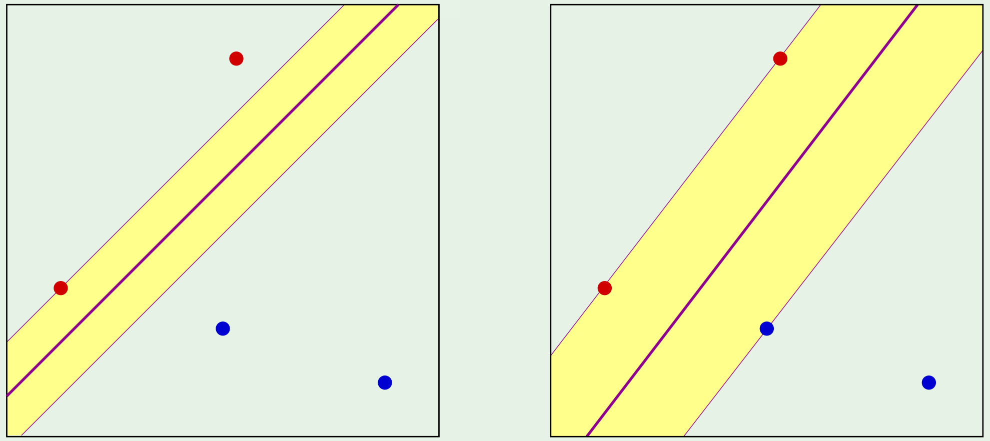
Review of Lecture 14

- The margin

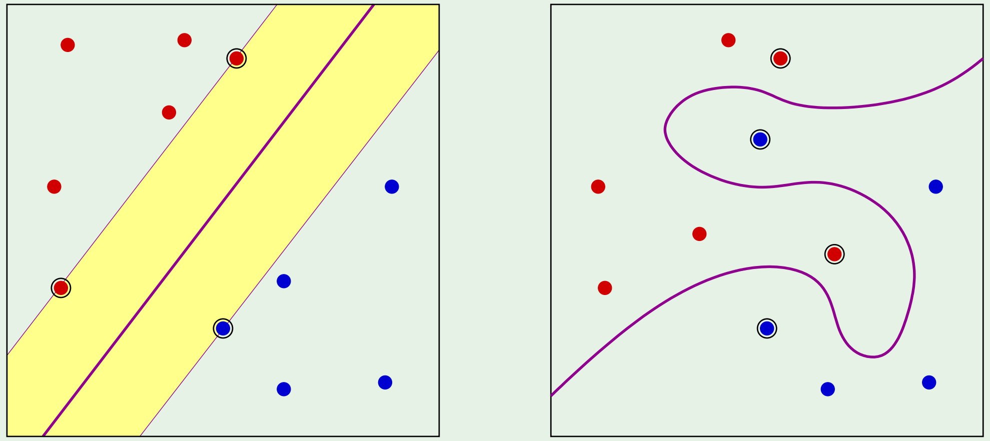


Maximizing the margin $\Longrightarrow$ dual problem:

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

quadratic programming

- Support vectors



$\mathbf{x}_n$ (or $\mathbf{z}_n$) with Lagrange $\alpha_n > 0$

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

(in-sample check of out-of-sample error)

- Nonlinear transform

Complex h, but simple $\mathcal{H}$ ☺

Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 15: Kernel Methods

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Tuesday, May 22, 2012

Outline

- The kernel trick

- Soft-margin SVM

What do we need from the $\mathcal{Z}$ space?

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

Constraints: $\alpha_n \geq 0$ for $n = 1, \cdots, N$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$

$$\boxed{g(\mathbf{x}) = \operatorname{sign}\left(\mathbf{w}^{\intercal} \mathbf{z} + b\right)}$$ need $\mathbf{z}_n^{\intercal} \mathbf{z}$

where $\mathbf{w} = \sum_{\mathbf{z}_n \text{ is SV}} \alpha_n y_n \mathbf{z}_n$

and b: $y_m \left(\mathbf{w}^{\intercal} \mathbf{z}_m + b\right) = 1$ need $\mathbf{z}_n^{\intercal} \mathbf{z}_m$

Generalized inner product

Given two points $\mathbf{x}$ and $\mathbf{x}' \in \mathcal{X}$, we need $\mathbf{z}^\intercal\mathbf{z}'$

Let $\mathbf{z}^\intercal\mathbf{z}' = K(\mathbf{x}, \mathbf{x}')$ (the kernel) "inner product" of $\mathbf{x}$ and $\mathbf{x}'$

Example: $\mathbf{x} = (x_1, x_2) \longrightarrow$ 2nd-order Φ

$$\mathbf{z} = \Phi(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$$

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{z}^\intercal\mathbf{z}' = 1 + x_1{x'}_1 + x_2{x'}_2 + x_1^2{x'}_1^2 + x_2^2{x'}_2^2 + x_1{x'}_1x_2{x'}_2$$

The trick

Can we compute $K(\mathbf{x}, \mathbf{x}')$ **without** transforming $\mathbf{x}$ and $\mathbf{x}'$?

Example: Consider $K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^\intercal \mathbf{x}')^2 = (1 + x_1 x'_1 + x_2 x'_2)^2$

$$= 1 + x_1^2 x'^2_1 + x_2^2 x'^2_2 + 2 x_1 x'_1 + 2 x_2 x'_2 + 2 x_1 x'_1 x_2 x'_2$$

This is an inner product!

$$(1 , x_1^2 , x_2^2 , \sqrt{2}\, x_1 , \sqrt{2}\, x_2 , \sqrt{2}\, x_1 x_2)$$

$$(1 , x'^2_1 , x'^2_2 , \sqrt{2} x'_1 , \sqrt{2} x'_2 , \sqrt{2} x'_1 x'_2)$$

The polynomial kernel

$\mathcal{X} = \mathbb{R}^d$ and $\Phi : \mathcal{X} \to \mathcal{Z}$ is polynomial of order Q

The "equivalent" kernel $K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^\intercal \mathbf{x}')^Q$

$$= (1 + x_1 x'_1 + x_2 x'_2 + \cdots + x_d x'_d)^Q$$

Compare for $d = 10$ and $Q = 100$

Can adjust scale: $K(\mathbf{x}, \mathbf{x}') = (a\mathbf{x}^\intercal \mathbf{x}' + b)^Q$

We only need $\mathcal{Z}$ to exist!

If $K(\mathbf{x}, \mathbf{x}')$ is an inner product in <u>some</u> space $\mathcal{Z}$, we are good.

Example: $$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|^2\right)$$

Infinite-dimensional $\mathcal{Z}$: take simple case

$$K(x, x') = \exp\left(-(x - x')^2\right)$$

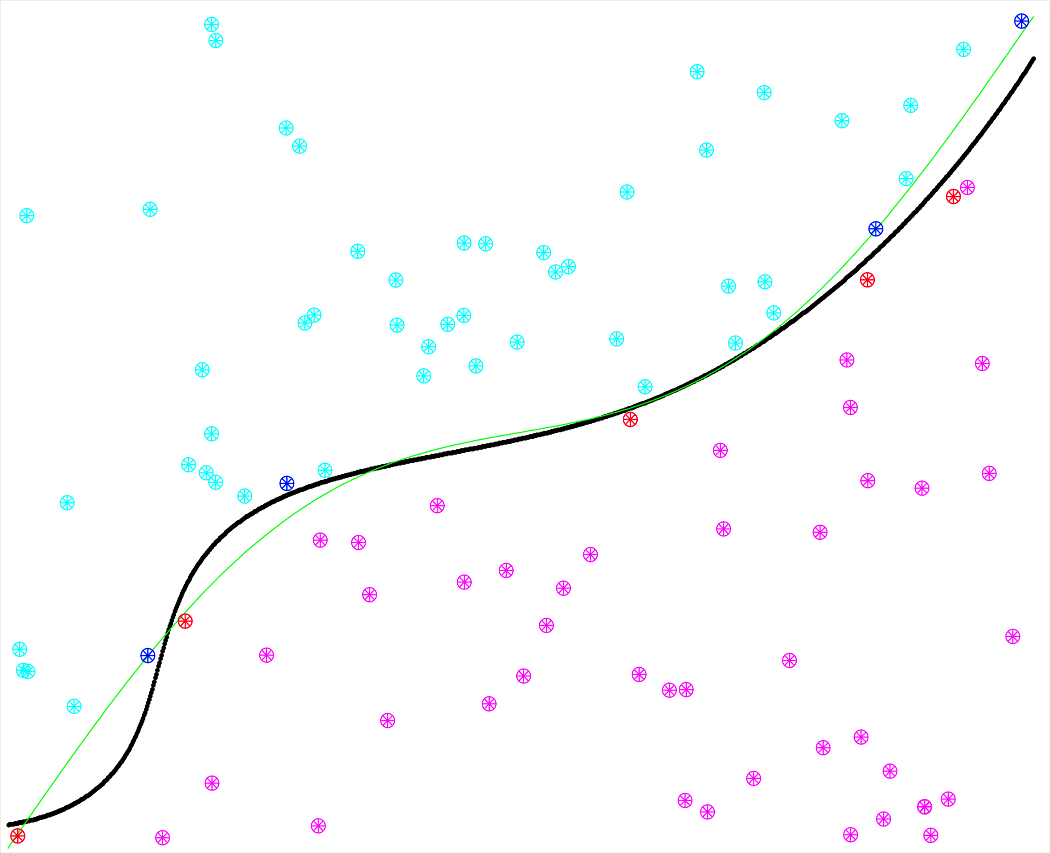
$$= \exp\left(-x^2\right) \exp\left(-x'^2\right) \underbrace{\sum_{k=0}^{\infty} \frac{2^k (x)^k (x')^k}{k!}}_{\exp(2xx')}$$

This kernel in action

Slightly non-separable case:

Transforming $\mathcal{X}$ into ∞-dimensional $\mathcal{Z}$

Overkill? Count the support vectors



Kernel formulation of SVM

Remember quadratic programming? The only difference now is:

$$\underbrace{\begin{bmatrix} y_1y_1\,K(\mathbf{x}_1,\mathbf{x}_1) & y_1y_2\,K(\mathbf{x}_1,\mathbf{x}_2) & \dots & y_1y_N K(\mathbf{x}_1,\mathbf{x}_N) \\ y_2y_1\,K(\mathbf{x}_2,\mathbf{x}_1) & y_2y_2\,K(\mathbf{x}_2,\mathbf{x}_2) & \dots & y_2y_N K(\mathbf{x}_2,\mathbf{x}_N) \\ \dots & \dots & \dots & \dots \\ y_Ny_1 K(\mathbf{x}_N,\mathbf{x}_1) & y_Ny_2 K(\mathbf{x}_N,\mathbf{x}_2) & \dots & y_Ny_N K(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix}}_{\text{quadratic coefficients}}$$

Everything else is the same.

The final hypothesis

Express $\quad g(\mathbf{x}) = \operatorname{sign}\left(\mathbf{w}^\intercal \mathbf{z} + b\right) \quad$ in terms of $K(-,-)$

$$\mathbf{w} = \sum_{\mathbf{z}_n \text{ is SV}} \alpha_n y_n \mathbf{z}_n \quad \Longrightarrow \quad g(\mathbf{x}) = \operatorname{sign}\left(\sum_{\alpha_n > 0} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b\right)$$

where $\; b = y_m - \sum_{\alpha_n > 0} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}_m)$

for any support vector $(\alpha_m > 0)$

How do we know that $\mathcal{Z}$ exists ...

... for a given $K(\mathbf{x}, \mathbf{x}')$? valid kernel

Three approaches:

1. By construction

2. Math properties (*Mercer's condition*)

3. Who cares? ☺

Design your own kernel

$K(\mathbf{x}, \mathbf{x}')$ is a valid kernel iff

1. It is symmetric and **2.** The matrix:

$$\begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \dots & K(\mathbf{x}_1, \mathbf{x}_N) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \dots & K(\mathbf{x}_2, \mathbf{x}_N) \\ \dots & \dots & \dots & \dots \\ K(\mathbf{x}_N, \mathbf{x}_1) & K(\mathbf{x}_N, \mathbf{x}_2) & \dots & K(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

is **positive semi-definite**

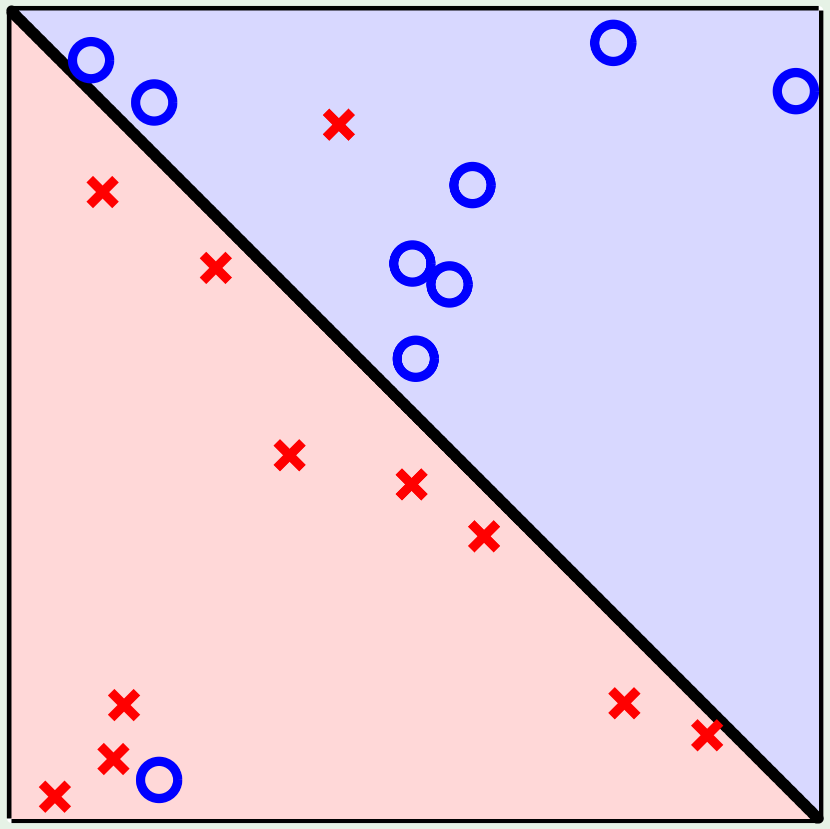
for any $\mathbf{x}_1, \cdots, \mathbf{x}_N$ (Mercer's condition)

Outline

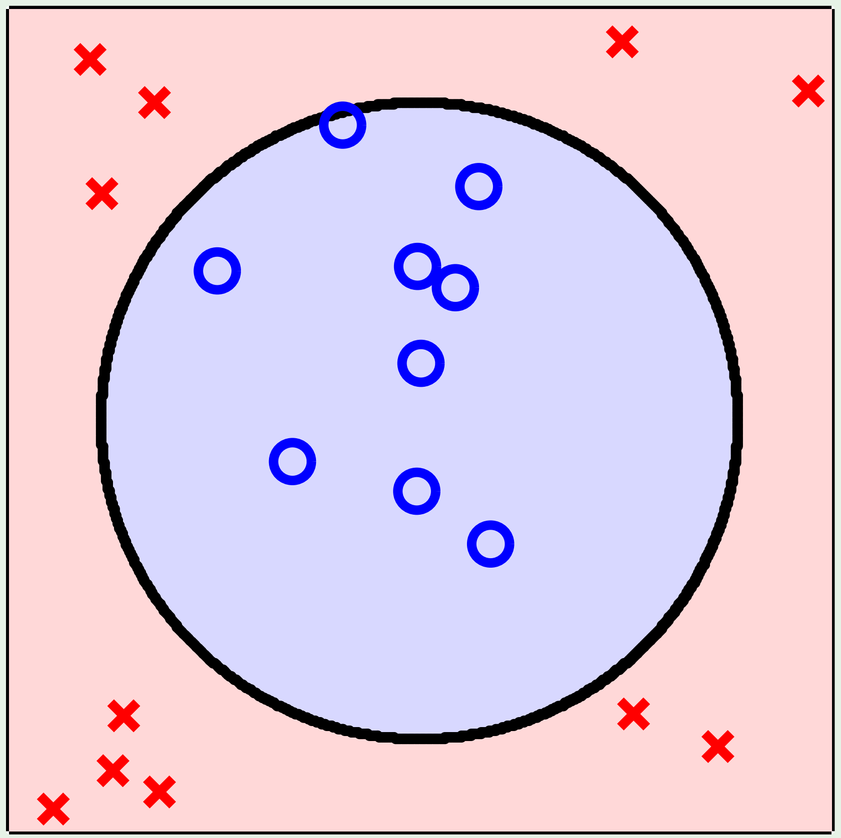
- The kernel trick
- Soft-margin SVM

Two types of non-separable

slightly:



seriously:

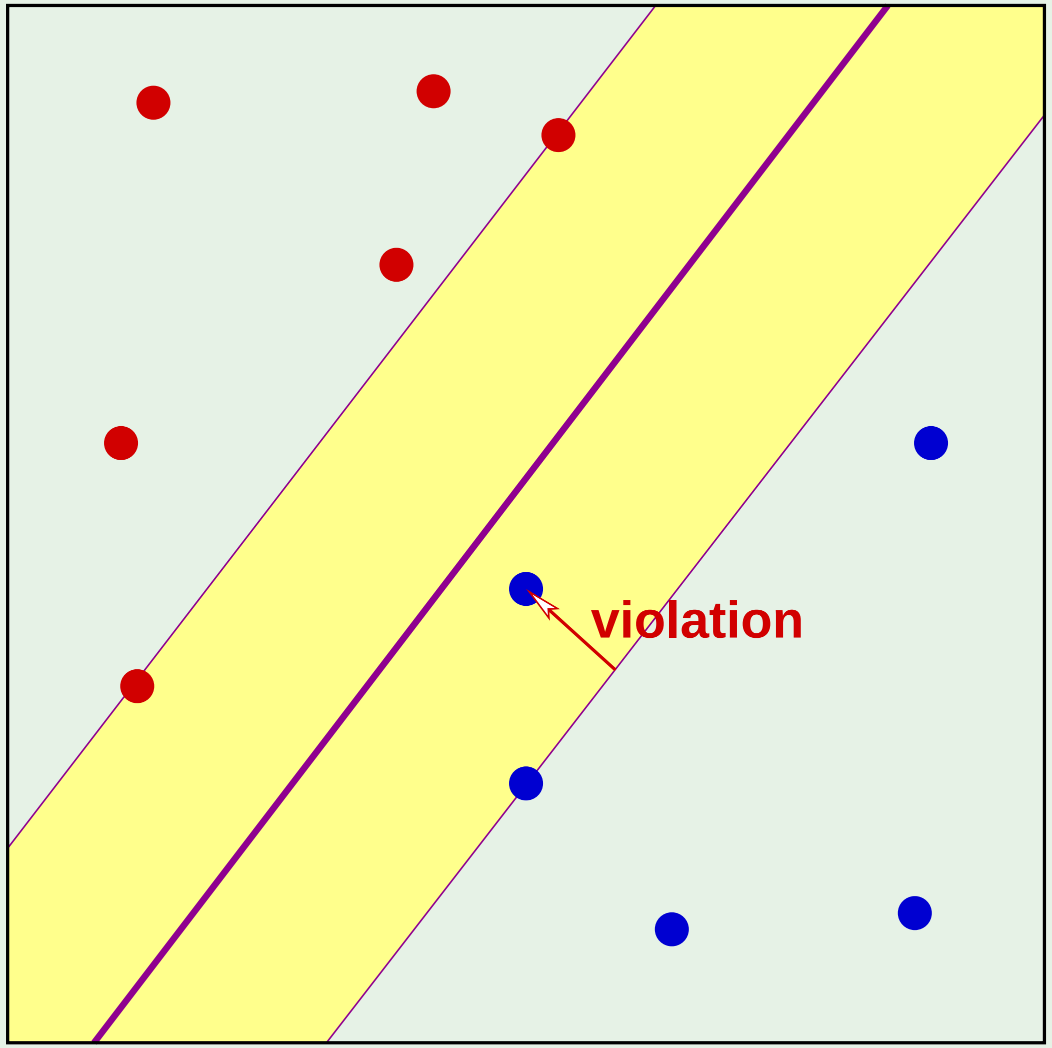


Error measure

Margin violation: $y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) \geq 1$ fails

Quantify: $y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) \geq 1 - \xi_n \qquad \xi_n \geq 0$

$$\text{Total violation} = \sum_{n=1}^{N} \xi_n$$



The new optimization

Minimize $\quad \frac{1}{2}\,\mathbf{w}^{\intercal}\mathbf{w} + C\sum_{n=1}^{N}\xi_n$

subject to $\quad y_n\left(\mathbf{w}^{\intercal}\mathbf{x}_n + b\right) \geq 1 - \xi_n \quad \text{for} \quad n = 1, \ldots, N$

and $\quad \xi_n \geq 0 \quad \text{for} \quad n = 1, \ldots, N$

$\mathbf{w} \in \mathbb{R}^d \ , \quad b \in \mathbb{R} \ , \quad \boldsymbol{\xi} \in \mathbb{R}^N$

Lagrange formulation

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \; = \frac{1}{2}\,\mathbf{w}^{\intercal}\mathbf{w} + C\sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \alpha_n (y_n\,(\mathbf{w}^{\intercal}\mathbf{x}_n + b) - 1 + \xi_n) - \sum_{n=1}^{N} \beta_n\,\xi_n$$

Minimize w.r.t. $\mathbf{w}$, b, and $\boldsymbol{\xi}$ and maximize w.r.t. each $\alpha_n \geq 0$ and $\beta_n \geq 0$

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

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

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \beta_n = 0$$

and the solution is . . .

Maximize $\mathcal{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \, \alpha_n \alpha_m \, \mathbf{x}_n^{\intercal} \mathbf{x}_m$ w.r.t. to $\boldsymbol{\alpha}$

subject to $0 \le \alpha_n \le C$ for $n = 1, \cdots, N$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$

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

minimizes $\frac{1}{2} \mathbf{w}^{\intercal} \mathbf{w} + C \sum_{n=1}^{N} \xi_n$

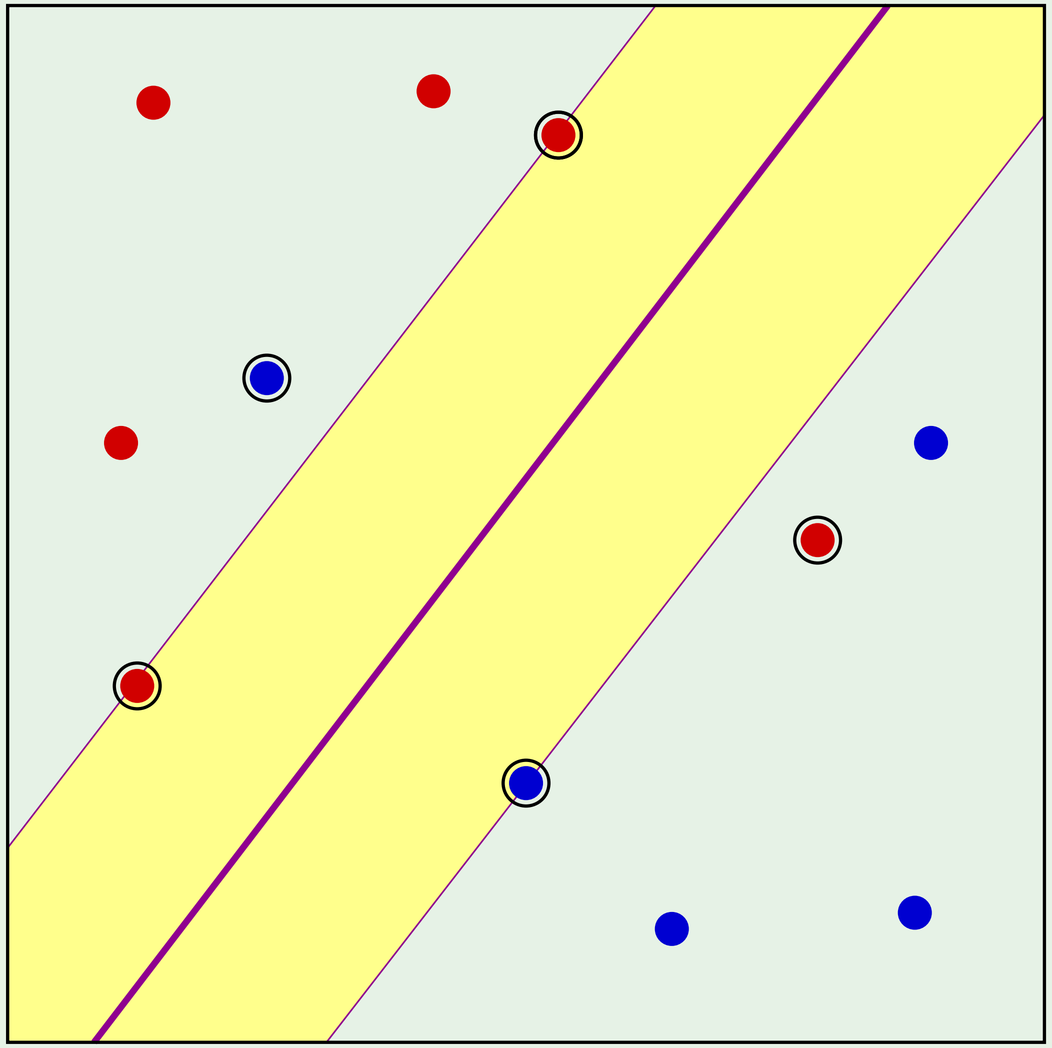
Types of support vectors

margin support vectors $\quad (0 < \alpha_n < C)$

$$y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) = 1 \qquad (\xi_n \;=\; 0)$$

non-margin support vectors $\quad (\alpha_n = C)$

$$y_n\left(\mathbf{w}^\intercal\mathbf{x}_n + b\right) < 1 \qquad (\xi_n \;>\; 0)$$



Two technical observations

1. Hard margin: What if data is not linearly separable?

   "primal $\longrightarrow$ dual" breaks down

2. $\mathcal{Z}$: What if there is w_0?

   All goes to b and $w_0 \rightarrow 0$