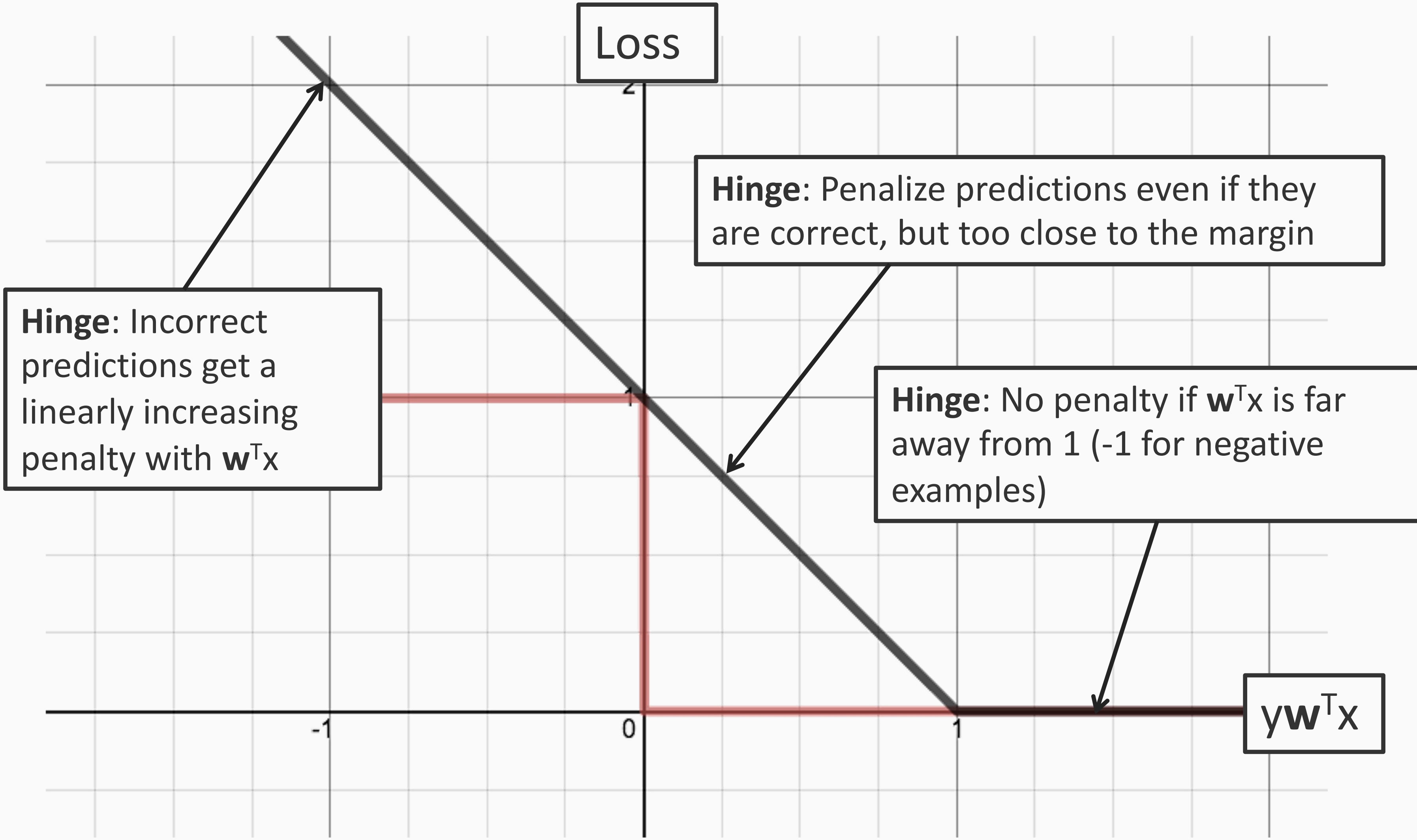


0-1 loss: If the sign of y and $\mathbf{w}^T x$ are different, then penalty = 1

0-1 loss: If the sign of y and $\mathbf{w}^T x$ is the same, then no penalty

$$L_{Hinge}(y, \mathbf{x}, \mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$$

The Hinge Loss



Hinge: Incorrect predictions get a linearly increasing penalty with $\mathbf{w}^T x$

Hinge: Penalize predictions even if they are correct, but too close to the margin

Hinge: No penalty if $\mathbf{w}^T x$ is far away from 1 (-1 for negative examples)

Maximizing margin and minimizing loss

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \boxed{\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)}$$

Maximize margin

Penalty for the prediction

We can consider three cases

- Example is correctly classified and is *outside* the margin: penalty = 0
- Example is incorrectly classified: penalty = $1 - y_i\mathbf{w}^T\mathbf{x}_i$
- Example is correctly classified but within the margin: penalty =$1 - y_i\mathbf{w}^T\mathbf{x}_i$

General learning principle

Risk minimization

Define the notion of “loss” over the training data as a function of a hypothesis

Learning = find the hypothesis that has lowest loss on the training data

General learning principle

Regularized risk minimization

Define a regularization function that penalizes over-complex hypothesis.

Capacity control gives better generalization

Define the notion of “loss” over the training data as a function of a hypothesis

Learning =
find the hypothesis that has lowest
[Regularizer + loss on the training data]

SVM objective function

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i} \max(0, 1 - y_i\mathbf{w}^T\mathbf{x})$$

Regularization term:
- Maximize the margin
- Imposes a preference over the hypothesis space and pushes for better generalization
- Can be replaced with other regularization terms which impose other preferences

Empirical Loss:
- Hinge loss
- Penalizes weight vectors that make mistakes
- Can be replaced with other loss functions which impose other preferences

SVM objective function

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i} \max(0, 1 - y_i\mathbf{w}^T\mathbf{x})$$

Regularization term:
- Maximize the margin
- Imposes a preference over the hypothesis space and pushes for better generalization
- Can be replaced with other regularization terms which impose other preferences

Empirical Loss:
- Hinge loss
- Penalizes weight vectors that make mistakes

- Can be replaced with other loss functions which impose other preferences

A hyper-parameter that controls the tradeoff between a large margin and a small hinge-loss