Support Vector Machine

Announcements

1. Prelim Conflict form is going out soon

2. Prelim practice: we will release previous semesters’ prelims w/ solutions

Outline for Today

1. Functional Margin & Geometric Margin

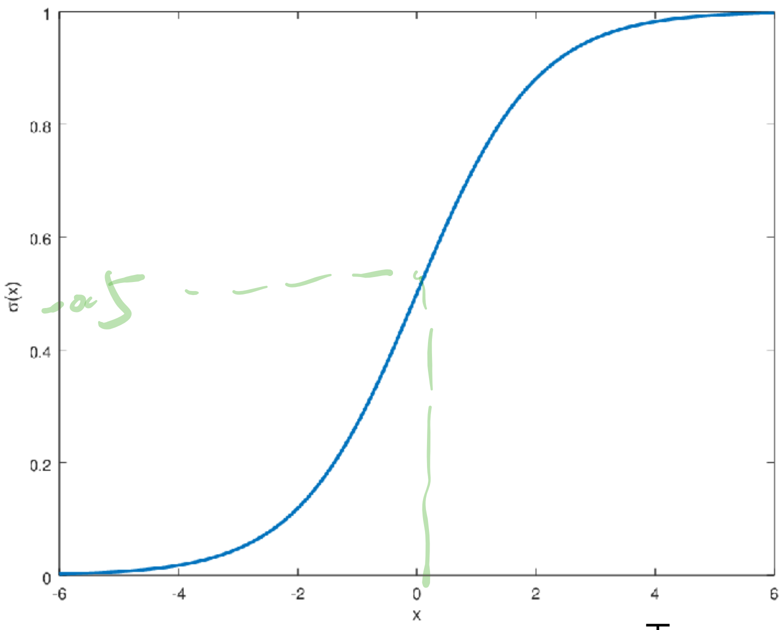
2. Support Vector Machine for separable data

3. SVM for non-separable data

Recall Logistic Regression

Binary classification with $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$

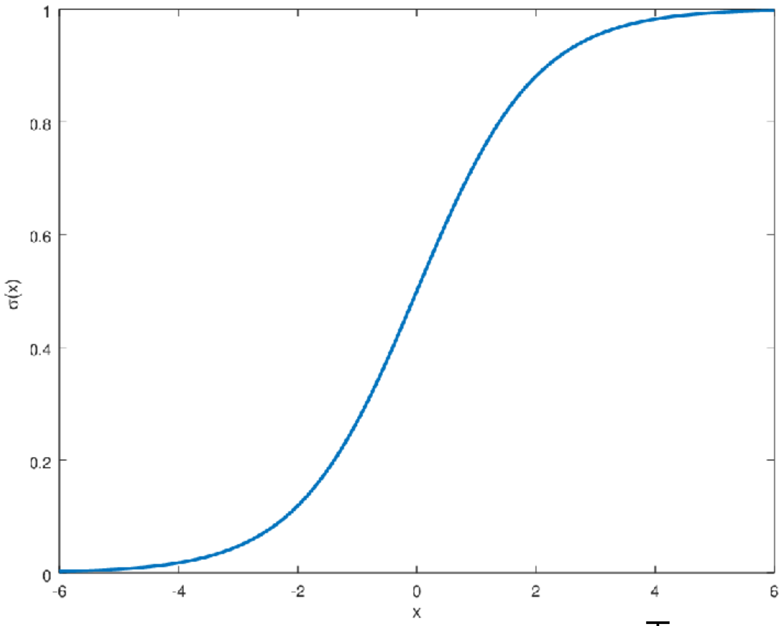


$z := y(w^\top x + b)$

Recall Logistic Regression

Binary classification with $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



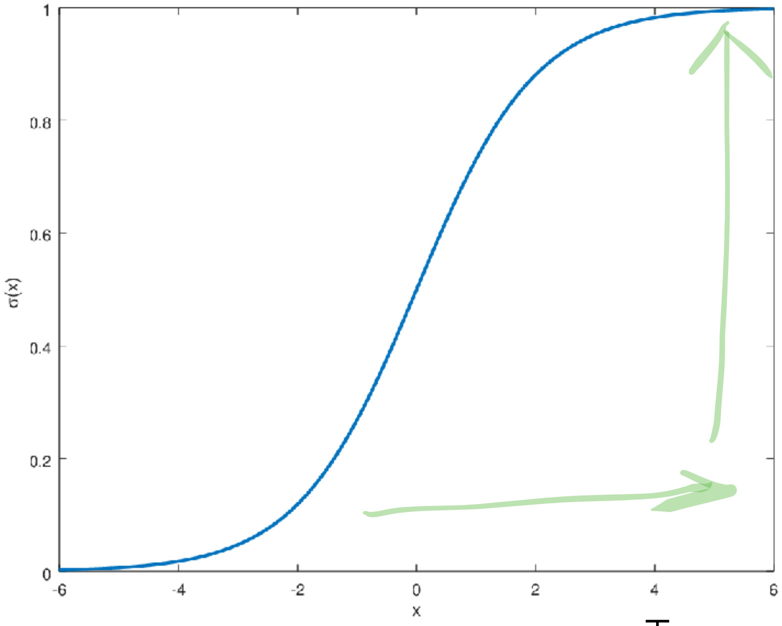
$z := y(w^\top x + b)$

Given (x, y), our model predict label y, if $P(y \mid x; w, b) > 0.5$, or equivalently $y(w^\top x + b) > 0$

Recall Logistic Regression

Binary classification with $\mathcal{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



$z := y(w^\top x + b)$

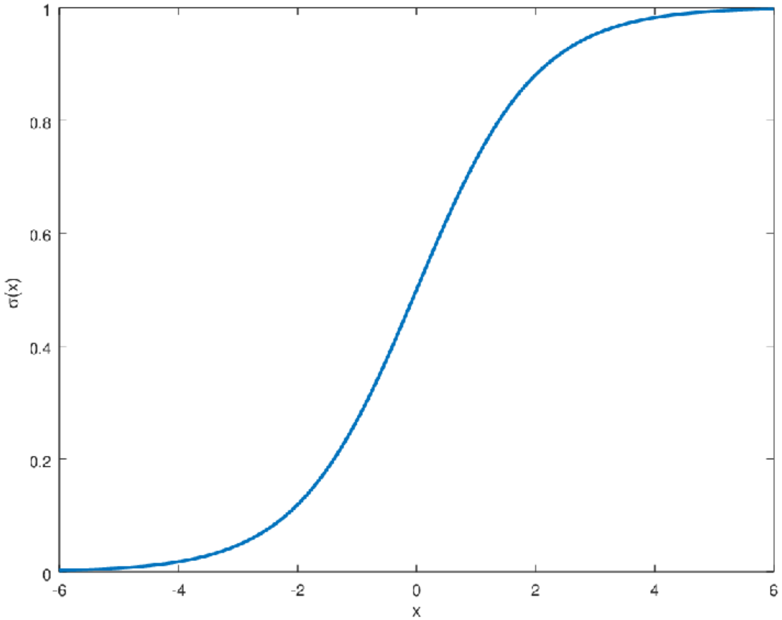
Given (x, y), our model predict label y, if $P(y \mid x; w, b) > 0.5$, or equivalently $y(w^\top x + b) > 0$

Larger $y(w^\top x + b) \rightarrow$ larger $P(y \mid x; w, b)$

Recall Logistic Regression

Binary classification with $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y\,|\,x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



$z := y(w^\top x + b)$

Given (x, y), our model predict label y, if

$P(y\,|\,x; w, b) > 0.5$, or equivalently $y(w^\top x + b) > 0$

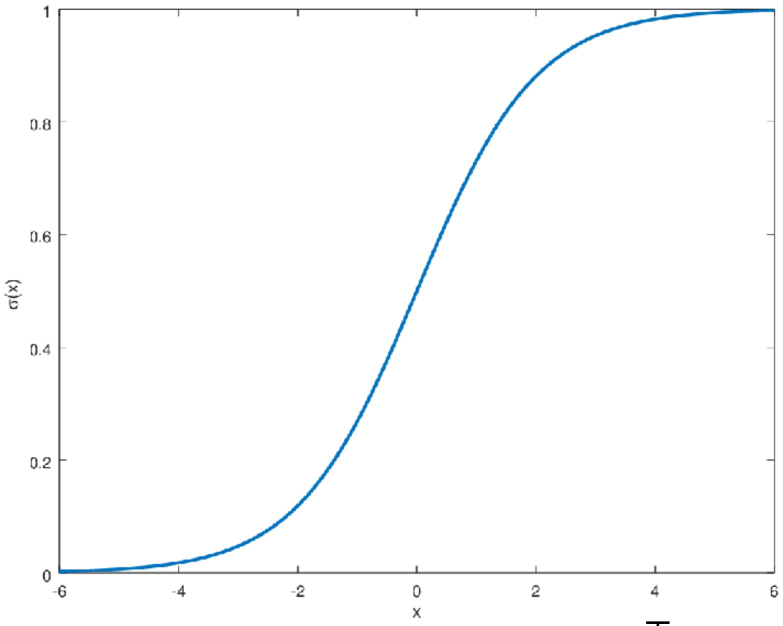
Larger $y(w^\top x + b)$ $\rightarrow$ larger $P(y\,|\,x; w, b)$

Functional margin
"confidence"

Recall Logistic Regression

Binary classification with $\mathcal{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y\,|\,x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$

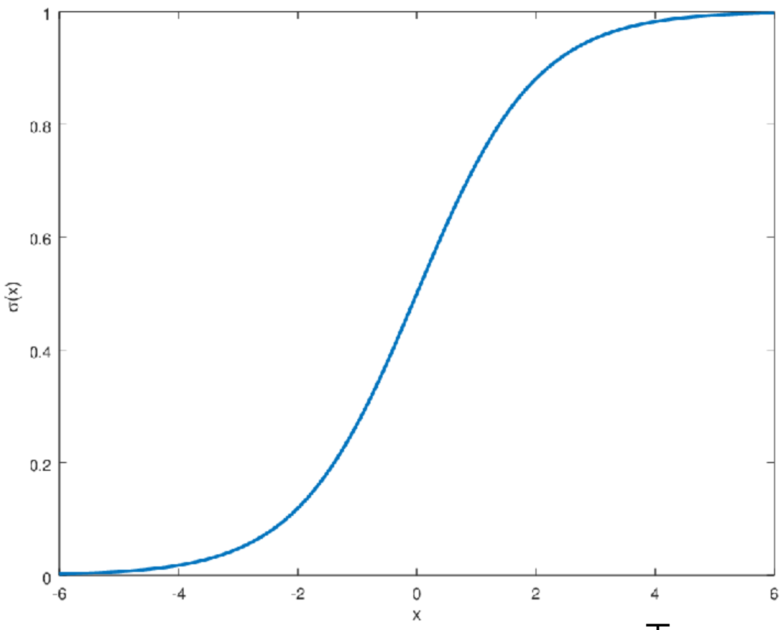


$z := y(w^\top x + b)$

Recall Logistic Regression

Binary classification with $\mathscr{D} = \{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



$z := y(w^\top x + b)$

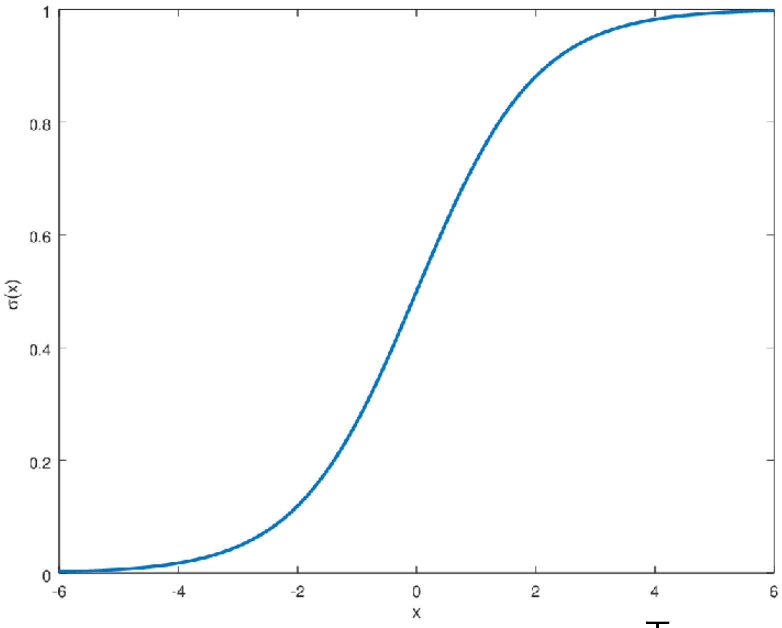
A good classifier should have large functional margin on training examples:

For all (x_i, y_i), $y_i(w^\top x_i + b) \gg 0$

Recall Logistic Regression

Binary classification with $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w, b) = \dfrac{1}{1 + \exp(-y(w^\top x + b))}$



$z := y(w^\top x + b)$

A good classifier should have large functional margin on training examples:

For all (x_i, y_i), $y_i(w^\top x_i + b) \gg 0$

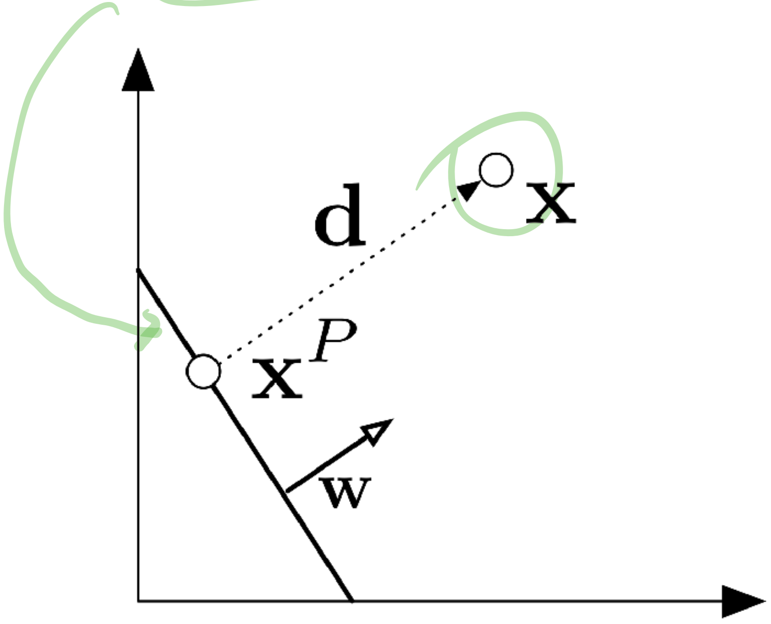
However, functional margin is NOT scale-invariant:

Consider $(2w, 2b)$: functional margin is doubled

Geometric Margin

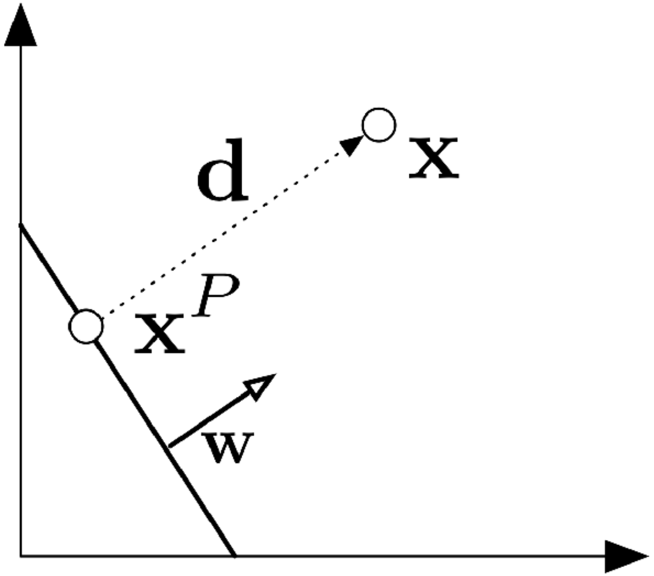
Hyperplane defined by (w, b), i.e.,

$$\{x : w^\top x + b = 0\}$$



Geometric Margin

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



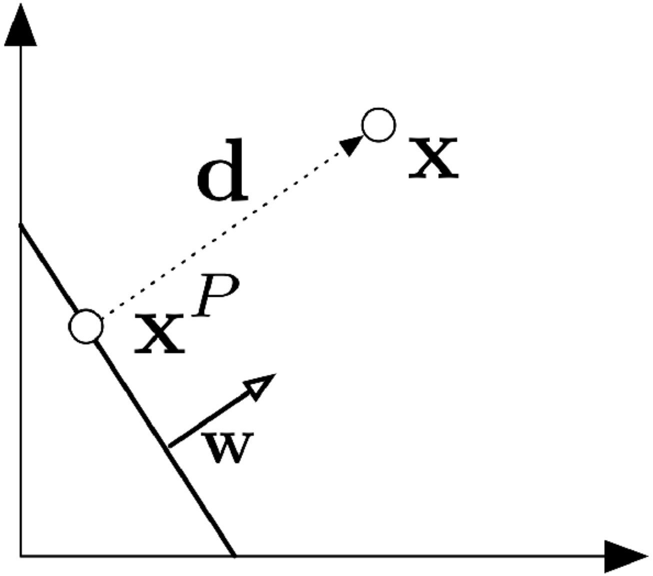
Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

Geometric Margin

Hyperplane defined by (w, b), i.e.,

$$\{x : w^\top x + b = 0\}$$



Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

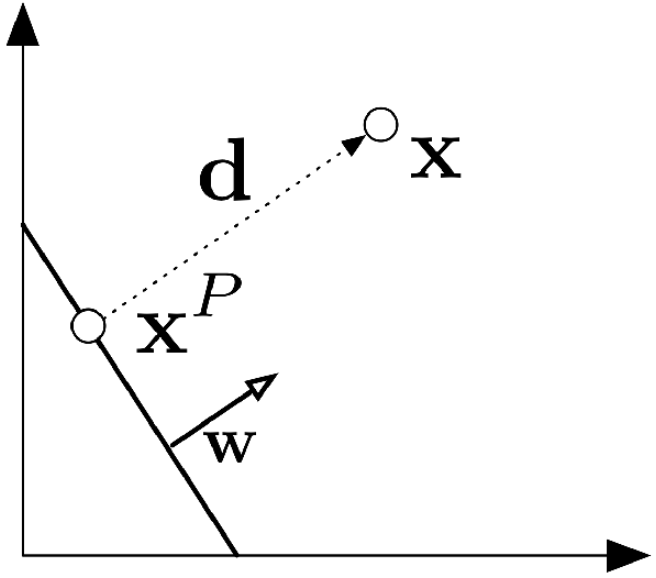
Fact 2. x^p is on the hyperplane:

$$w^\top x^P + b = 0$$

$x^p = x - \alpha w$

Geometric Margin

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

Fact 2. x^p is on the hyperplane:

$$w^\top x^P + b = 0$$

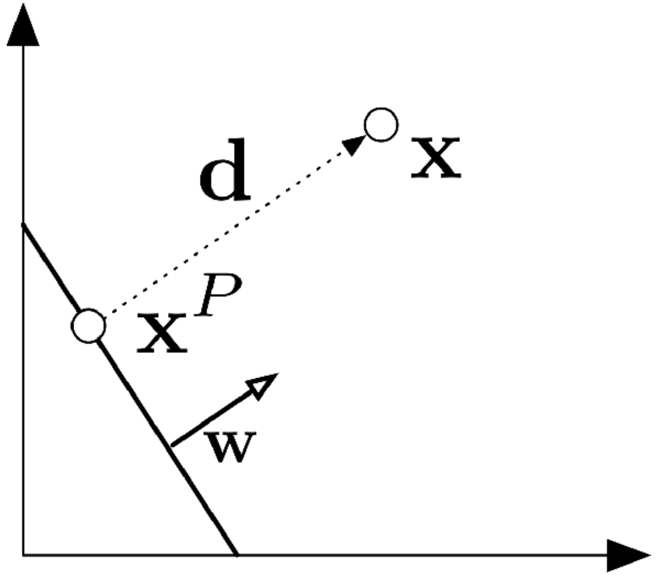
Fact 1 + fact 2 implies:

$$w^\top (x - \alpha w) + b = 0$$

x_p

Geometric Margin

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

Fact 2. x^p is on the hyperplane:

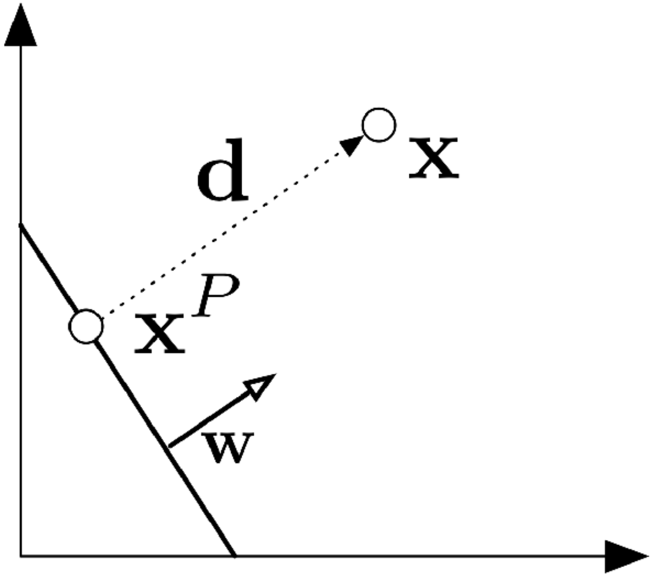
$$w^\top x^P + b = 0$$

Fact 1 + fact 2 implies:

$$w^\top (x - \alpha w) + b = 0 \ \rightarrow \alpha = (w^\top x + b)/\|w\|_2^2$$

Geometric Margin

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

Fact 2. x^p is on the hyperplane:

$$w^\top x^P + b = 0$$

Fact 1 + fact 2 implies:

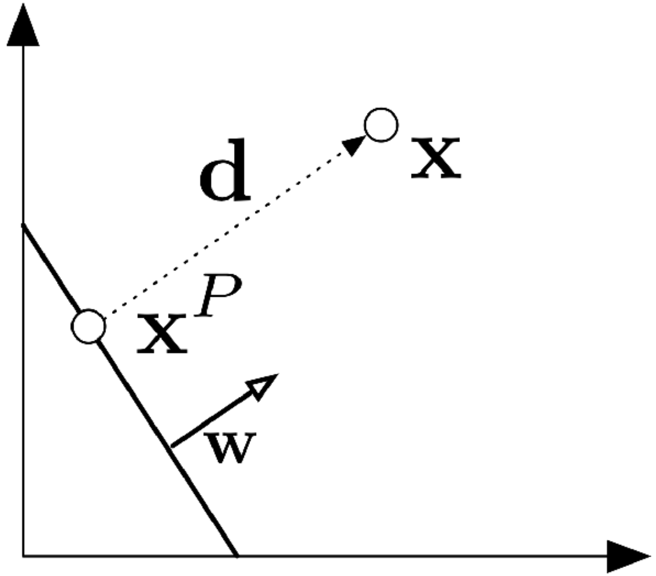
$$w^\top(x - \alpha w) + b = 0 \;\rightarrow\; \alpha = (w^\top x + b)/\|w\|_2^2$$

Final step:

$$\|d\|_2 = \|x - x^p\|_2 = \|\alpha w\|_2$$

Geometric Margin

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



Fact 1. $x - x^P$ is parallel to w:

$$x - x^p = \alpha w$$

Fact 2. x^p is on the hyperplane:

$$w^\top x^P + b = 0$$

Fact 1 + fact 2 implies:

$$w^\top (x - \alpha w) + b = 0 \;\rightarrow \alpha = (w^\top x + b)/\|w\|_2^2$$

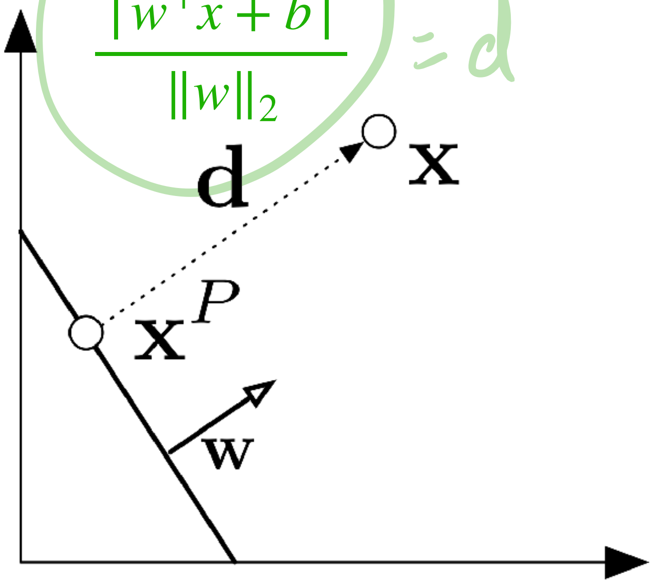
Final step:

$$\|d\|_2 = \|x - x^p\|_2 = \|\alpha w\|_2 \;= \frac{|w^\top x + b|}{\|w\|_2}$$

Geometric Margin is Scale Invariant

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$

$$\frac{|w^\top x + b|}{\|w\|_2} = d$$

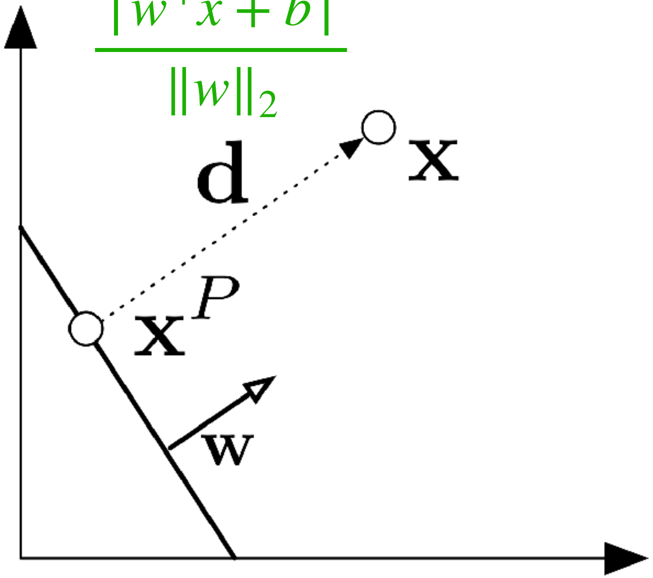


We scale (w, b) by a constant $\gamma \in \mathbb{R}^+$

Geometric Margin is Scale Invariant

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$

$$\frac{|w^\top x + b|}{\|w\|_2}$$



We scale (w, b) by a constant $\gamma \in \mathbb{R}^+$

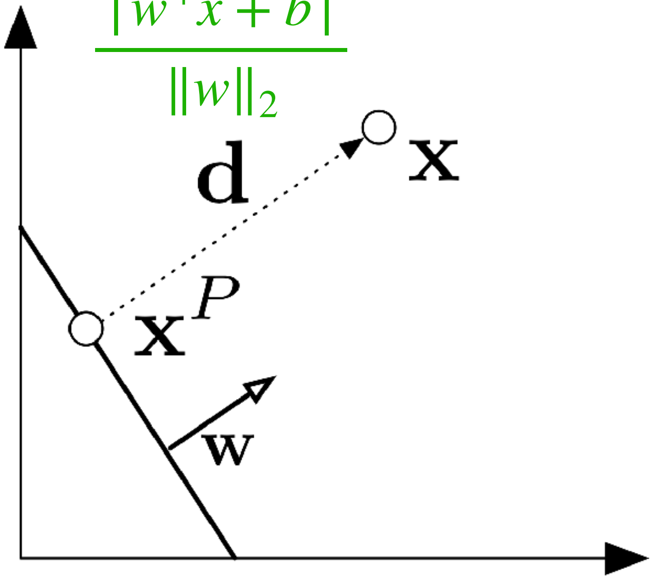
Q: is the hyperplane defined by $(\gamma w, \gamma b)$ different?

x: $w^T x + b = 0 \Rightarrow (\gamma w)^T x + \gamma b = 0$ $\Leftarrow$

Geometric Margin is Scale Invariant

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$

$$\frac{|w^\top x + b|}{\|w\|_2}$$



We scale (w, b) by a constant $\gamma \in \mathbb{R}^+$

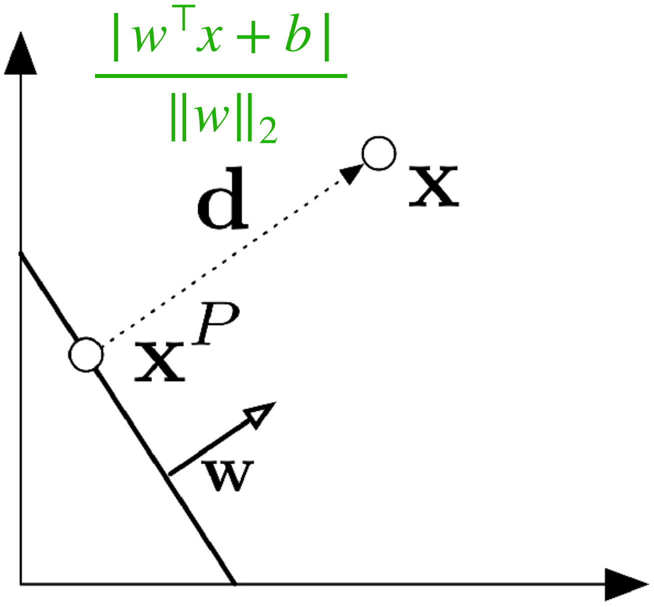
Q: is the hyperplane defined by $(\gamma w, \gamma b)$ different?

Q: does the margin change?

$$\frac{|\gamma w^\top x + \gamma b|}{\|\gamma w\|_2}$$

Geometric Margin is Scale Invariant

Hyperplane defined by (w, b), i.e.,
$\{x : w^\top x + b = 0\}$



We scale (w, b) by a constant $\gamma \in \mathbb{R}^+$

Q: is the hyperplane defined by $(\gamma w, \gamma b)$ different?

Q: does the margin change?

Hyperplane & Geometric margin are **scale invariant**!

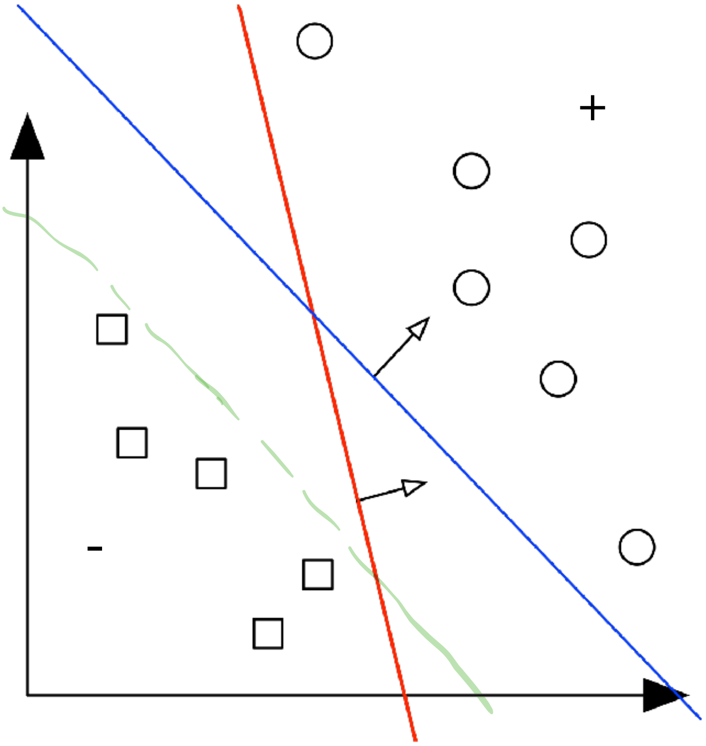
Outline for Today

1. Functional Margin & Geometric Margin

2. Support Vector Machine for separable data

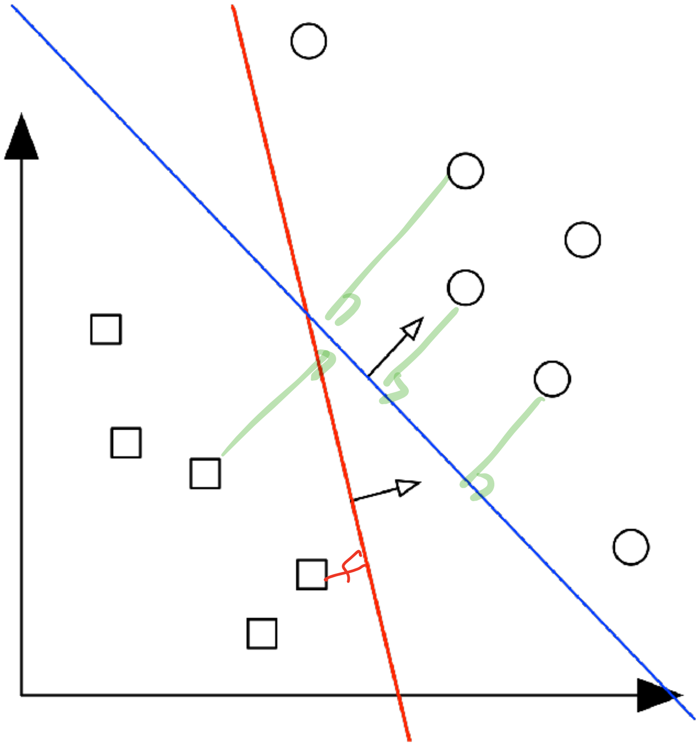
3. SVM for non-separable data

Which linear classifier is Better?



Both hyperplanes correctly separate the data

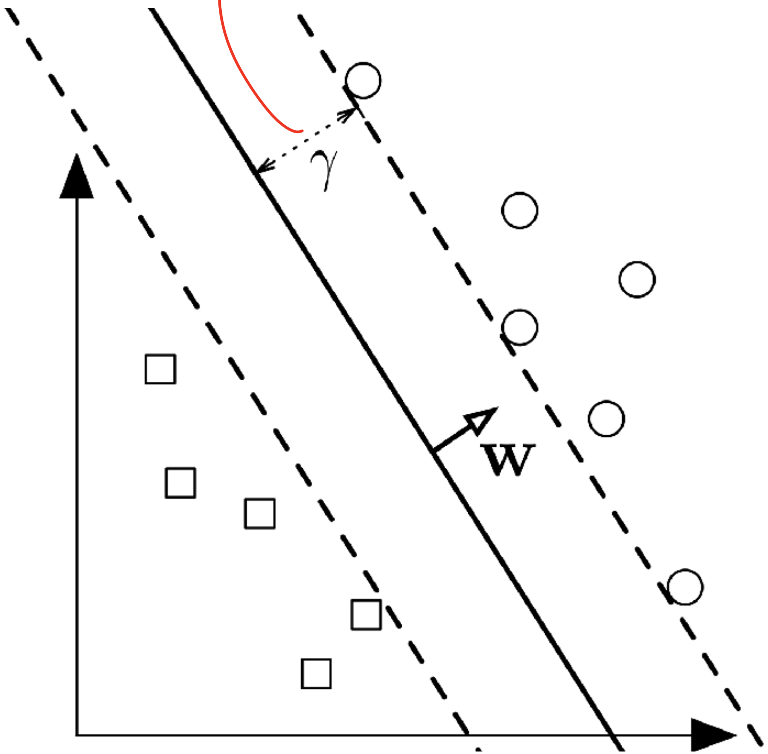
Max Margin Classifier



The Goal of SVM:

Find a hyperplane that has the largest Geometric margin

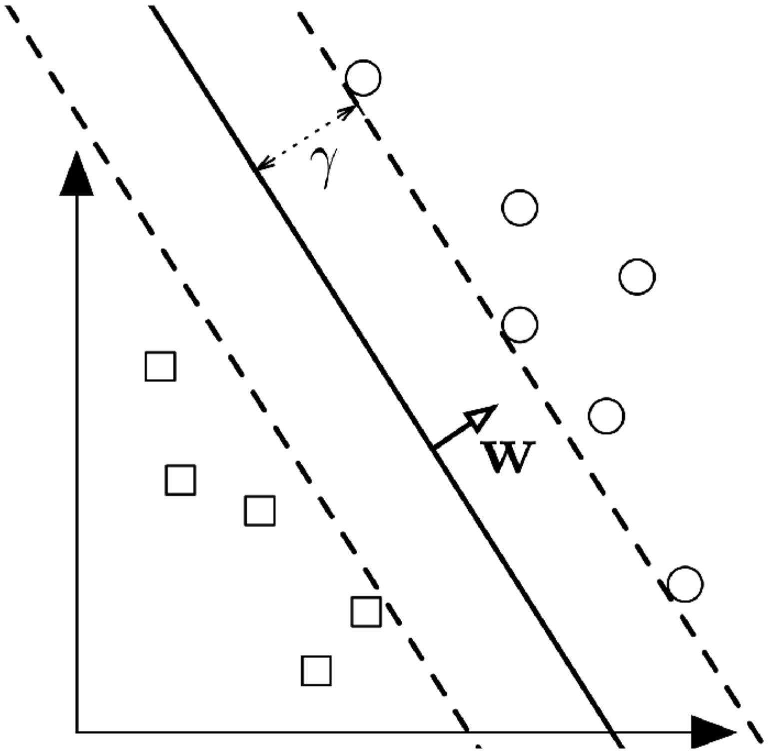
Max Margin Classifier



Given a linearly separable dataset $\{x_i, y_i\}_{i=1}^{n}$, the minimum geometric margin is defined as

$$\gamma(w, b) := \min_{x_i \in \mathcal{D}} \frac{|x_i^\top w + b|}{\|w\|_2}$$

Max Margin Classifier

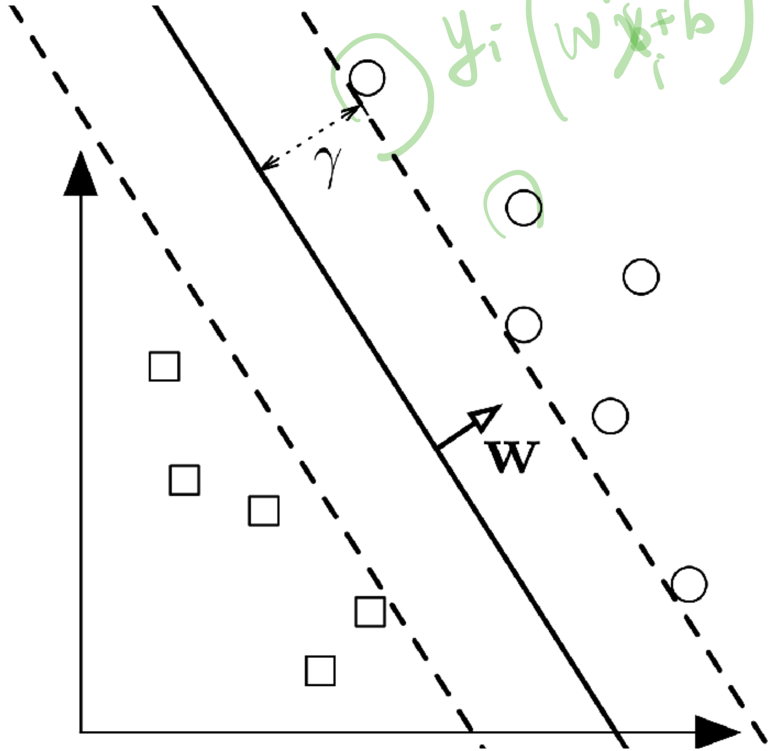


Given a linearly separable dataset $\{x_i, y_i\}_{i=1}^n$, the minimum geometric margin is defined as

$$\gamma(w, b) := \min_{x_i \in \mathcal{D}} \frac{|x_i^\top w + b|}{\|w\|_2}$$

Goal: we want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

Max Margin Classifier

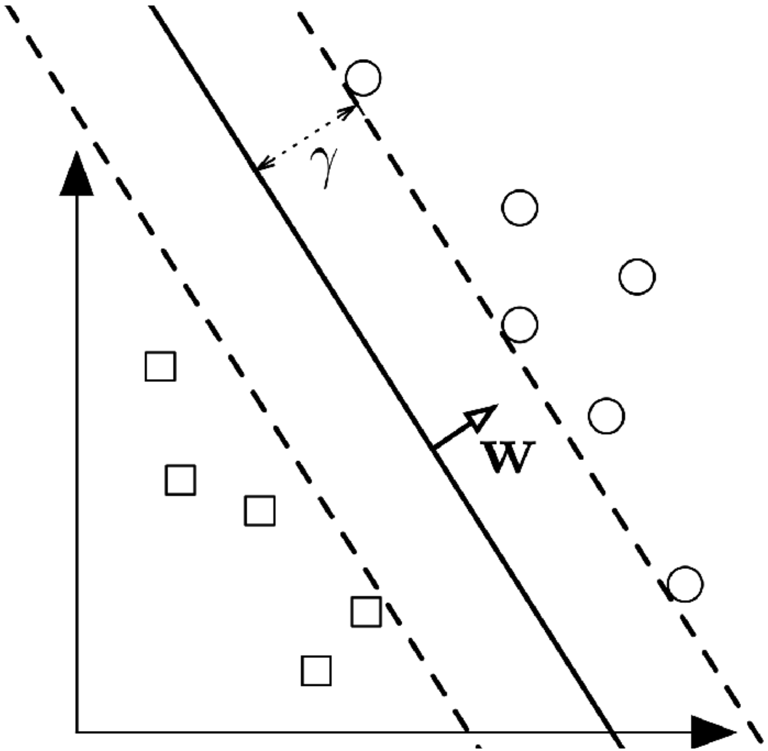


We want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

$$\max_{w,b} \gamma(w, b)$$

s.t. $\forall i,\ y_i(w^\top x_i + b) \geq 0$

Max Margin Classifier



We want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

$$\max_{w,b} \gamma(w, b)$$

s.t. $\forall i,\ y_i(w^\top x_i + b) \geq 0$

Plug in the def of $\gamma(w, b)$:

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} |w^\top x_i + y_i|$$

s.t. $\forall i,\ y_i(w^\top x_i + b) \geq 0$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} |w^\top x_i + y_i|$$

$$\text{s.t.} \forall i,\ y_i(w^\top x_i + b) \geq 0$$

Geometric

Recall that margin & hyperplane is scale invariant

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} |w^\top x_i + y_i|$$

s.t. $\forall i,\ y_i(w^\top x_i + b) \geq 0$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have

$$\min_{x_i} |w^\top x + b| = 1$$

$$(w, b) \rightarrow \min_{x_i} |w^\top x + b| = c$$

$$\tilde{w} = w/c, \quad \tilde{b} = b/c \rightarrow (\tilde{w}, \tilde{b}) \rightarrow \min_{x_i} |\tilde{w}^\top x + \tilde{b}| = 1$$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} | w^\top x_i + y_i |$$

$$\text{s.t.} \ \forall i, \ y_i(w^\top x_i + b) \geq 0$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have
$$\min_{x_i} | w^\top x + b | = 1$$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} | w^\top x_i + b | = 1$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2} \min_{x_i} | w^\top x_i + y_i |$$

$$\text{s.t. } \forall i,\ y_i(w^\top x_i + b) \geq 0$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have

$$\min_{x_i} | w^\top x + b | = 1$$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} | w^\top x_i + b | = 1$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\text{s.t.} \; \forall i, \; y_i(w^\top x_i + b) \geq 0$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have

$$\min_{x_i} |w^\top x + b| = 1$$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} |w^\top x_i + b| = 1$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2}$$

$$\text{s.t.}\ \forall i,\ y_i(w^\top x_i + b) \geq 0$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have

$$\min_{x_i} |w^\top x + b| = 1$$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} |w^\top x_i + b| = 1$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\max_{w,b} \frac{1}{\|w\|_2}$$

$$\text{s.t.} \forall i, \; y_i(w^\top x_i + b) \geq 0$$

$$\min_i |w^\top x_i + b| = 1$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have $\min_{x_i} |w^\top x + b| = 1$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} |w^\top x_i + b| = 1$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\text{s.t.}\ \forall i,\ y_i(w^\top x_i + b) \geq 0$$

$$\min_i |w^\top x_i + b| = 1$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have $\min_{x_i} |w^\top x + b| = 1$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} |w^\top x_i + b| = 1$

SVM for separable data: Max Margin Classifier

We want to find (w, b) s.t. it separates the data, and maximize $\gamma(w, b)$

$$\min_{w,b} \|w\|_2^2$$

$$\text{s.t.} \; \forall i, \; y_i(w^\top x_i + b) \geq 0$$

$$\min_i |w^\top x_i + b| = 1$$

Recall that margin & hyperplane is scale invariant

For any (w, b), we can always scale it by some constant to have

$$\min_{x_i} |w^\top x + b| = 1$$

Without loss of generality, let's just focus on such (w, b) pairs with $\min_{x_i} |w^\top x_i + b| = 1$

SVM for separable data: Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

s.t. $\forall i,\ y_i(w^\top x_i + b) \geq 0$

$$\min_i |w^\top x_i + b| = 1$$

SVM for separable data: Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

$$\text{s.t.} \; \forall i, \; y_i(w^\top x_i + b) \geq 0$$

$$\min_i |w^\top x_i + b| = 1$$

We can further simplify the constraint

SVM for separable data: Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

s.t. $\forall i,\ y_i(w^\top x_i + b) \geq 0$

$$\min_i |w^\top x_i + b| = 1$$

We can further simplify the constraint

$$\min_{w,b} \|w\|_2^2$$

$$\forall i:\ \ y_i(w^\top x_i + b) \geq 1$$

SVM for separable data: Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

s.t. $\forall i, y_i(w^\top x_i + b) \geq 0$

$$\min_i |w^\top x_i + b| = 1$$

when $y_i(w^\top x_i + b) \geq 0$

$\Rightarrow |w^\top x_i + b| = y_i(w^\top x_i + b)$

We can further simplify the constraint

quadratic convex

$$\min_{w,b} \|w\|_2^2$$

$$\forall i : \ y_i(w^\top x_i + b) \geq 1$$

$\rightarrow$ linear

You will prove that in HW4!

$\min_i y_i(w^\top x_i + b) = 1$

$\Rightarrow \forall i, y_i(w^\top x_i + b) \geq 1$

Summary for Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1$$

Summary for Max Margin Classifier

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1$$

Not only linearly separable, but also
has functional margin no less than 1

Summary for Max Margin Classifier

Avoids "cheating" (i.e., scale w, b up by large constant)

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \;\; y_i(w^\top x_i + b) \geq 1$$

Not only linearly separable, but also has functional margin no less than 1

Summary for Max Margin Classifier

Avoids "cheating" (i.e., scale w, b up by large constant)

$$\min_{w,b} \|w\|_2^2$$

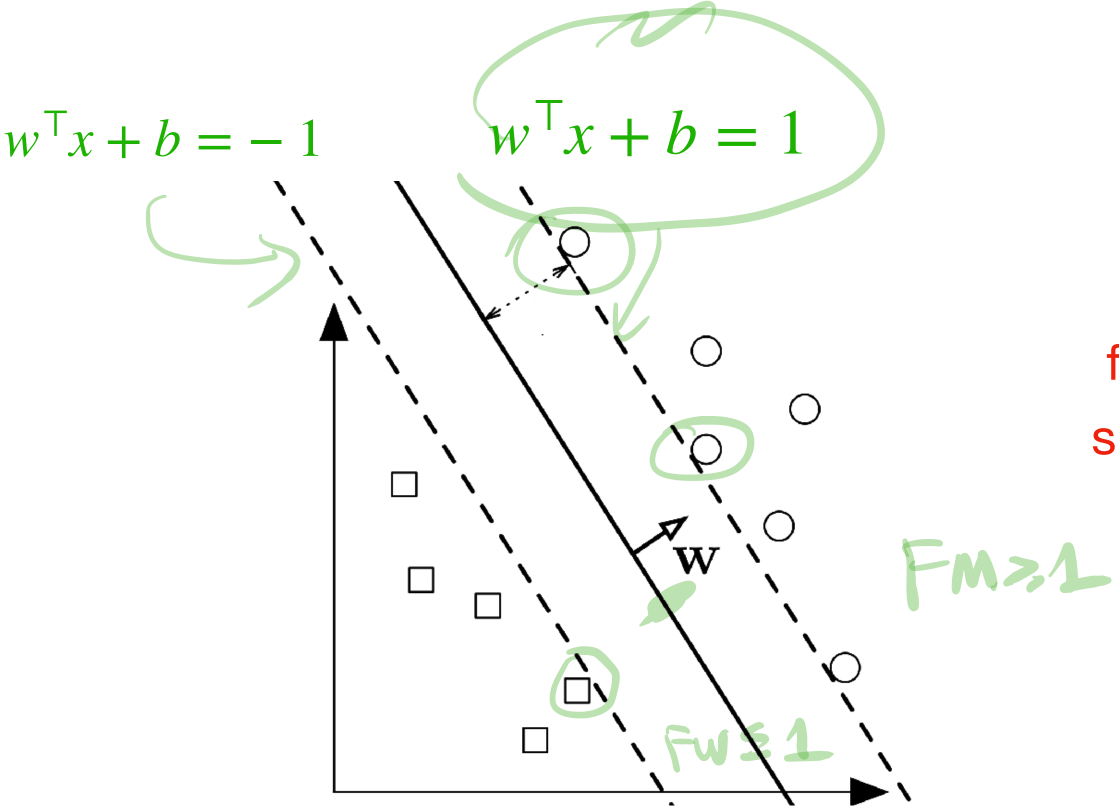
$$\forall i: \ y_i(w^\top x_i + b) \geq 1$$

Not only linearly separable, but also has functional margin no less than 1

Always remember **where we started**:

We want to find (w, b) s.t. it separates the data, and maximizes $\gamma(w, b)$

Support Vectors



for the optimal (w, b) pair, points x_i such that $y_i(w^\top x_i + b) = 1$ are called **support vectors**

Outline for Today

1. Functional Margin & Geometric Margin

2. Support Vector Machine for separable data

3. SVM for non-separable data

SVM for non-separable data

If data is not linearly separable, then **there is no** (w, b)
can satisfy $\forall i: \; y_i(w^\top x_i + b) \geq 1$

linear-separable $\Leftrightarrow \exists (w, b)$
s.t. $y_i(w^\top x_i + b) \geq 0, \forall i$

SVM for non-separable data

If data is not linearly separable, then **there is no** (w, b) can satisfy $\forall i: \; y_i(w^\top x_i + b) \geq 1$

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

SVM for non-separable data

If data is not linearly separable, then **there is no** (w, b) can satisfy $\forall i: \; y_i(w^\top x_i + b) \geq 1$

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \forall i$$

SVM for non-separable data

If data is not linearly separable, then **there is no** (w, b) can satisfy $\forall i: \; y_i(w^\top x_i + b) \geq 1$

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \forall i$$

$\Rightarrow y_i(w^\top x_i + b) \geq -\infty$

This always has feasible solutions (e.g., take $\xi_i = +\infty$)

SVM for non-separable data

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

$$\forall i: \ y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

SVM for non-separable data

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

We still want our margin to be somewhat large, i.e., we want slack variables to be as small as possible

SVM for non-separable data

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

We still want our margin to be somewhat large, i.e., we want slack variables to be as small as possible

$$\min_{w,b,\xi} \|w\|_2^2 + c \sum_{i=1}^{n} \xi_i$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

SVM for non-separable data

Idea: introducing slack variables to relax the constraint, i.e., find (w, b), st,

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

We still want our margin to be somewhat large, i.e., we want slack variables to be as small as possible

$$\min_{w,b,\xi} \|w\|_2^2 + c\sum_{i=1}^{n} \xi_i$$

Penalizing large slacks

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

SVM for non-separable data

$$\min_{w,b,\xi_i} \|w\|_2^2 + c\sum_{i=1}^{n} \xi_i$$

Penalizing large slacks

$$\forall i: \ \ y_i(w^\top x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0$$

We can turn this constrained opt to a unconstraint opt w/ a single objective.

SVM for non-separable data

$$\min_{w,b,\xi_i} \|w\|_2^2 + c\sum_{i=1}^{n} \xi_i$$

Penalizing large slacks

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

We can turn this constrained opt to a unconstraint opt w/ a single objective.

Q: For any fixed (w, b) pair, how to set ξ_i, such that the obj is minimized?

$$\begin{cases} \xi_i \geq 1 - y_i(w^\top x_i + b) \\ \xi_i \geq 0 \end{cases} \Rightarrow \begin{cases} \xi_i = 1 - y_i(w^\top x_i + b) & \text{if } 1 - y_i(w^\top x_i + b) \geq 0 \\ \xi_i = 0 & \text{else} \end{cases}$$

SVM for non-separable data

$$\min_{w,b,\xi_i} \|w\|_2^2 + c\sum_{i=1}^{n} \xi_i$$

Penalizing large slacks

$$\forall i: \; y_i(w^\top x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

We can turn this constrained opt to a unconstraint opt w/ a single objective.

Q: For any fixed (w, b) pair, how to set ξ_i, such that the obj is minimized?

A: set $\xi_i = \max\{0, 1 - y_i(w^\top x_i + b)\}$

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

$$\xi_i = \max\left\{0,\ 1 - y_i(w^\top x_i + b)\right\}$$

SVM for non-separable data

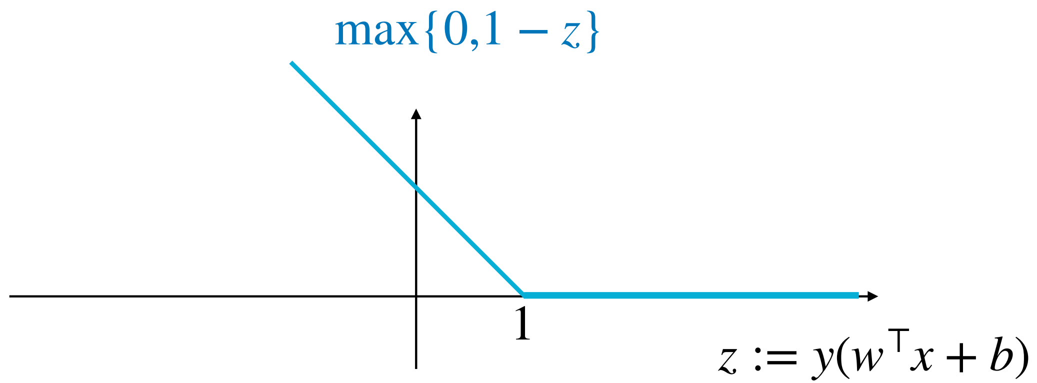
$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Hinge loss

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

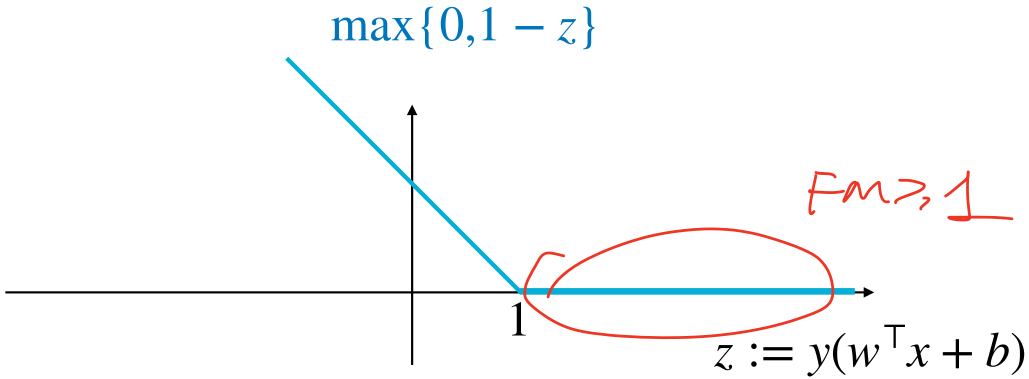
Hinge loss



SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Hinge loss



Hinge loss starts penalizing when functional margin falls below 1

SVM for non-separable data

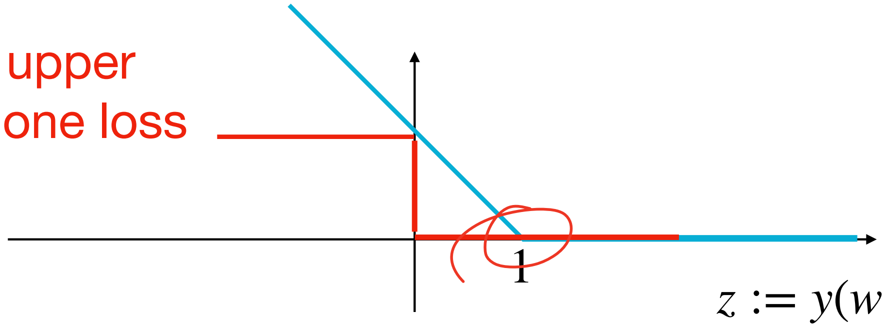
$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Hinge loss

→ Convex wrt (w, b)

$\max\{0, 1 - z\}$

Hinge loss upper bounds zero-one loss



$z := y(w^\top x + b)$

Hinge loss starts penalizing when functional margin falls below 1

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

Set $C = +\infty$

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

When $c \to +\infty$:

forcing $y_i(w^\top x_i + b) \geq 1$ for as many data points as possible

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

When $c \to +\infty$:

forcing $y_i(w^\top x_i + b) \geq 1$ for as many data points as possible

When $c \to 0^+$:

The solution $w \to \mathbf{0}$ (i.e., we do not care about hinge loss part)

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

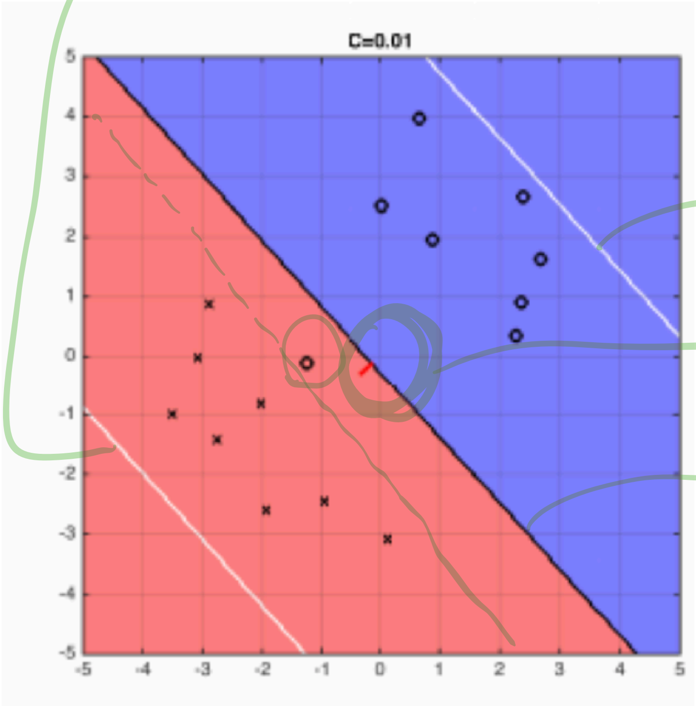
Trades off $\|w\|_2^2$ and functional margins over data

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 0.01

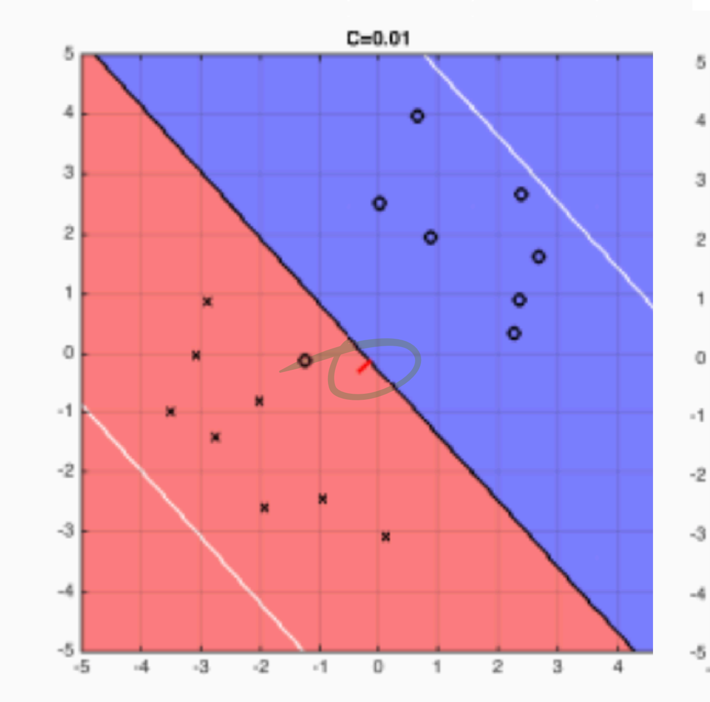


SVM for non-separable data

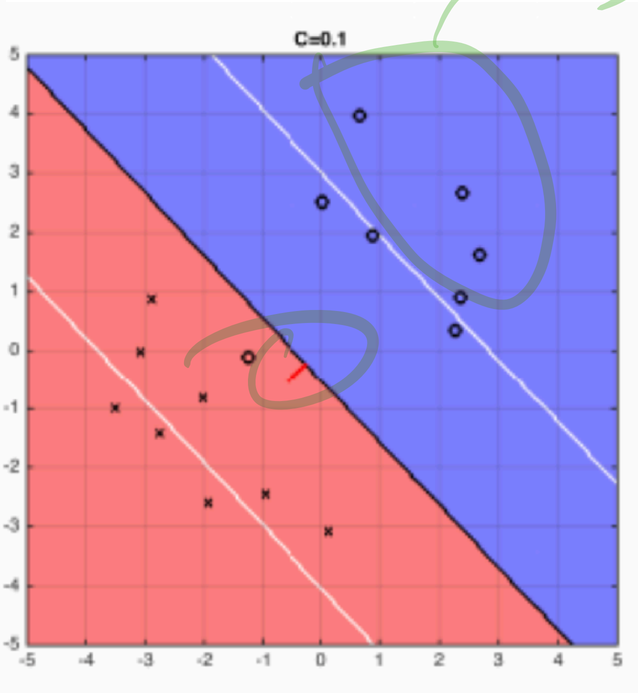
$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 0.01



C = 1



⇒ FM ≥ 1

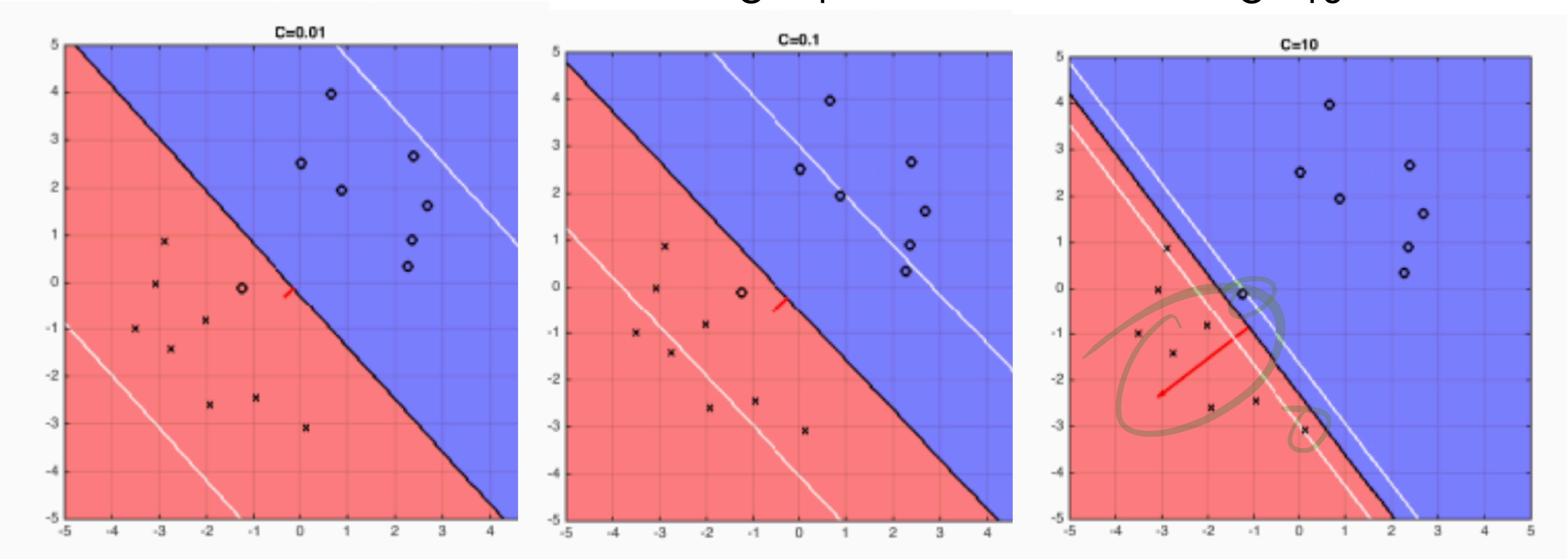
⇒ hingeloss = 0

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 0.01 C = 1 C =10

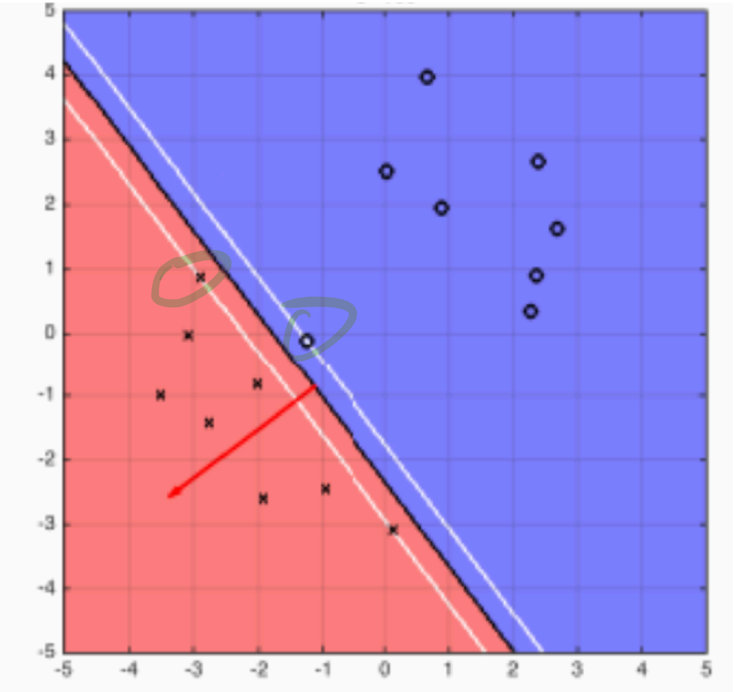


SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 100

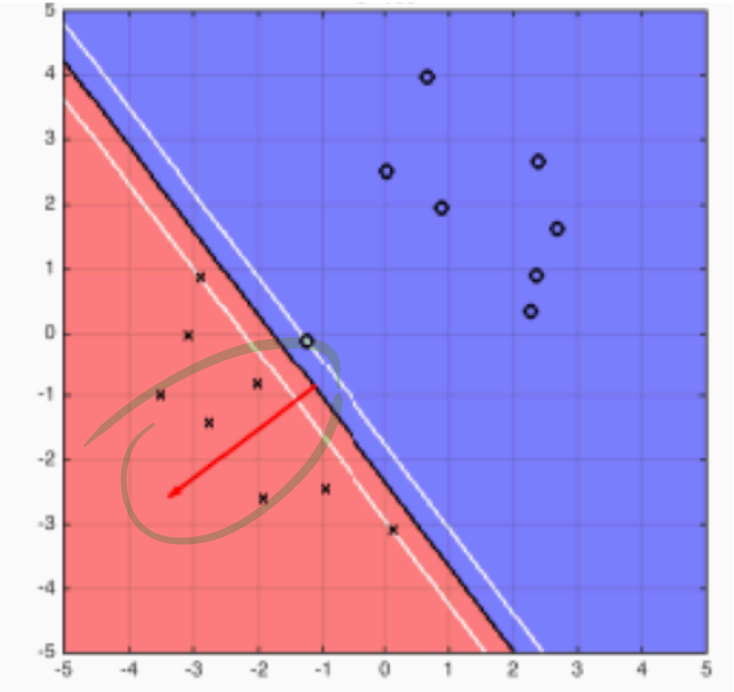


SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 100



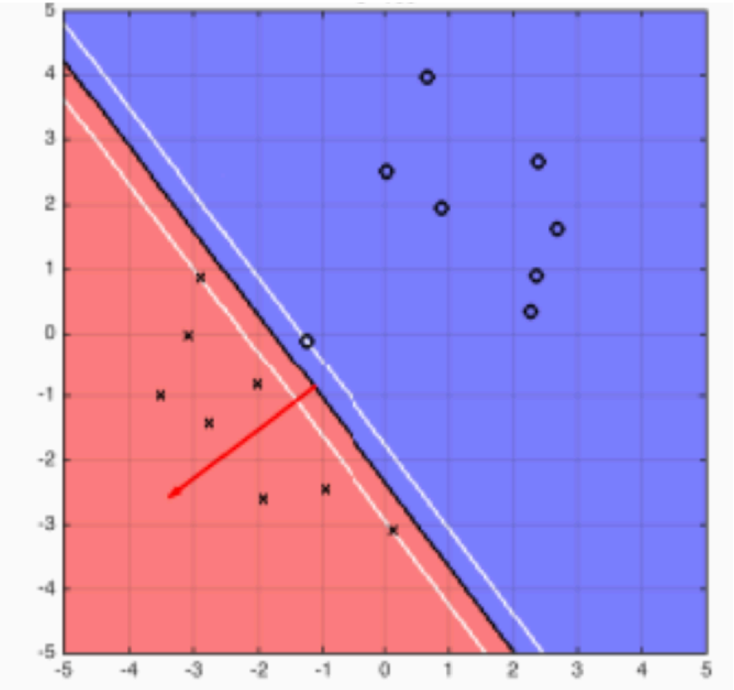
all examples have zero Hinge loss, but w has large norm

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c\sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 100



all examples have zero Hinge loss, but w has large norm

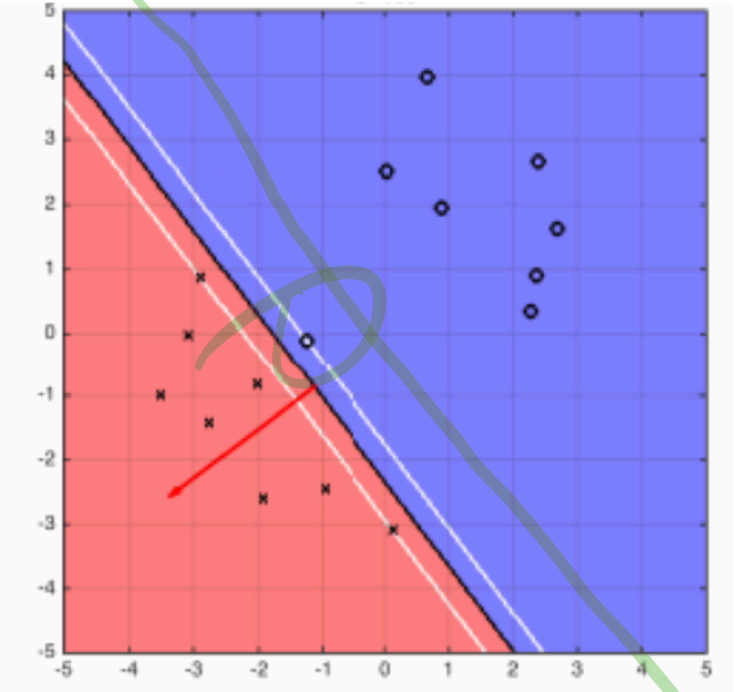
Bad geometric margin but good functional margin (achieved by “cheating”)

SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\}$$

Trades off $\|w\|_2^2$ and functional margins over data

C = 100



all examples have zero Hinge loss, but w has large norm

Bad geometric margin but good functional margin (achieved by "cheating")

Potentially overfitting to the noise, not a good classifier in test time maybe

Summary for today

1. SVM for linearly separable data

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1$$

Summary for today

1. SVM for linearly separable data

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \;\; y_i(w^\top x_i + b) \geq 1$$

2. SVM for non-separable data

$$\min_{w,b} \|w\|_2^2 + c \sum_{i=1}^{n} \boxed{\max\left\{0, 1 - y_i(w^\top x_i + b)\right\}}$$

Hinge loss