MAC 0460 / 5832
Introduction to Machine Learning

17 — Support Vector Machines (SVM)

• hyperplane • margin • margin violation •
• QP problems • dual problems • kernel trick •

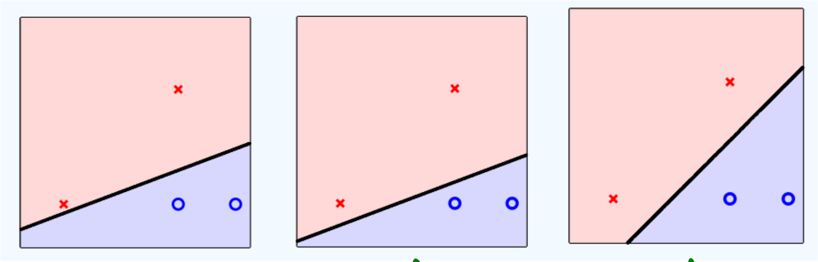
IME/USP (16/06/2021)

SVM roadmap

- Binary linear classification: The linearly separable case
- hard-margin SVM: Maximum margin formulation
- Binary linear classification: The non-linearly separable case
- soft-margin SVM: allows margin violation
- hard-margin/soft-margin SVM is a QP problem
- Dual of hard-margin/soft-margin SVM is also QP
- How to solve QP problems
- Non-linear classification: the kernel trick

Linearly separable case

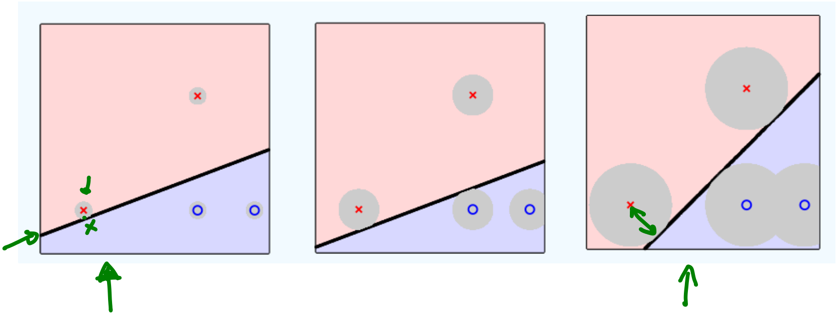
Given a linearly separable D, a linear decision boundary separating negatives from positives can be obtained using, for instance, **PLA** or **logistic regression**



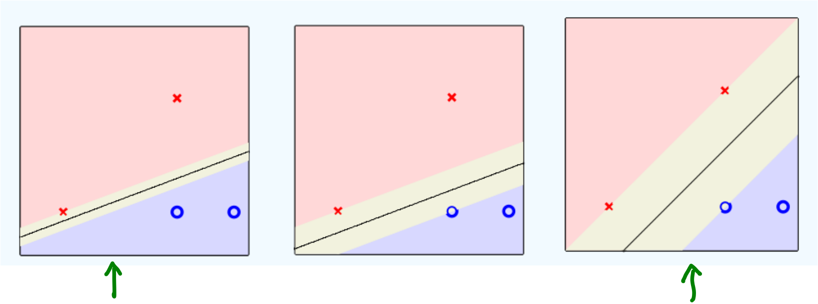
Is there one that is preferable than others ?

Intuition

Depending on where the separating line is, it is more or less robust to noise



Maximum margin



Any of these lines separate the negatives from the positives
They have margins of different sizes

Problem formulation

How to find the separating **hyperplane that maximizes the margin** ?

In **SVM**, this is achieved by formulating the problem as a quadratic programming (QP) optimization problem

QP: optimization of quadratic functions with linear constraints on the variables

Notations

Previous Chapters

$\mathbf{x} \in \{1\} \times \mathbb{R}^d; \ \mathbf{w} \in \mathbb{R}^{d+1}$

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

$$h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

This Chapter

$\mathbf{x} \in \mathbb{R}^d; \ b \in \mathbb{R}, \ \mathbf{w} \in \mathbb{R}^d$

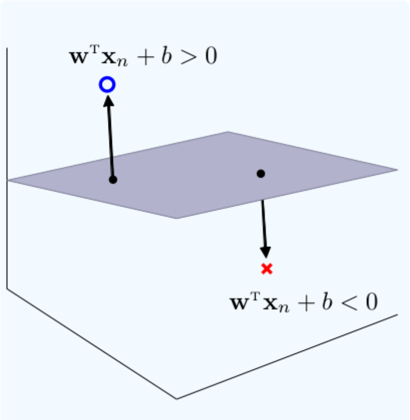
$$b = \text{bias}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

$$h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b)$$

$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}_n + b = 0$ defines a hyperplane H

Classification based on H



Output class: $h(\mathbf{x}) = \underbrace{\mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)}$

Relate parameters to margin

The classifier has parameters $(\mathbf{w}, b)$:

$$h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$$

We need to somehow relate $\mathbf{w}$ and b with the margin

Margin is the distance between H and the closest point among all points in D

$\Longrightarrow$ Let us examine $d(\mathbf{x}, H)$!

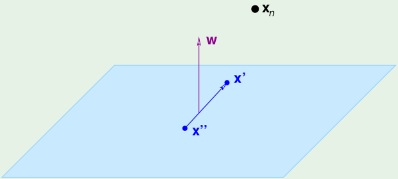
Recap: vector normal to the hyperplane

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

The vector $\mathbf{w}$ is $\perp$ to the plane in the $\mathcal{X}$ space:

Take $\mathbf{x}'$ and $\mathbf{x}''$ on the plane

??



Recap: vector normal to the hyperplane

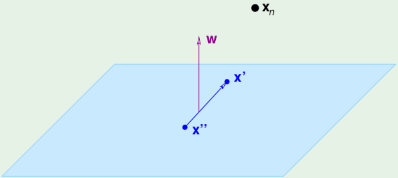
$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$

The vector $\mathbf{w}$ is $\perp$ to the plane in the $\mathcal{X}$ space:

Take $\mathbf{x}'$ and $\mathbf{x}''$ on the plane

$\mathbf{w}^{\intercal}\mathbf{x}' + b = 0$ and $\mathbf{w}^{\intercal}\mathbf{x}'' + b = 0$

$\implies \mathbf{w}^{\intercal}(\mathbf{x}' - \mathbf{x}'') = 0$



Recap: distance between point and hyperplane

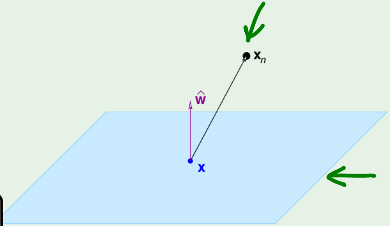
$d(\mathbf{x}_n, H) = ?$

Distance between $\mathbf{x}_n$ and the plane:

Take any point $\mathbf{x}$ on the plane

Projection of $\mathbf{x}_n - \mathbf{x}$ on $\mathbf{w}$

??



Recap: distance between point and hyperplane

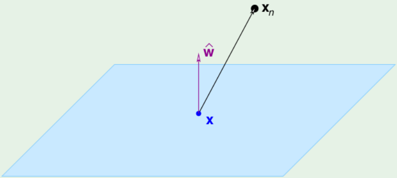
$d(\mathbf{x}_n, H) = ?$

Distance between $\mathbf{x}_n$ and the plane:

Take any point $\mathbf{x}$ on the plane

Projection of $\mathbf{x}_n - \mathbf{x}$ on $\mathbf{w}$

$$\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \implies \text{distance} = \left|\hat{\mathbf{w}}^\intercal(\mathbf{x}_n - \mathbf{x})\right|$$



The points in D may be in one or in the other side of H

Thus, distance is given by the absolute value $\left| \hat{\mathbf{w}}^T (\mathbf{x}_n - \mathbf{x}) \right|$

The need to treat the two cases (if-else situation) is not convenient

Remember logistic regression? There we used a trick o avoid if-else:

$$\boxed{P(y|\mathbf{x}) = \theta(y\,\mathbf{w}^T\mathbf{x})}$$

$$\boxed{P(y|\mathbf{x}) = P(y=1|\mathbf{x})^y\,[1 - P(y=1|\mathbf{x})]^{1-y}}$$

$p(y|x)$

$w^T x$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})|$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

$$\mathbf{w}^T(\mathbf{x_n} - \mathbf{x}) = \mathbf{w}^T\mathbf{x_n} - \mathbf{w}^T\mathbf{x}$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

$$\mathbf{w}^T(\mathbf{x_n} - \mathbf{x}) = \mathbf{w}^T\mathbf{x_n} - \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{x_n} + b - (\mathbf{w}^T\mathbf{x} + b)$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

$$\mathbf{w}^T(\mathbf{x_n} - \mathbf{x}) = \mathbf{w}^T\mathbf{x_n} - \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{x_n} + b - (\mathbf{w}^T\mathbf{x} + b)$$
$$= \mathbf{w}^T\mathbf{x_n} + b - 0 = \mathbf{w}^T\mathbf{x_n} + b$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\underbrace{\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})}|$$

$$\mathbf{w}^T(\mathbf{x_n} - \mathbf{x}) = \mathbf{w}^T\mathbf{x_n} - \mathbf{w}^T\mathbf{x} = \underbrace{\mathbf{w}^T\mathbf{x_n} + b} - \underbrace{(\mathbf{w}^T\mathbf{x} + b)}_{0}$$

$$= \mathbf{w}^T\mathbf{x_n} + b - 0 = \mathbf{w}^T\mathbf{x_n} + b$$

Why $\mathbf{w}^T\mathbf{x} + b = 0$ **?**

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

$$\mathbf{w}^T(\mathbf{x_n} - \mathbf{x}) = \mathbf{w}^T\mathbf{x_n} - \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{x_n} + b - (\mathbf{w}^T\mathbf{x} + b)$$

$$= \mathbf{w}^T\mathbf{x_n} + b - 0 = \mathbf{w}^T\mathbf{x_n} + b$$

$$dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x_n} + b| = \frac{1}{||\mathbf{w}||}y_n(\mathbf{w}^T\mathbf{x_n} + b)$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

$\mathbf{w}^T$

$= \mathbf{w}$

Why I can do $|\mathbf{w}^T\mathbf{x_n} + b| = y_n(\mathbf{w}^T\mathbf{x}_n + b)$ **?**

$$dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x_n} + b| = \frac{1}{||\mathbf{w}||}y_n(\mathbf{w}^T\mathbf{x_n} + b)$$

Rewriting the distance between point and hyperplane

$$dist(\mathbf{x}_n, H) = |\hat{\mathbf{w}}^T(\mathbf{x_n} - \mathbf{x})| = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T(\mathbf{x_n} - \mathbf{x})|$$

$$\mathbf{w}^T(\mathbf{x_n} - \mathbf{x}) = \mathbf{w}^T\mathbf{x_n} - \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{x_n} + b - (\mathbf{w}^T\mathbf{x} + b)$$
$$= \mathbf{w}^T\mathbf{x_n} + b - 0 = \mathbf{w}^T\mathbf{x_n} + b$$

$$dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x_n} + b| = \frac{1}{||\mathbf{w}||}\underbrace{y_n(\mathbf{w}^T\mathbf{x_n} + b)}$$

(because if $\mathbf{x_n}$ is at the correct side $\implies \underbrace{y_n(\mathbf{w}^T\mathbf{x_n})} > 0$)

Choosing a convenient hyperplane representation (weights)

Distance as seen before:

$$dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x_n} + b| = \frac{1}{||\mathbf{w}||}y_n(\mathbf{w}^T\mathbf{x_n} + b)$$

If I manage to make $\boxed{|\mathbf{w}^T\mathbf{x}_n + b| = 1}$, then I will have

$$\boxed{dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||}}$$

We can always rescale $(\mathbf{w}, b)$ so as to have $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$.

Let us do that with respect to the closest point to the hyperplane:

$$\rho = \min_{n=1,\ldots,N} y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b),$$

If we divide $(\mathbf{w}, b)$ by ρ, the hyperplane does not change:

$$\min_{n=1,\ldots,N} y_n\left(\frac{\mathbf{w}}{\rho}^{\mathrm{T}}\mathbf{x}_n + \frac{b}{\rho}\right) = \frac{1}{\rho}\boxed{\min_{n=1,\ldots,N} y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)} = \frac{\rho}{\rho} = 1.$$

Homework

Exercise 8.2

Consider the data below and a 'hyperplane' $(b, \mathbf{w})$ that separates the data.

$$\mathrm{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} 1.2 \\ -3.2 \end{bmatrix} \quad b = -0.5$$

(a) Compute $\rho = \min_{n=1,\ldots,N} y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)$.

(b) Compute the weights $\frac{1}{\rho}(b, \mathbf{w})$ and show that they satisfy (8.2).

(c) Plot both hyperplanes to show that they are the *same* separator.

(8.2) $$\min_{n=1,\ldots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$

Wrapping up

Let D be a linearly separable set of points, $\mathbf{x}_n \in D$, and a separating hyperplane H characterized by $(\mathbf{w}, b)$. Then

$$dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||} y_n(\mathbf{w}^T \mathbf{x_n} + b)$$

We can always choose $(\mathbf{w}, b)$ such that the closest point $\mathbf{x}_n$ to H satisfies

$$y_n(\mathbf{w}^T \mathbf{x_n} + b) = 1$$

In such case

$$dist(\mathbf{x}_n, H) = \frac{1}{||\mathbf{w}||}$$

The problem we want to solve

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{maximize}} && \frac{1}{||\mathbf{w}||} \\ &\text{subject to} && \min_{i=1,\ldots,N} y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 \end{aligned}$$

The problem we want to solve

$$\underset{\mathbf{w},b}{\text{maximize}} \quad \frac{1}{||\mathbf{w}||}$$
$$\text{subject to} \quad \min_{i=1,\ldots,N} y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$$

- The constraint $\min_{i=1,\ldots,N} y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$ implies $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ which has the effect of forcing all examples to be classified correctly
- The equality $\min_{i=1,\ldots,N} \boxed{y_i(\mathbf{w}^T\mathbf{x}_i + b)} = 1$ implies that the distance of the closest point to the hyperplane is $\frac{1}{||\mathbf{w}||}$ (a nice objective function!)

The problem we want to solve

$$
\begin{aligned}
&\underset{\mathbf{w},b}{\text{maximize}} && \frac{1}{||\mathbf{w}||}\\
&\text{subject to} && \min_{n=1,\ldots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1
\end{aligned}
$$

Equivalent formulation

Quadratic function

$$
\begin{aligned}
&\underset{\mathbf{w},b}{\text{minimize}} && \frac{1}{2}\mathbf{w}^T\mathbf{w}\\
&\text{subject to} && \min_{n=1,\ldots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1
\end{aligned}
$$

Relaxed formulation

Original minimization formulation:

$$
\begin{array}{ll}
\underset{\mathbf{w},b}{\text{minimize}} & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\
\text{subject to} & \min_{n=1,\ldots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1
\end{array}
$$

Equivalent relaxed formulation:

$$
\begin{array}{ll}
\underset{\mathbf{w},b}{\text{minimize}} & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\
\text{subject to} & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \; n = 1, \ldots, N
\end{array}
$$

The equivalence can be proved by contradiction (see Chapter on SVM, page 7)

A toy example

N=4

Constraints: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$?

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

A toy example

Constraints: $\underset{\uparrow}{y_i}(\underbrace{\mathbf{w}^T\mathbf{x}_i} + b) \geq 1$?

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

$$\begin{aligned} -b &\geq 1 \quad (1) \\ -(2w_1 + 2w_2 + b) &\geq 1 \quad (2) \\ 2w_1 + b &\geq 1 \quad (3) \\ 3w_1 + b &\geq 1 \quad (4) \end{aligned}$$

$\frac{1}{2}\,w^Tw$

Solving it by hand

$$
\begin{aligned}
-b &\geq 1 \quad (1)\\
-(2w_1 + 2w_2 + b) &\geq 1 \quad (2)\\
2w_1 + b &\geq 1 \quad (3)\\
3w_1 + b &\geq 1 \quad (4)
\end{aligned}
$$

Solving it by hand

$$
\begin{aligned}
-b &\geq 1 \quad (1)\\
-(2w_1 + 2w_2 + b) &\geq 1 \quad (2)\\
2w_1 + b &\geq 1 \quad (3)\\
3w_1 + b &\geq 1 \quad (4)
\end{aligned}
$$

- From (3) and (1)
  $2w_1 + b \geq 1 \rightsquigarrow 2w_1 \geq 1 - b \rightsquigarrow w_1 \geq \frac{1}{2}(1-b)$ && $b \leq -1$
  $\Longrightarrow w_1 \geq 1$

$$
\begin{aligned}
-b &\geq 1 \quad (1)\\
-(2w_1 + 2w_2 + b) &\geq 1 \quad (2)\\
2w_1 + b &\geq 1 \quad (3)\\
3w_1 + b &\geq 1 \quad (4)
\end{aligned}
$$

- From (3) and (1)
  $2w_1 + b \geq 1 \rightsquigarrow 2w_1 \geq 1 - b \rightsquigarrow w_1 \geq \frac{1}{2}(1 - b)$ && $b \leq -1$
  $\Longrightarrow w_1 \geq 1$
- From (2) and (3):
  $-(2w_1 + 2w_2 + b) \geq 1 \rightsquigarrow -2w_1 - 2w_2 - b \geq 1 \rightsquigarrow$
  $2w_2 \leq -2w_1 - b - 1$ && $2w_1 + b \geq 1 \Longrightarrow w_2 \leq -1$

Solving it by hand

$$
\begin{aligned}
-b &\geq 1 \quad (1)\\
-(2w_1 + 2w_2 + b) &\geq 1 \quad (2)\\
2w_1 + b &\geq 1 \quad (3)\\
3w_1 + b &\geq 1 \quad (4)
\end{aligned}
$$

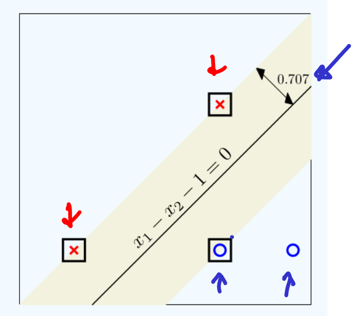
- From (3) and (1)
  $2w_1 + b \geq 1 \rightsquigarrow 2w_1 \geq 1 - b \rightsquigarrow w_1 \geq \frac{1}{2}(1-b)$ && $b \leq -1$
  $\Longrightarrow w_1 \geq 1$

- From (2) and (3):
  $-(2w_1 + 2w_2 + b) \geq 1 \rightsquigarrow -2w_1 - 2w_2 - b \geq 1 \rightsquigarrow$
  $2w_2 \leq -2w_1 - b - 1$ && $2w_1 + b \geq 1 \Longrightarrow w_2 \leq -1$

Thus, $\boxed{\frac{1}{2}\mathbf{w}^T\mathbf{w}} = \frac{1}{2}(w_1^2 + w_2^2) \geq 1$ and the minimum is at $\underbrace{\mathbf{w} = (1, -1)}$;
$\underbrace{(b = -1, w_1 = 1, w_2 = -1)}$ satisfies the 4 constraints

The separating hyperplane H with maximum margin is given by $x_1 - x_2 - 1 = 0$.

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

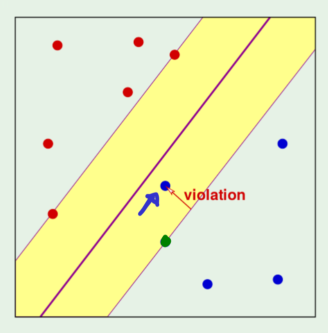


The margin is $\frac{1}{||\mathbf{w}||} = \frac{1}{\sqrt{2}} \approx 0.707$

Summary (linearly separable case)

- The goal is to find a hyperplane that maximizes the margin
- We examined the formulation of the hard margin SVM
- It can be written as a QP otimization (quadratic objective function with linear inequality constraints)
- We solved a toy example by hand
- We still do not know how to solve QP problems

This case is dealt by considering a **soft margin** formulation as opposed to the (previuous) **hard margin** formulation:



Soft margin: $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n$

(**Hard margin**: $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$)

Optimization problem

$$
\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{n=1}^{N}\xi_n \\
\text{subject to} \quad & y_n\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b\right) \geq 1 - \xi_n \text{ for } n = 1, 2, \ldots, N; \\
& \xi_n \geq 0 \text{ for } n = 1, 2, \ldots, N.
\end{aligned}
$$

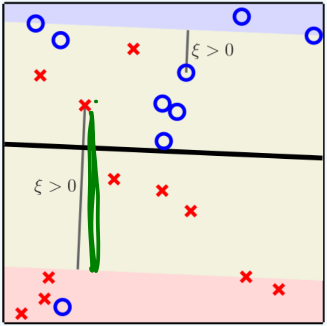
$C >= 0$ is an user-specified parameter; the larger it is, the smaller the allowed margin violation

---

Compare to the hard-margin formulation:

$$
\begin{aligned}
\underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\
\text{subject to} \quad & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, n = 1, \ldots, N
\end{aligned}
$$

Intuition on constant C



(a) $C = 1$ (b) $C = 500$

$$\underset{\mathbf{w},b,\xi}{\text{minimize}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

How to solve QP optimization problems ?

Both cases, hard and soft margin SVM, can be formulated as a QP optimization problem

Primal formulation: Standard QP optimization

Dual formulation: based on Lagrange formulation, dual QP

Standard QP optimization

Standard form of QP problems

M inequality constraints and Q positive semi-definite

$$\begin{aligned} &\underset{\mathbf{u}}{\text{minimize}} && \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} \\ &\text{subject to:} && \mathbf{a}_m^T\mathbf{u} \geq c_m \quad (m = 1, \ldots, M) \end{aligned}$$

In matrix form

$$\begin{aligned} &\underset{\mathbf{u}}{\text{minimize}} && \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u} \\ &\text{subject to:} && A\mathbf{u} \geq \mathbf{c} \end{aligned}$$

QP solvers can be used to compute the optimal solution $\mathbf{u}^*$:

$$\mathbf{u}^* \leftarrow \mathrm{QP}(Q, \mathbf{p}, A, \mathbf{c})$$

QP problem formulation

$$\underset{\mathbf{u}}{\text{minimize}} \quad \frac{1}{2}\mathbf{u}^T Q\mathbf{u} + \mathbf{p}^T\mathbf{u}$$

$$\text{subject to:} \quad \mathbf{a}_m^T\mathbf{u} \geq c_m$$

$$i = m, \ldots, M$$

QP of hard-margin SVM

$$\underset{\mathbf{w},b}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{subject to:} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

$$i = 1, \ldots, N$$

Denoting $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$, we have

$$\mathbf{w}^{\mathrm{T}}\mathbf{w} = \begin{bmatrix} b & \mathbf{w}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}_d^{\mathrm{T}} \\ \mathbf{0}_d & \mathrm{I}_d \end{bmatrix} \begin{bmatrix} b \\ \mathbf{w}^{\mathrm{T}} \end{bmatrix} = \mathbf{u}^{\mathrm{T}} \begin{bmatrix} 0 & \mathbf{0}_d^{\mathrm{T}} \\ \mathbf{0}_d & \mathrm{I}_d \end{bmatrix} \mathbf{u},$$

$$\mathbf{a}_n^{\mathrm{T}} = y_n \begin{bmatrix} 1 & \mathbf{x}_n^{\mathrm{T}} \end{bmatrix} \text{ and } c_n = 1$$

Linear Hard-Margin SVM with QP

1: Let $\mathbf{p} = \mathbf{0}_{d+1}$ ($(d+1)$-dimensional zero vector) and $\mathbf{c} = \mathbf{1}_N$ (N-dimensional vector of ones). Construct matrices Q and A, where

$$\underset{(d+1)\times(d+1)}{\mathrm{Q}} = \begin{bmatrix} 0 & \overset{1\times d}{\mathbf{0}_d^{\mathrm{T}}} \\ \underset{d\times 1}{\mathbf{0}_d} & \mathrm{I}_d \end{bmatrix}, \qquad \mathrm{A} = \underbrace{\begin{bmatrix} y_1 & -y_1\mathbf{x}_1^{\mathrm{T}}- \\ \vdots & \vdots \\ y_N & -y_N\mathbf{x}_N^{\mathrm{T}}- \end{bmatrix}}_{\text{signed data matrix}}.$$

2: Calculate $\begin{bmatrix} b^* \\ \mathbf{w}^* \end{bmatrix} = \mathbf{u}^* \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$.

3: Return the hypothesis $\boxed{g(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{*\mathrm{T}}\mathbf{x} + b^*)}$.

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$v = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}$$

b

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$v^T Q v = \begin{bmatrix} b & w_1 & w_2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & w_1 & w_2 \end{bmatrix} \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}$$

$$= w^T w$$