FSAN/ELEG815: Statistical Learning

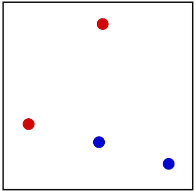
Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

Support Vector Machines

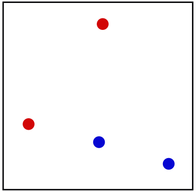
Support Vector Machines - Better linear separation

- Linearly separable data.



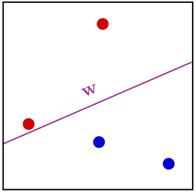
Support Vector Machines - Better linear separation

- Linearly separable data.
- Different separating lines.



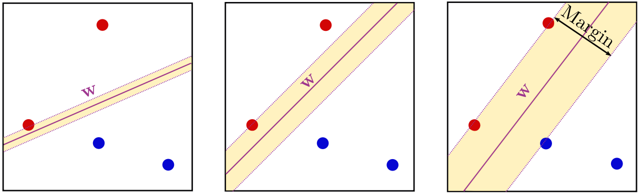
Support Vector Machines - Better linear separation

- Linearly separable data.
- Different separating lines.



Support Vector Machines - Better linear separation

- Linearly separable data.
- Different separating lines.
- Which one is best?
- Intuitively, bigger margin is better.



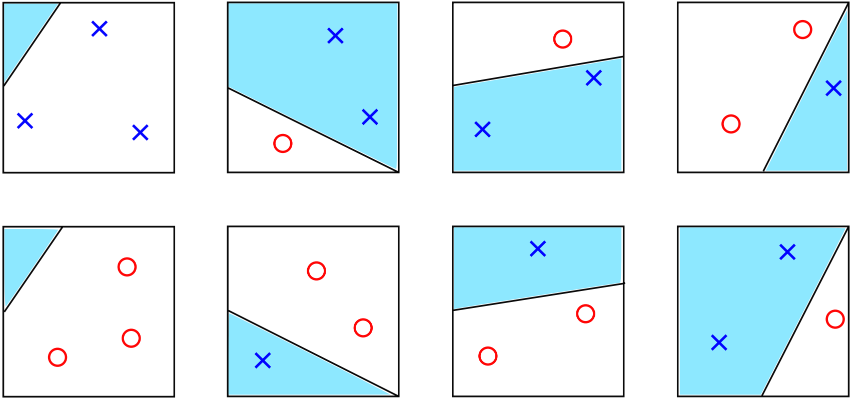
Two questions:

1. Why is bigger margin better?
2. Which $\mathbf{w}$ maximizes the margin?

,

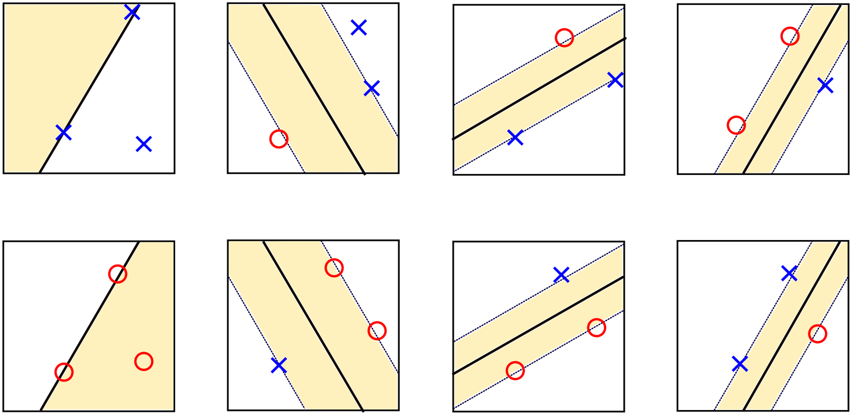
Support Vector Machines - Growth Function

All Possible Dichotomies with a line.



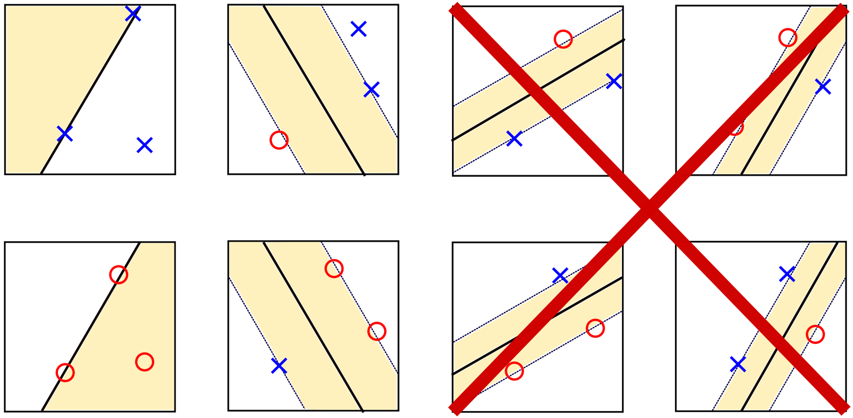
Bad news!

Support Vector Machines - Growth Function



Let's consider a classifier that requires a minimum margin.

Support Vector Machines - Growth Function



Let's consider a classifier that requires a minimum margin.

Fat margins imply fewer dichotomies $\Longrightarrow$ smaller growth function

Support Vector Machines - Finding $\mathbf{w}$ with large margin

Let $\mathbf{x}_n$ be the nearest data point to the line/plane (given by $\mathbf{w}^\top \mathbf{x} = 0$)

How far is it?

Two preliminary techniques:

1. **Normalize** $\mathbf{w}$: For any point:

$$|\mathbf{w}^\top \mathbf{x}_n| > 0.$$

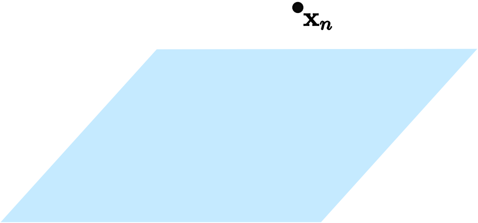
   Does scalar multiplication change the plane? NO! Pick one:

$$|\mathbf{w}^\top \mathbf{x}_n| = 1.$$

2. **Pull out** w_0:
   $\mathbf{w} = (w_1, \ldots w_d)$ apart from $w_0 = b$.
   The plane is now $\boxed{\mathbf{w}\mathbf{x} + b = 0}$ (no x_0).

Support Vector Machines - Computing the distance

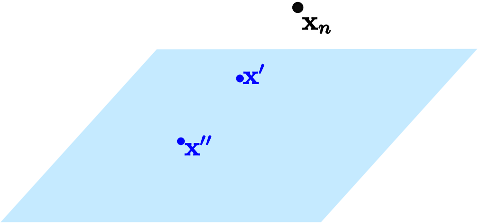
The distance between $\mathbf{x}_n$ and the plane $\mathbf{w}^\top\mathbf{x}+b=0$, where $|\mathbf{w}^\top\mathbf{x}_n+b|=1$.



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

Support Vector Machines - Computing the distance

The distance between $\mathbf{x}_n$ and the plane $\mathbf{w}^\top\mathbf{x}+b=0$, where $|\mathbf{w}^\top\mathbf{x}_n+b|=1$.



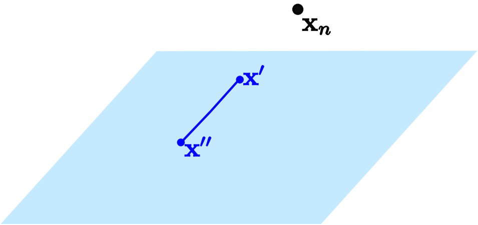
The vector $\mathbf{w}$ is $\perp$ to the plane in the $\mathcal{X}$ space:

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

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

Support Vector Machines - Computing the distance

The distance between $\mathbf{x}_n$ and the plane $\mathbf{w}^\top\mathbf{x}+b=0$, where $|\mathbf{w}^\top\mathbf{x}_n+b|=1$.



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

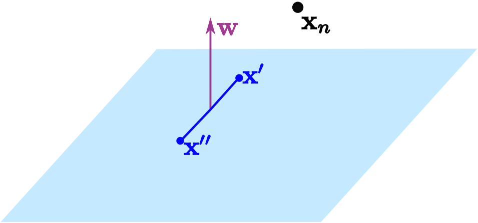
Take $\mathbf{x}'$ and $\mathbf{x}''$ on the plane.

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

$$\Longrightarrow \mathbf{w}^\top(\mathbf{x}'-\mathbf{x}'')=0.$$

Support Vector Machines - Computing the distance

The distance between $\mathbf{x}_n$ and the plane $\mathbf{w}^\top\mathbf{x}+b=0$, where $|\mathbf{w}^\top\mathbf{x}_n+b|=1$.



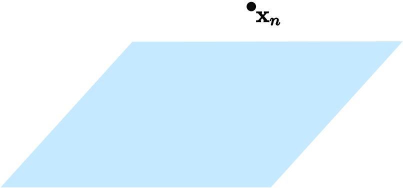
The vector $\mathbf{w}$ is $\perp$ to the plane in the $\mathcal{X}$ space:

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

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

$$\Longrightarrow \mathbf{w}^\top(\mathbf{x}'-\mathbf{x}'')=0.$$

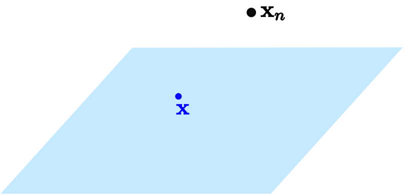
... and the distance is...



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

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

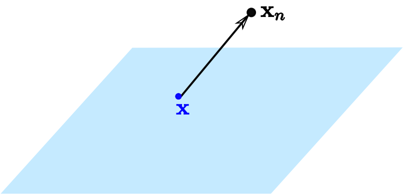
... and the distance is...



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

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

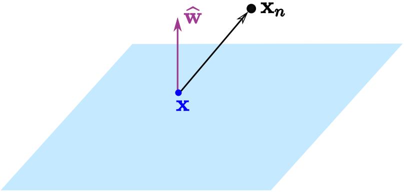
… and the distance is…



Distance between $\mathbf{x}_n$ and the plane:
Take any point $\mathbf{x}$ on the plane.

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

… and the distance is…



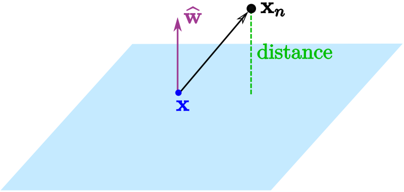
Distance between $\mathbf{x}_n$ and the plane:

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

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

$$\widehat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \Rightarrow \text{distance} = |\widehat{\mathbf{w}}^\top(\mathbf{x}_n - \mathbf{x})|.$$

... and the distance is...



Distance between $\mathbf{x}_n$ and the plane:
Take any point $\mathbf{x}$ on the plane.

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

$$\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \Rightarrow \text{distance} = |\hat{\mathbf{w}}^\top(\mathbf{x}_n - \mathbf{x})|.$$

$$\text{distance} = \frac{1}{\|\mathbf{w}\|}|\mathbf{w}^\top\mathbf{x}_n - \mathbf{w}^\top\mathbf{x}| \Longrightarrow \frac{1}{\|\mathbf{w}\|}|\underbrace{\mathbf{w}^\top\mathbf{x}_n + b}_{=1.} - \underbrace{\mathbf{w}^\top\mathbf{x} - b}_{\substack{=0. \\ \text{Point on the plain}}}| = \frac{1}{\|\mathbf{w}\|}.$$

Support Vector Machines - The optimization problem

Maximize the margin:

$$\text{maximize}_{\mathbf{w},b} \quad \frac{1}{\|\mathbf{w}\|} \qquad \Longrightarrow \text{Hard to solve}$$

$$\text{subject to} \quad \min_{n=1,2,\ldots,N} |\mathbf{w}^\top \mathbf{x}_n + b| = 1.$$

We need to get rid of the $\min$.

Support Vector Machines - The optimization problem

Maximize the margin:

$$\text{maximize}_{\mathbf{w},b} \quad \frac{1}{\|\mathbf{w}\|} \qquad \Longrightarrow \text{Hard to solve}$$

$$\text{subject to} \quad \min_{n=1,2,\ldots,N} |\mathbf{w}^\top \mathbf{x}_n + b| = 1.$$

We need to get rid of the $\min$.

Notice: $|\mathbf{w}^\top \mathbf{x}_n + b| = y_n(\mathbf{w}^\top \mathbf{x}_n + b)$.

$\mathbf{x}_n$ is classified correctly.

Support Vector Machines - The optimization problem

Maximize the margin:

$$\text{maximize}_{\mathbf{w},b} \quad \frac{1}{\|\mathbf{w}\|} \qquad \Longrightarrow \text{Hard to solve}$$

$$\text{subject to} \quad \min_{n=1,2,\dots,N} |\mathbf{w}^\top \mathbf{x}_n + b| = 1.$$

We need to get rid of the $\min$.

Notice: $|\mathbf{w}^\top \mathbf{x}_n + b| = y_n(\mathbf{w}^\top \mathbf{x}_n + b)$.

$\mathbf{x}_n$ is classified correctly.

$$\text{minimize}_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w} \qquad \Longrightarrow \text{Equivalent problem}$$

$$\text{subject to} \quad y_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1 \ \text{ for } n = 1, 2, \dots, N;.$$

Support Vector Machines - Constrained optimization

$$\text{minimize}_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

$$\text{subject to} \quad y_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1 \quad \text{for } n = 1, 2, \ldots, N,$$

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

Lagrange? inequality instead of equality constraints

$\Longrightarrow$ KKT: Lagrange under inequality constraints

Support Vector Machines - We saw this before

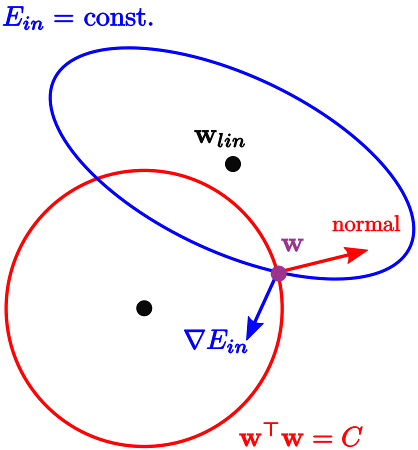
Remember regularization?

$$\text{minimize} \quad E_{in}(\mathbf{w}) = \frac{1}{N}(\mathbf{Z}\mathbf{w}-\mathbf{y})^\top(\mathbf{Z}\mathbf{w}-\mathbf{y})$$

$$\text{subject to} \quad \mathbf{w}^\top\mathbf{w} \le C.$$

Condition for the solution:

∇E_{in} relates to constraint.
∇E_{in} parallel to $\mathbf{w}_{reg}$ but in the opposite direction.

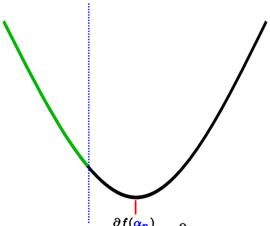


	Optimize	**Constrain**
Regularization	E_{in}	$\mathbf{w}^\top\mathbf{w}$
SVM	$\mathbf{w}^\top\mathbf{w}$	E_{in}

Support Vector Machines - Lagrange formulation

$$\text{minimize} \quad \mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^\top\mathbf{w} - \sum_{n=1}^{N} \alpha_n \underbrace{\left(y_n(\mathbf{w}^\top\mathbf{x}_n + b) - 1\right)}_{\substack{\text{constrain} \\ y_n(\mathbf{w}^\top\mathbf{x}_n + b) - 1 \geq 0}}$$

w.r.t to $\mathbf{w}$ and b and maximize w.r.t each $\alpha_n \geq 0$.



$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = \mathbf{0}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$$

Support Vector Machines - Lagrange formulation

Substituting

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \quad \text{and} \quad \sum_{n=1}^{N} \alpha_n y_n = 0$$

In the Lagrangian:

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_{n=1}^{N} \alpha_n (y_n(\mathbf{w}^\top \mathbf{x}_n + \underbrace{b}_{\sum_{n=1}^{N} \alpha_n (y_n) b = 0}) - 1),$$

Support Vector Machines - Lagrange formulation

Substituting

$$\mathbf{w} = \textstyle\sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \quad \text{and} \quad \textstyle\sum_{n=1}^{N} \alpha_n y_n = 0$$

In the Lagrangian:

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \tfrac{1}{2}\mathbf{w}^\top \mathbf{w} - \textstyle\sum_{n=1}^{N} \alpha_n (y_n(\mathbf{w}^\top \mathbf{x}_n + \underbrace{b}_{\sum_{n=1}^{N} \alpha_n (y_n) b = 0}) - 1),$$

we get:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \alpha) &= \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_{n=1}^{N} \alpha_n (y_n \mathbf{w}^\top \mathbf{x}_n - 1) \\ &= \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^\top \mathbf{x}_n + \sum_{n=1}^{N} \alpha_n \end{aligned}$$

Support Vector Machines - Lagrange formulation

Substituting

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

In:

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^\top \mathbf{x}_n + \sum_{n=1}^{N} \alpha_n,$$

Support Vector Machines - Lagrange formulation

Substituting

$$\mathbf{w} = \textstyle\sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

In: $$\mathcal{L}(\mathbf{w}, b, \alpha) = \tfrac{1}{2}\mathbf{w}^\top \mathbf{w} - \textstyle\sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^\top \mathbf{x}_n + \textstyle\sum_{n=1}^{N} \alpha_n,$$

we get:

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^\top \mathbf{x}_m.$$

Now maximize w.r.t α *subject to* $\alpha_n \geq 0$ for $n = 1, \ldots, N$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$.

Support Vector Machines - The solution

Notice: $\max \ \mathcal{L} \ = \ \min \ -\mathcal{L}.$

Quadratic programming:

$$\min_{\alpha} \frac{1}{2}\alpha^\top \underbrace{\begin{bmatrix} y_1y_1\mathbf{x}_1^\top\mathbf{x}_1 & y_1y_2\mathbf{x}_1^\top\mathbf{x}_2 & \dots & y_1y_N\mathbf{x}_1^\top\mathbf{x}_N \\ y_2y_1\mathbf{x}_2^\top\mathbf{x}_1 & y_2y_2\mathbf{x}_2^\top\mathbf{x}_2 & \dots & y_2y_N\mathbf{x}_2^\top\mathbf{x}_N \\ y_Ny_1\mathbf{x}_N^\top\mathbf{x}_1 & y_Ny_2\mathbf{x}_N^\top\mathbf{x}_2 & \dots & y_Ny_N\mathbf{x}_N^\top\mathbf{x}_N \end{bmatrix}}_{\text{quadratic coefficients}} \alpha + \underbrace{(-\mathbf{1}^\top)}_{\text{linear}}\alpha$$

$$\text{subject to} \quad \underbrace{\mathbf{y}^\top\alpha = 0}_{\text{linear constraint}} \ ,$$

$$\underbrace{\mathbf{0}}_{\text{lower bounds}} \leq \alpha \leq \underbrace{\infty}_{\text{upper bounds}} \quad .$$

Support Vector Machines - QP hands us α

Solution: $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_N$

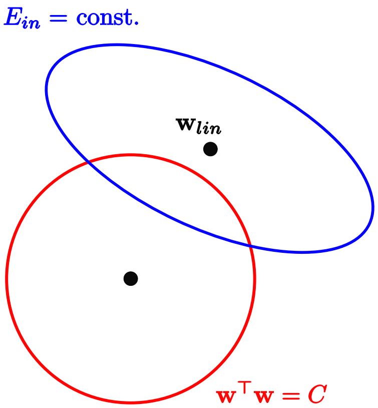
$\Longrightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n.$

Many α_n equal to zero. KKT condition:

For $n = 1, \ldots, N$

$$\alpha_n (y_n (\mathbf{w}^\top \mathbf{x}_n + b) - 1) = 0.$$

We saw this before!



Support Vector Machines - QP hands us α

Solution: $\alpha = \alpha_1, \alpha_2, ..., \alpha_N$

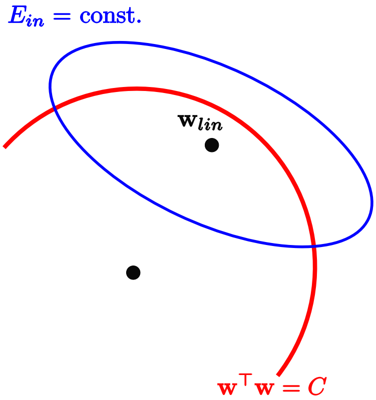
$\Longrightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n.$

Many α_n equal to zero. KKT condition:

For $n = 1, ..., N$

$$\alpha_n(y_n(\mathbf{w}^\top \mathbf{x}_n + b) - 1) = 0.$$

We saw this before!



Support Vector Machines - QP hands us α

Solution: $\alpha = \alpha_1, \alpha_2, ..., \alpha_N$

$\Longrightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n.$

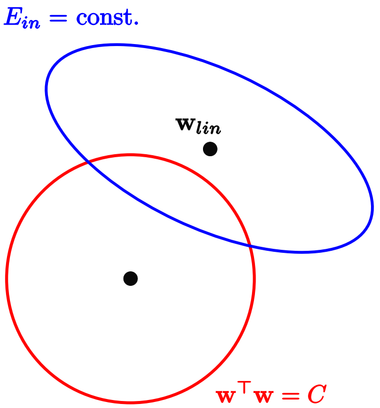
Many α_n equal to zero. KKT condition:

For $n = 1, ..., N$

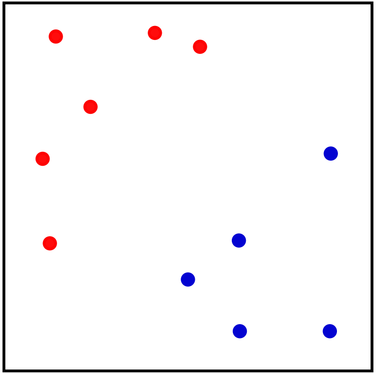
$$\alpha_n(y_n(\mathbf{w}^\top \mathbf{x}_n + b) - 1) = 0.$$

We saw this before!

$\alpha_n > 0 \Longrightarrow \mathbf{x}_n$ is a **support vector.**

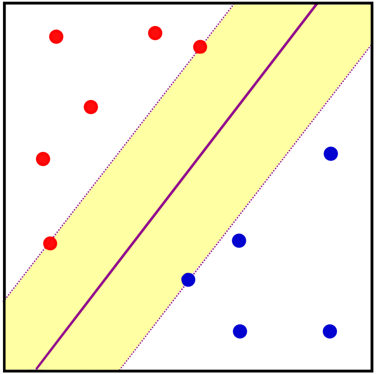


Support Vector Machines - Support vectors



Support Vector Machines - Support vectors

Closest $\mathbf{x}_n$'s to the plane.
Support vectors $\Longrightarrow$ achieve the margin.



Support Vector Machines - Support vectors

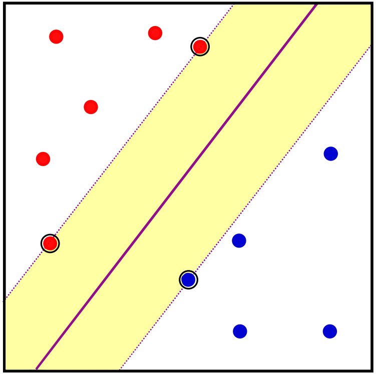
Closest $\mathbf{x}_n$'s to the plane.
Support vectors $\Longrightarrow$ achieve the margin.

$$\Longrightarrow y_n(\mathbf{w}^\top \mathbf{x}_n + b)) = 1.$$

$$\mathbf{w} = \sum_{\mathbf{x}_n \text{is SV}} \alpha_n y_n \mathbf{x}_n.$$

Solve b using any support vector:

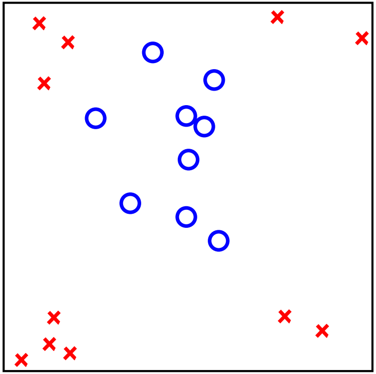
$$y_n(\mathbf{w}^\top \mathbf{x}_n + b)) = 1.$$



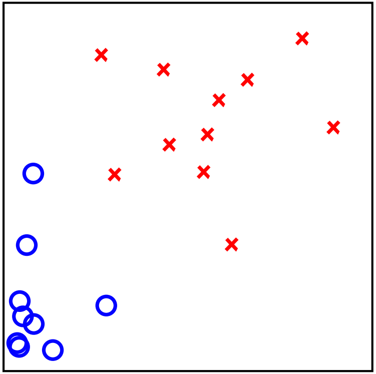
Support Vector Machines - Nonlinear transformation

$\mathbf{z}$ instead of $\mathbf{x}$

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^\top \mathbf{x}_m.$$



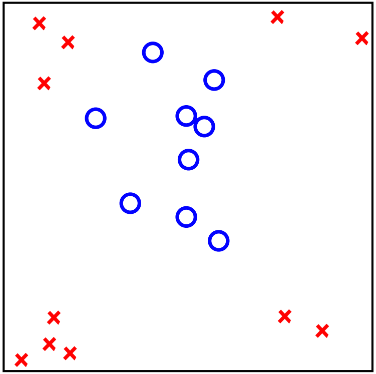
$\mathcal{X} \longrightarrow \mathcal{Z}$



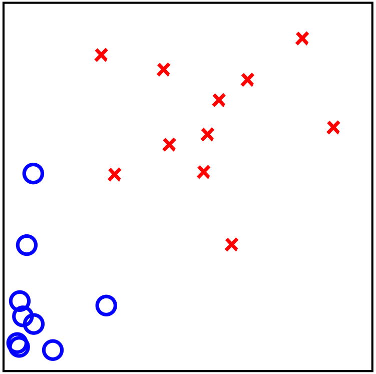
Support Vector Machines - Nonlinear transformation

$\mathbf{z}$ instead of $\mathbf{x}$

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{z}_n^\top \mathbf{z}_m.$$

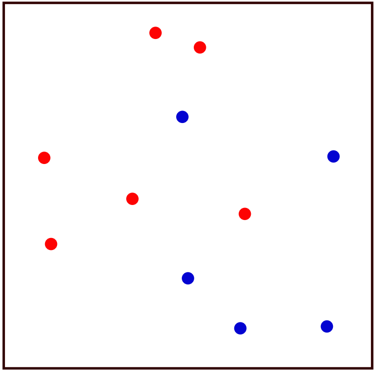


$$\mathcal{X} \longrightarrow \mathcal{Z}$$



Support Vector Machines - "Support vectors" in $\mathcal{X}$ space

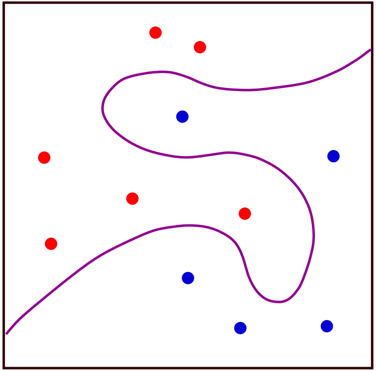
Support vectors live in the $\mathcal{Z}$ space. In the $\mathcal{X}$ space, "pre-images" of support vectors.



Support Vector Machines - "Support vectors" in $\mathcal{X}$ space

Support vectors live in the $\mathcal{Z}$ space. In the $\mathcal{X}$ space, "pre-images" of support vectors.

The margin is maintened in the $\mathcal{Z}$ space.



Support Vector Machines - "Support vectors" in $\mathcal{X}$ space

Support vectors live in the $\mathcal{Z}$ space. In the $\mathcal{X}$ space, "pre-images" of support vectors.

The margin is maintened in the $\mathcal{Z}$ space.

Generalization result

$$\mathbb{E}\left[E_{out}\right] \leq \frac{\mathbb{E}[\#\text{ of SV's}]}{N-1}$$