

PS5841

# Data Science in Finance & Insurance

## Support Vector Classifier

Yubo Wang

Autumn 2021

# Geometry (2D)

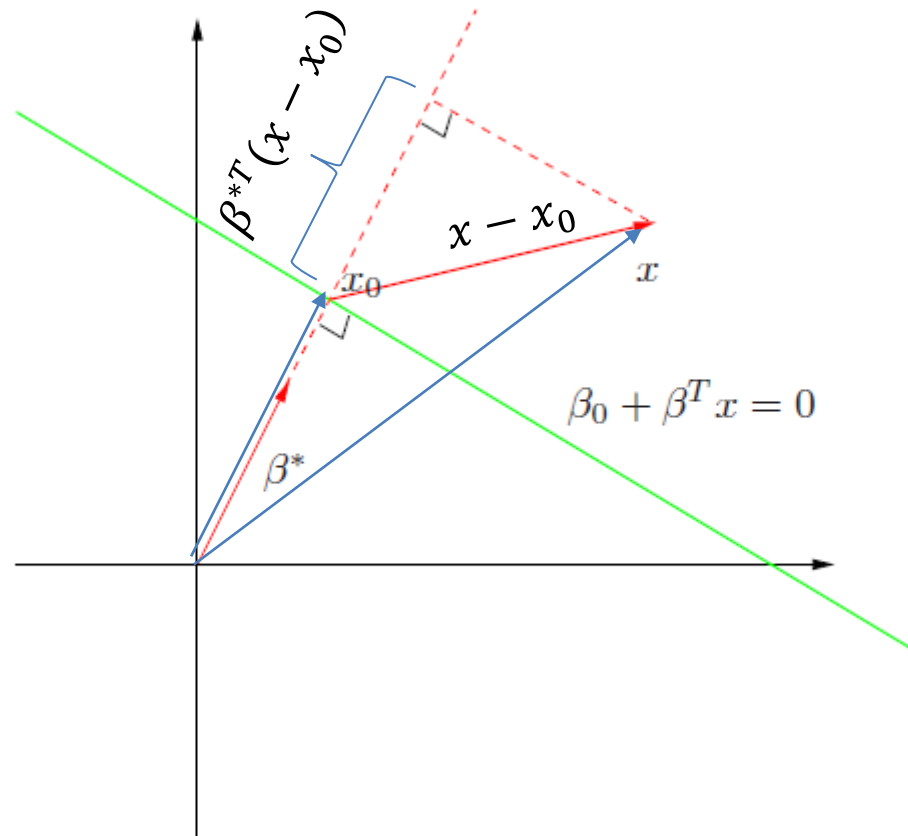
- Perpendicular lines

$$\text{slope}_1 \times \text{slope}_2 = -1$$

- Example

- The vector  $(\beta_1, \beta_2)$  has slope  $\frac{\beta_2}{\beta_1}$
- The line  $\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$  has slope  $-\frac{\beta_1}{\beta_2}$
- Are perpendicular to each other

# Geometry (higher dimensions)



$\beta^*$  is a unit vector

# Geometry (higher dimensions)

- $\boldsymbol{\beta}$  is normal to the hyperplane  $L$  defined by
$$\{\boldsymbol{x} | \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} = 0\}$$

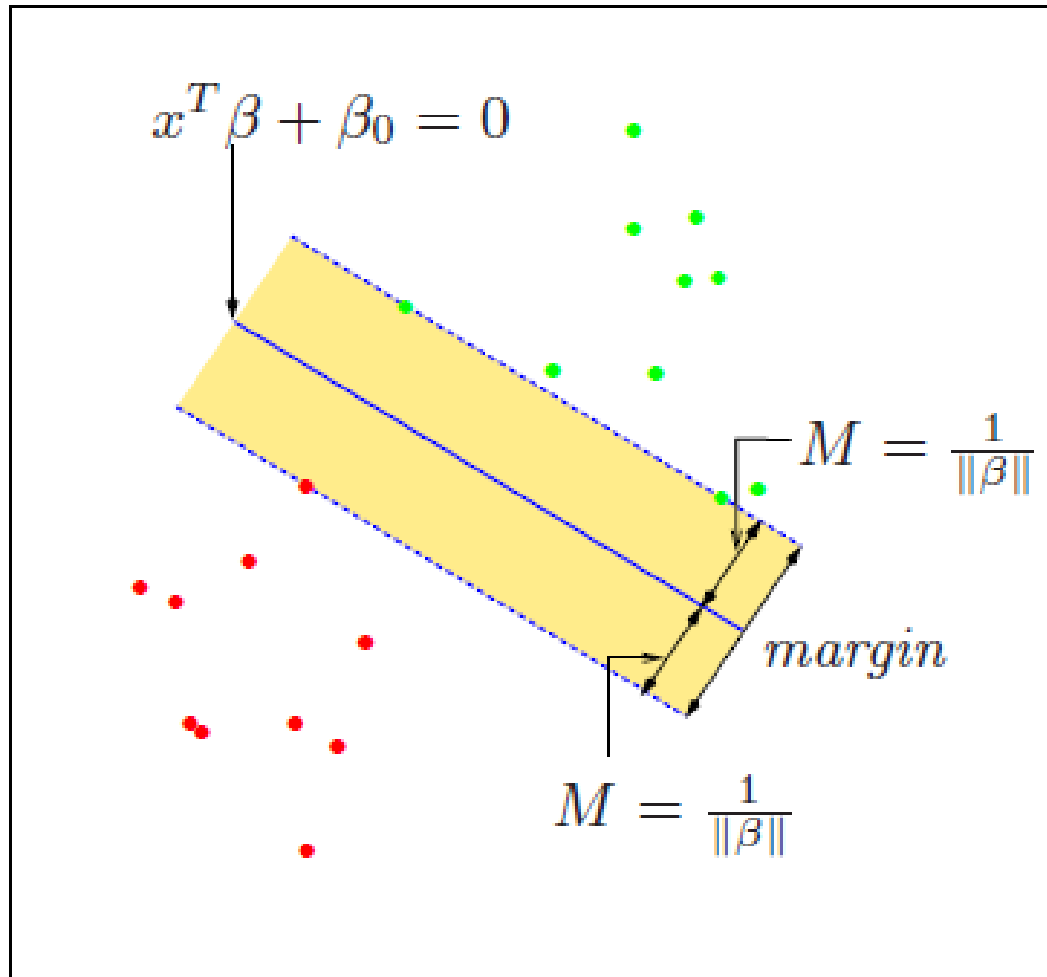
- For  $\boldsymbol{x}_1$  and  $\boldsymbol{x}_2$  on  $L$ ,
$$\boldsymbol{\beta}^T (\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$$

- The signed distance from  $\boldsymbol{x}$  to  $L$  is

$$\left( \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right)^T (\boldsymbol{x} - \boldsymbol{x}_0) = \frac{1}{\|\boldsymbol{\beta}\|} (\boldsymbol{\beta}^T \boldsymbol{x} - \boldsymbol{\beta}^T \boldsymbol{x}_0) = \frac{1}{\|\boldsymbol{\beta}\|} (\boldsymbol{x}^T \boldsymbol{\beta} + \beta_0)$$

- $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta} + \beta_0$  is proportional to the signed distance from  $\boldsymbol{x}$  to  $L$

# Separable Case



# Optimal Separating Hyperplane

- Label  $y_i$  indicates where  $\mathbf{x}_i$  is in relation to  $L$

$$y_i = \begin{cases} +1, & \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} > 0 \\ -1, & \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} < 0 \end{cases}$$

- $\mathbf{x}_i$  is at least a distance  $M$  from  $L$

$$y_i \frac{1}{\|\boldsymbol{\beta}\|} (\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq M \rightarrow y_i (\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq \|\boldsymbol{\beta}\| M$$

- The optimal separating hyperplane is the maximum margin hyperplane that maximizes  $M$

# Optimization Problem

- Maximizing  $M$  is equivalent to minimizing  $\|\boldsymbol{\beta}\|$ . WLOG, set  $\|\boldsymbol{\beta}\| = \frac{1}{M}$

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2$$

subject to

$$y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq 1, \quad \forall i$$

# Optimization (1)

- Lagrange primal

$$\begin{aligned} L_P &= \frac{1}{2} \|\boldsymbol{\beta}\|^2 - \sum_{i=1}^N \alpha_i [y_i (\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - 1] \\ &= \frac{1}{2} (\beta_0^2 + \dots + \beta_p^2) - \sum_{i=1}^N [\alpha_i y_i (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) - \alpha_i] \end{aligned}$$

- Set partial derivatives to zero

$$\begin{aligned} \frac{\partial L_P}{\partial \beta_0} = 0 &\rightarrow \sum_{i=1}^N \alpha_i y_i = 0 \\ \frac{\partial L_P}{\partial \beta_{j \neq 0}} = 0 &\rightarrow \boldsymbol{\beta} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \end{aligned}$$

- Get  $L_D$  by substitute these into  $L_P$



# Optimization (2)

- Wolfe dual

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k$$

$$\text{subject to } \alpha_i \geq 0, \quad 0 = \sum_{i=1}^N \alpha_i y_i, \quad \forall i$$

- Solutions satisfy the Karush-Kuhn-Tucker (KKT) conditions

$$\sum_{i=1}^N \alpha_i y_i = 0$$

$$\boldsymbol{\beta} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$$

$$\alpha_i \geq 0, \forall i$$

$$\alpha_i [y_i (\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - 1] = 0, \forall i$$

# Support Vectors

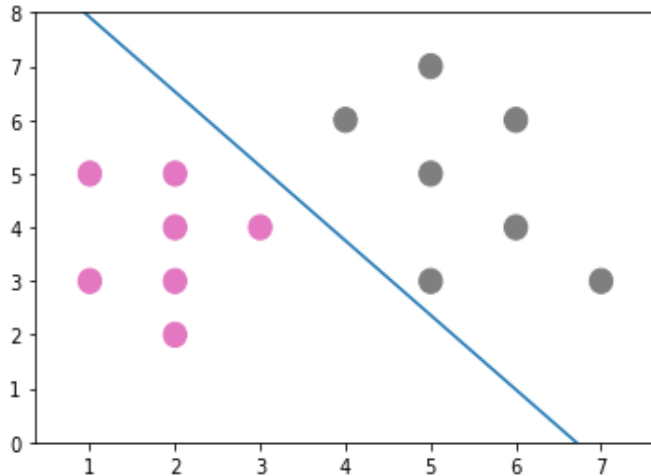
- The margin around the linear decision boundary has thickness  $M = \frac{1}{\|\boldsymbol{\beta}\|}$
- For any  $\mathbf{x}_i$  more than  $M$  away from the boundary,  $y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) > 1$   
$$\alpha_i [y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - 1] = 0 \rightarrow \alpha_i = 0$$
- The support vectors, those on the margin and  $\alpha_i > 0$ , define the decision boundary
- For SVs,  $y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - 1 = 0$
- $\boldsymbol{\beta} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$  is a linear combination of support vectors

# Maximal Margin Classifier

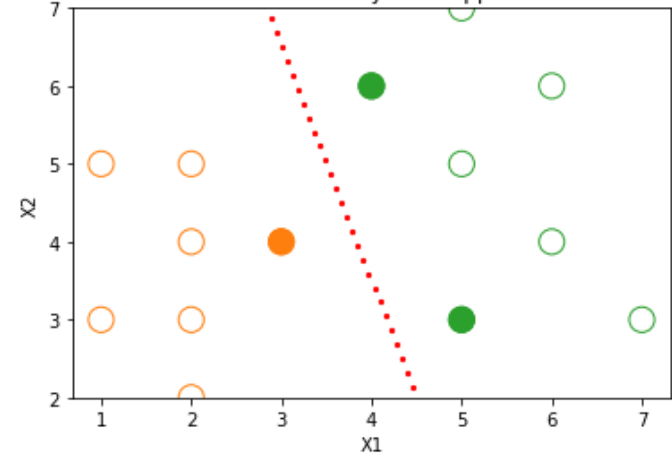
$$\text{clf}(\mathbf{x}) = \text{sign}[\mathbf{x}^T \hat{\boldsymbol{\beta}} + \hat{\beta}_0]$$

# Decision Boundary separable case

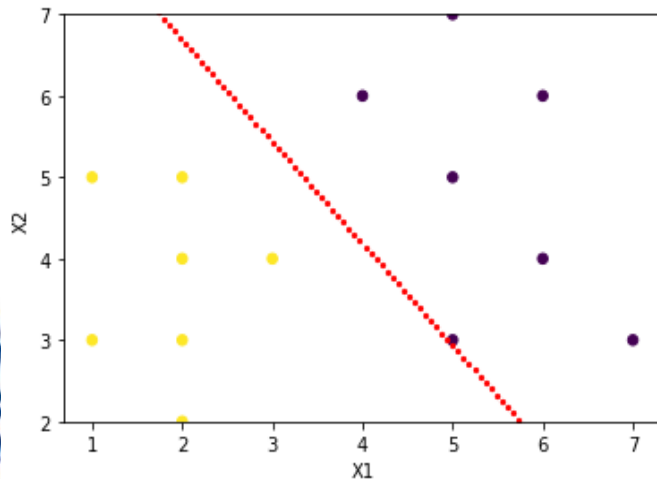
perceptron via coding



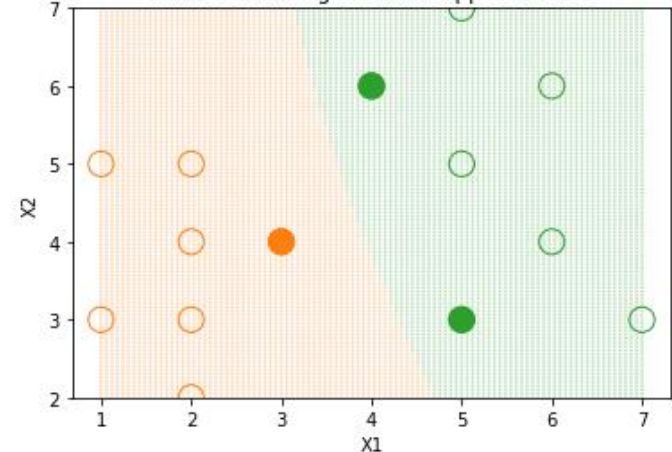
SVC decision boundary and support vectors



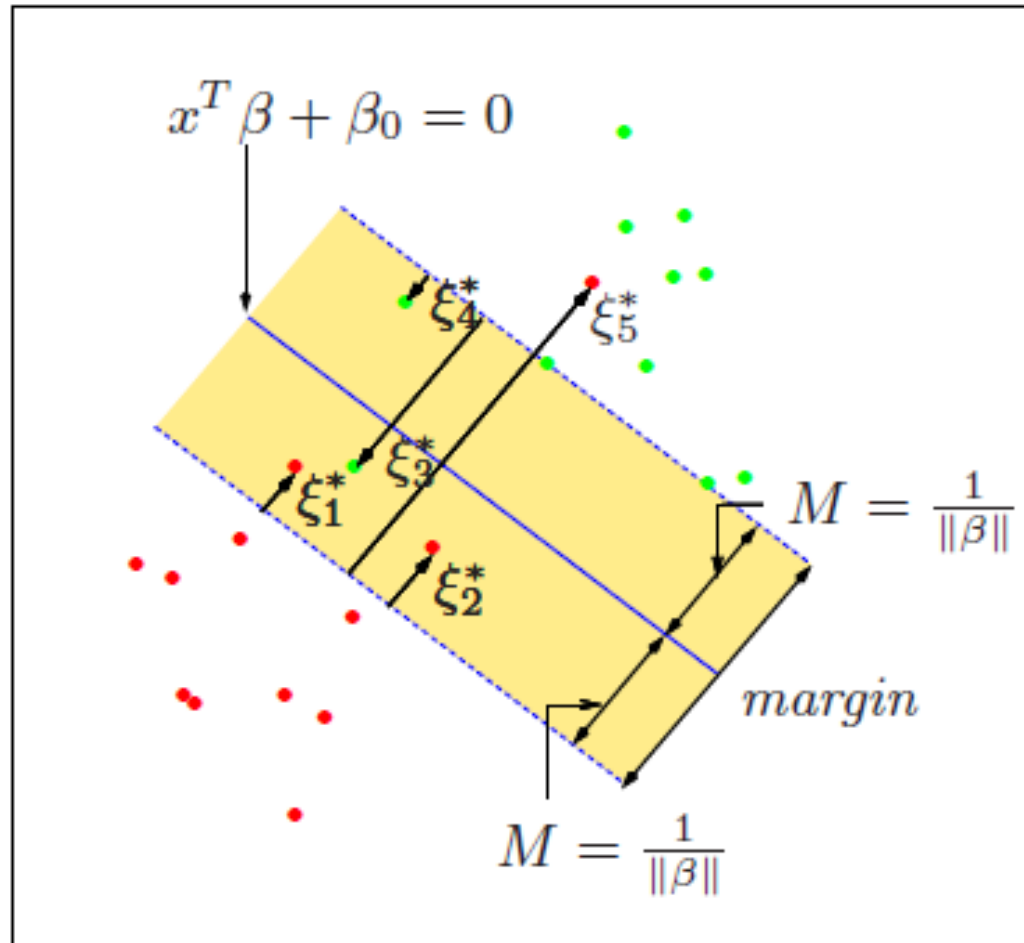
perceptron via ffn



SVM decision regions and support vectors



# Non-Separable Case



# Optimal Separating Hyperplane

- Label  $y_i$  indicates where  $\mathbf{x}_i$  is in relation to  $L$

$$y_i = \begin{cases} +1, & \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} > 0 \\ -1, & \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} < 0 \end{cases}$$

- $\mathbf{x}_i$  is at least a distance  $M$  from  $L$ , with allowance for some margin violation

$$y_i \frac{1}{\|\boldsymbol{\beta}\|} (\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq M(1 - \epsilon_i) \rightarrow$$
$$y_i (\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq \|\boldsymbol{\beta}\| M(1 - \epsilon_i)$$

- The slack variable  $\epsilon_i \geq 0$ , is the proportional amount by which  $\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0$  is on the wrong side of its margin
- The decision boundary is one that maximizes  $M$

# Slack Variable

- The slack variable  $\epsilon_i \geq 0$ , is the proportional amount by which  $\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0$  is on the wrong side of its margin

$$\epsilon_i = \begin{cases} = 0, & \text{OK wrt margin and } L \\ > 0, & \text{violates margin, OK wrt } L \\ > 1, & \text{misclassification} \end{cases}$$

# Optimization Problem

- Maximizing  $M$  is equivalent to minimizing  $\|\boldsymbol{\beta}\|$ .  
WLOG, set  $\|\boldsymbol{\beta}\| = \frac{1}{M}$

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^N \epsilon_i$$

subject to

$$\begin{aligned} y_i(\mathbf{x}^T \boldsymbol{\beta} + \beta_0) &\geq 1 - \epsilon_i, & \forall i \\ \epsilon_i &\geq 0 \end{aligned}$$

- cost parameter  $C$ 
  - penalty for margin violation
  - $C = \infty$  is for the separable case



# Optimization

- Lagrange primal

$$L_P = \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^N \epsilon_i - \sum_{i=1}^N \alpha_i [y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - (1 - \epsilon_i)] - \sum_{i=1}^N \mu_i \epsilon_i$$

- Wolfe dual

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k$$

- Solutions satisfy

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad \boldsymbol{\beta} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$$

$$\alpha_i = C - \mu_i,$$

$$\alpha_i [y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - (1 - \epsilon_i)] = 0$$

$$\mu_i \epsilon_i = 0$$

$$y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - (1 - \epsilon_i) \geq 0$$

$$\alpha_i, \mu_i, \epsilon_i \geq 0$$

# Support Vectors

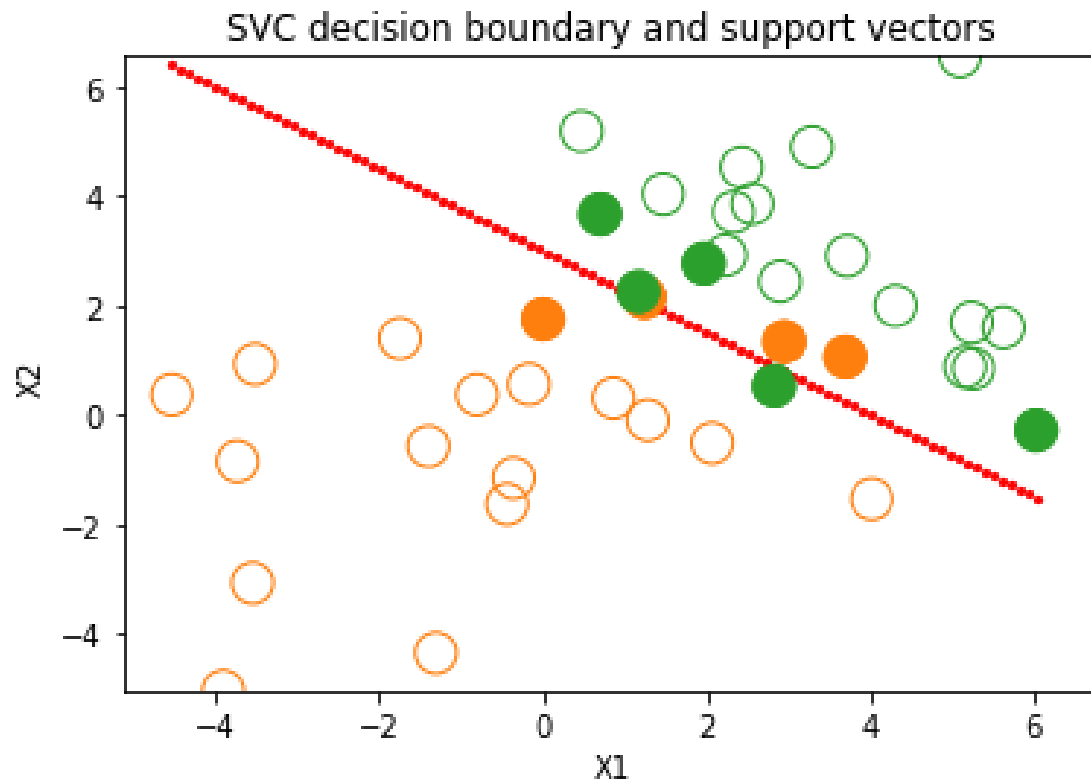
- Support vectors alone define the decision boundary
- Support vectors are those on or violate the margin
- Support vectors satisfy

$$y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - (1 - \epsilon_i) = 0$$

# Support Vector Classifier

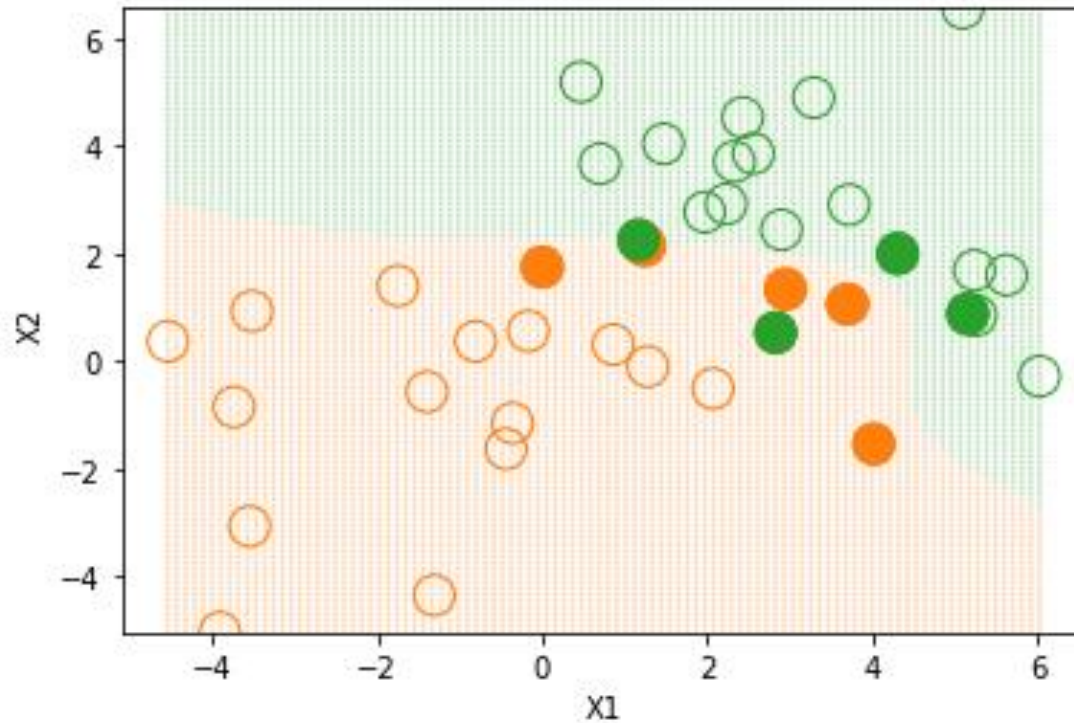
$$\text{clf}(\mathbf{x}) = \text{sign}[\mathbf{x}^T \hat{\boldsymbol{\beta}} + \hat{\beta}_0]$$

# Decision Boundary (SVC) non-separable case



# Decision Boundary (SVM) non-separable case

- Poly3, C=1



That was

