

Support Vector Machines: Soft Margin

Leandro L. Minku

Overview

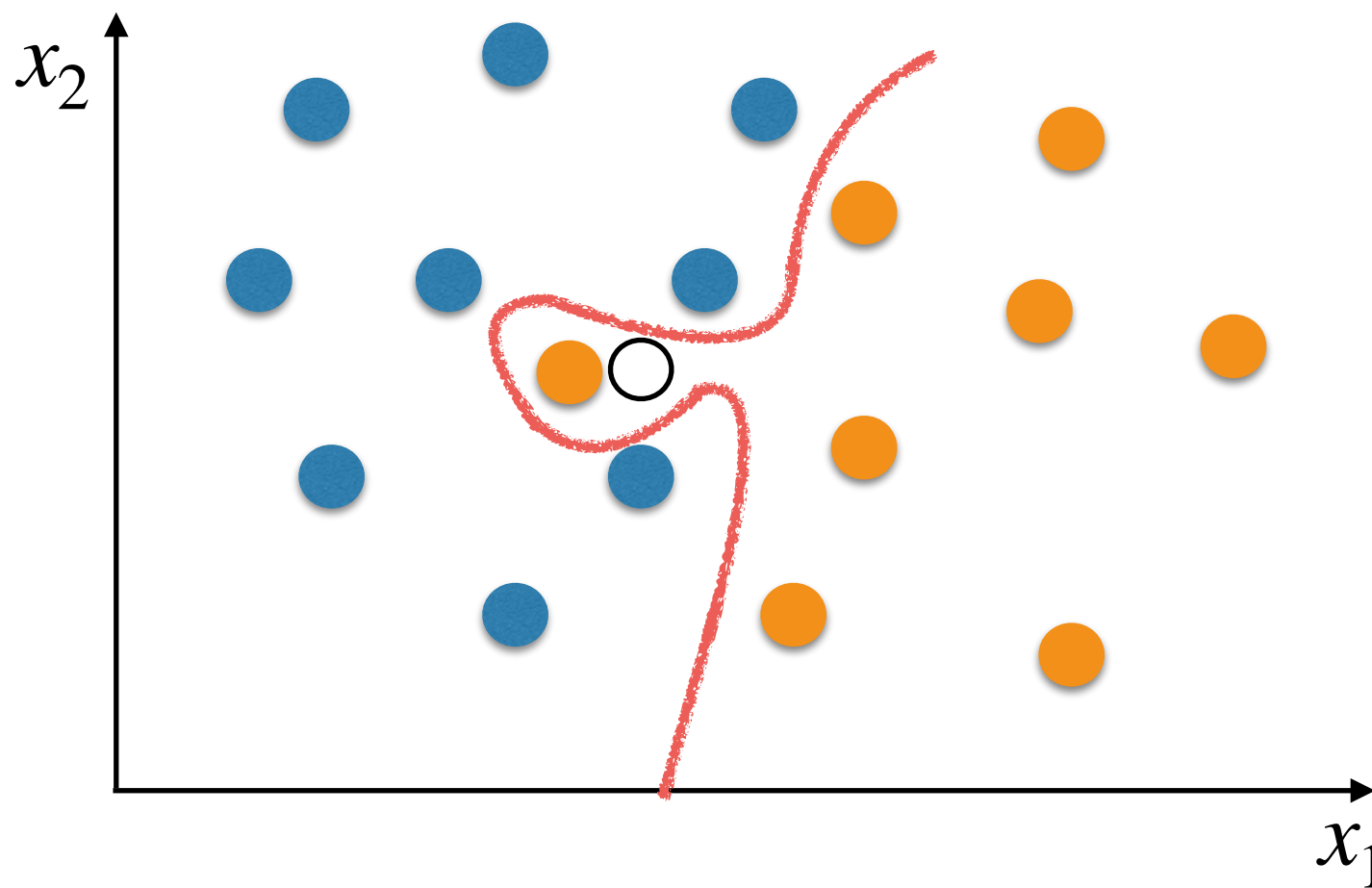
- Soft margin SVM
 - Primal
 - Dual
 - Making predictions

Overfitting?

- Overfitting happens when we fit noise in our training data and as a result worsen the generalisation capability.
- One may be concerned with overfitting if we are using such high dimensional embedding as the one underlying the Gaussian kernel.
- Maximising the margin can help coping with overfitting.
- Still, some overfitting may occur. For that, we will learn about the soft margin SVM next.

General Idea

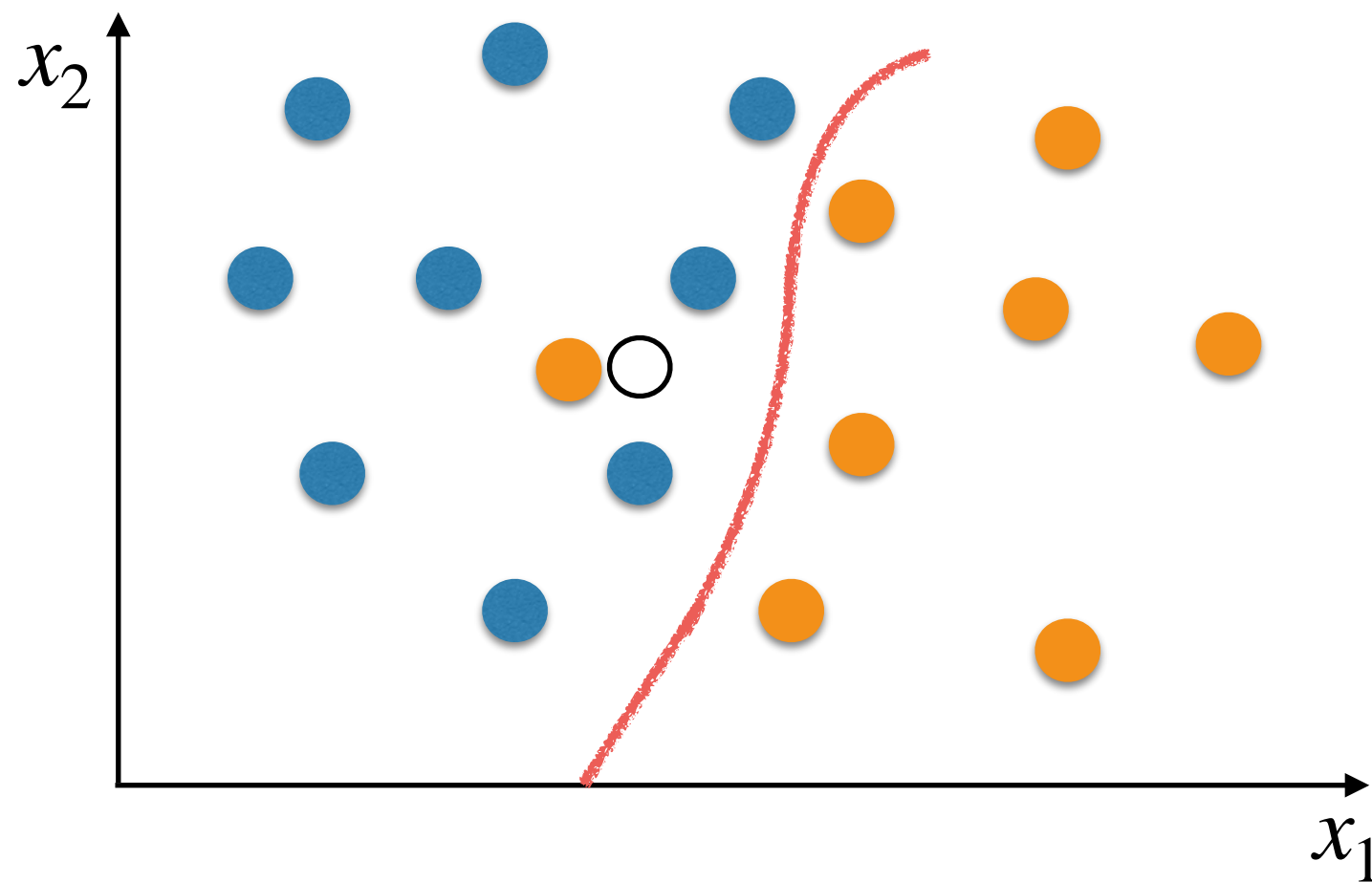
Our (kernelised) maximum margin classifiers assume that the training data are linearly separable (in the higher dimensional embedding).



They will try to perfectly separate the training examples, which may lead to poor generalisation.

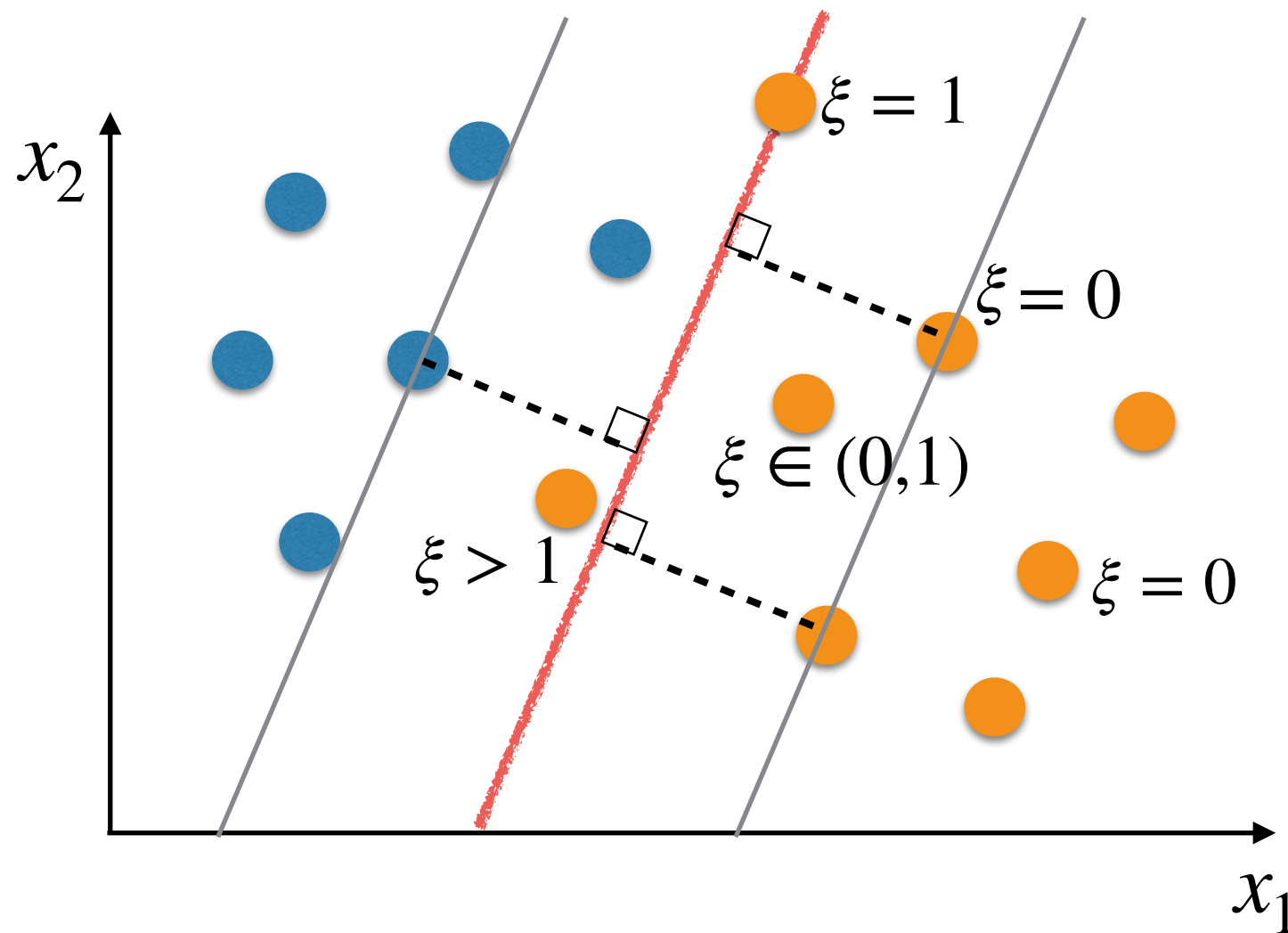
General Idea

It may be better to misclassify some training examples!



Slack Variables ξ

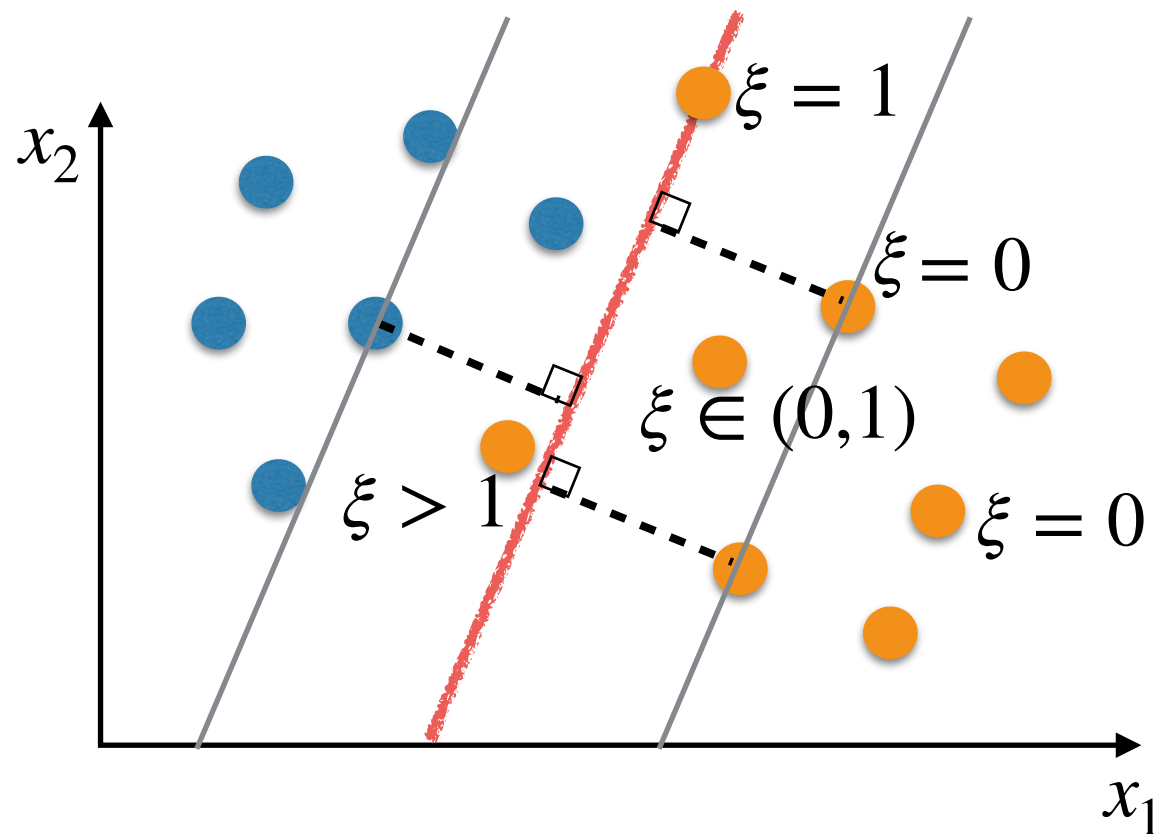
One slack variable $\xi^{(n)} \geq 0$ is associated to each training example $(\mathbf{x}^{(n)}, y^{(n)})$.



These variables tell us by how much an example can be within the margin or on the wrong side of the decision boundary.

$$y^{(n)}h(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)}$$

The Effect of ξ



$$y^{(n)}h(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)}$$

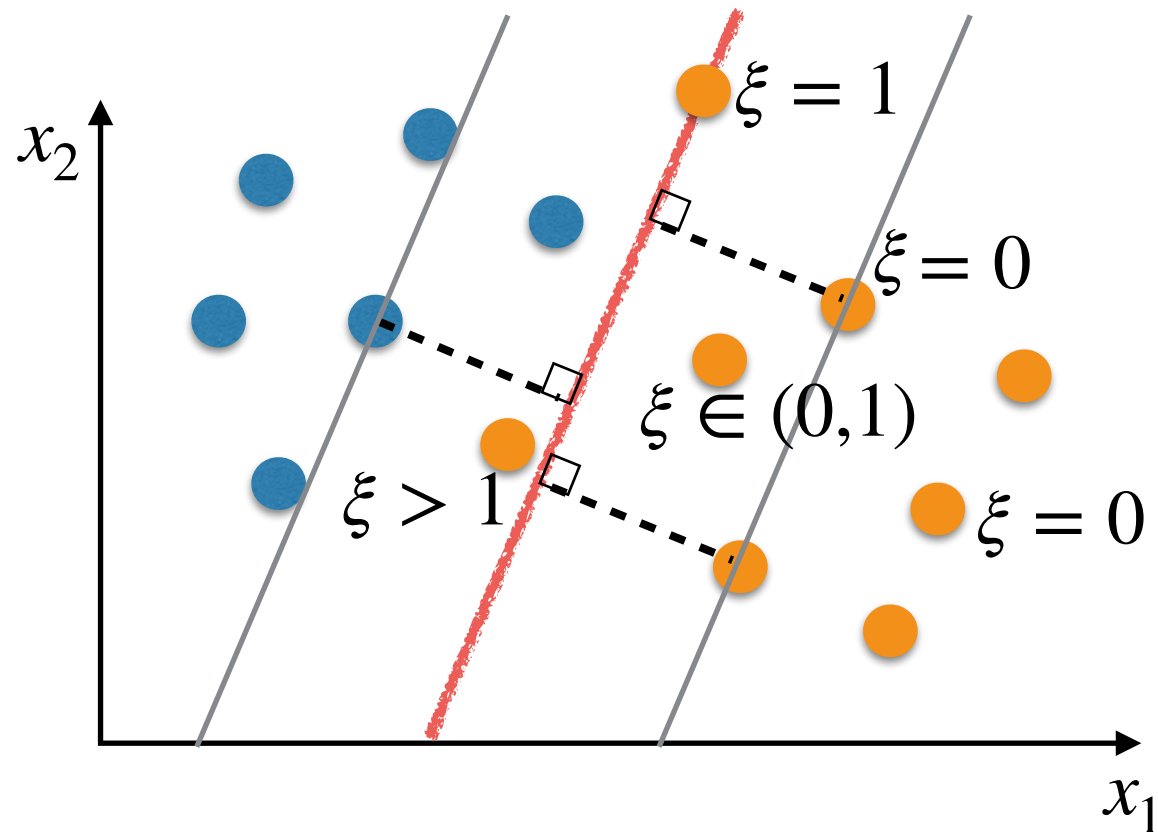
$$\xi^{(n)} = 0 \longrightarrow y^{(n)}h(\mathbf{x}^{(n)}) \geq 1$$

$$\xi^{(n)} = 1 \longrightarrow y^{(n)}h(\mathbf{x}^{(n)}) \geq 0$$

$$\xi^{(n)} \in (0,1) \longrightarrow y^{(n)}h(\mathbf{x}^{(n)}) \geq v \in (0,1)$$

$$\xi^{(n)} > 1 \longrightarrow y^{(n)}h(\mathbf{x}^{(n)}) \geq -v, v > 0$$

Margin



Our margin was previously defined by

$$\text{dist}(h, \mathbf{x}^{(k)}) = \frac{y^{(k)}h(\mathbf{x}^{(k)})}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

where $(\mathbf{x}^{(k)}, y^{(k)})$ was the closest example to the decision boundary.

Now, our margin is simply defined as $1/\|\mathbf{w}\|$.

Our New Optimisation Problem

- Recap of our optimisation problem (primal representation):

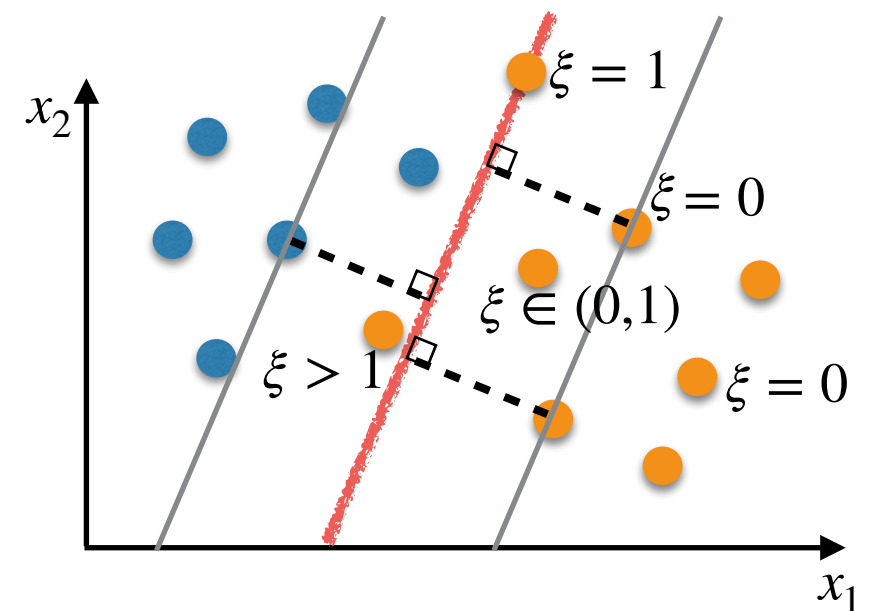
$$\operatorname{argmin}_{\mathbf{w}, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\}$$

Subject to: $y^{(n)} h(\mathbf{x}^{(n)}) \geq 1, \forall n \in \{1, 2, \dots, N\}$

- Using Slack

$$\operatorname{argmin}_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} \right\}$$

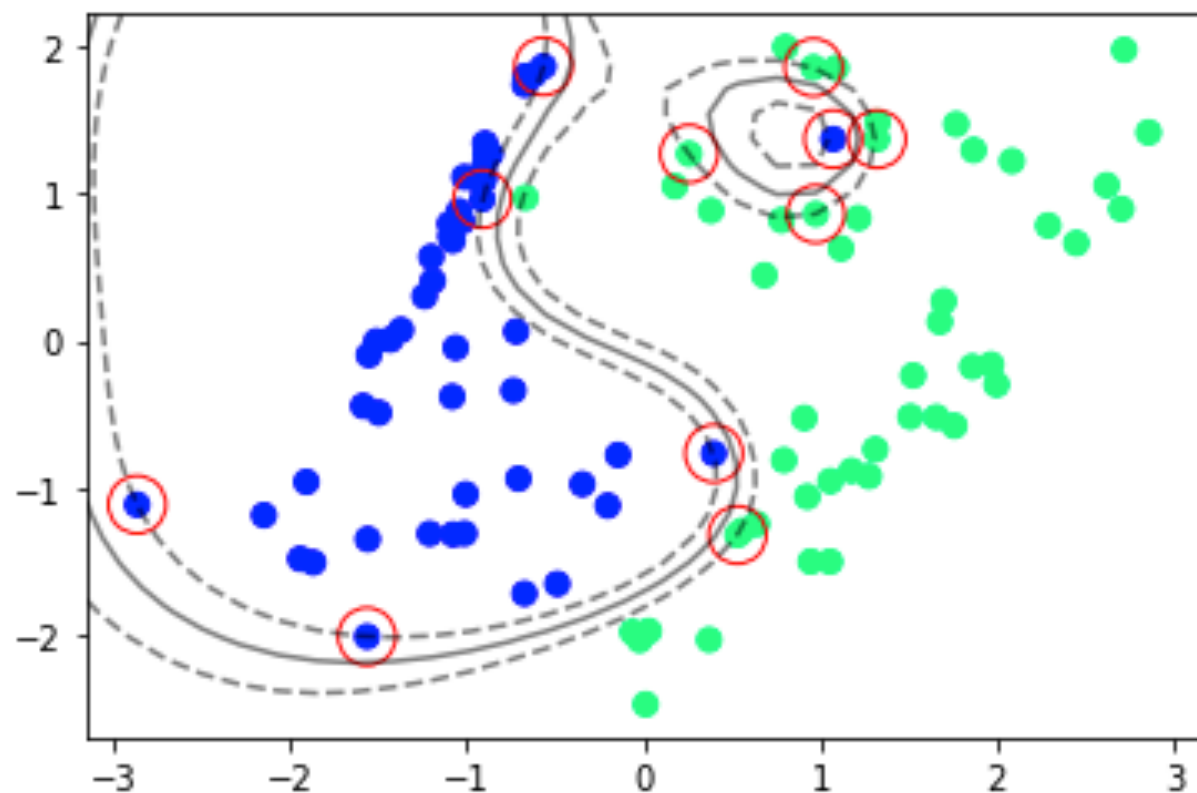
Subject to: $y^{(n)} h(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)}, \forall n \in \{1, 2, \dots, N\}$
 $\xi^{(n)} \geq 0$



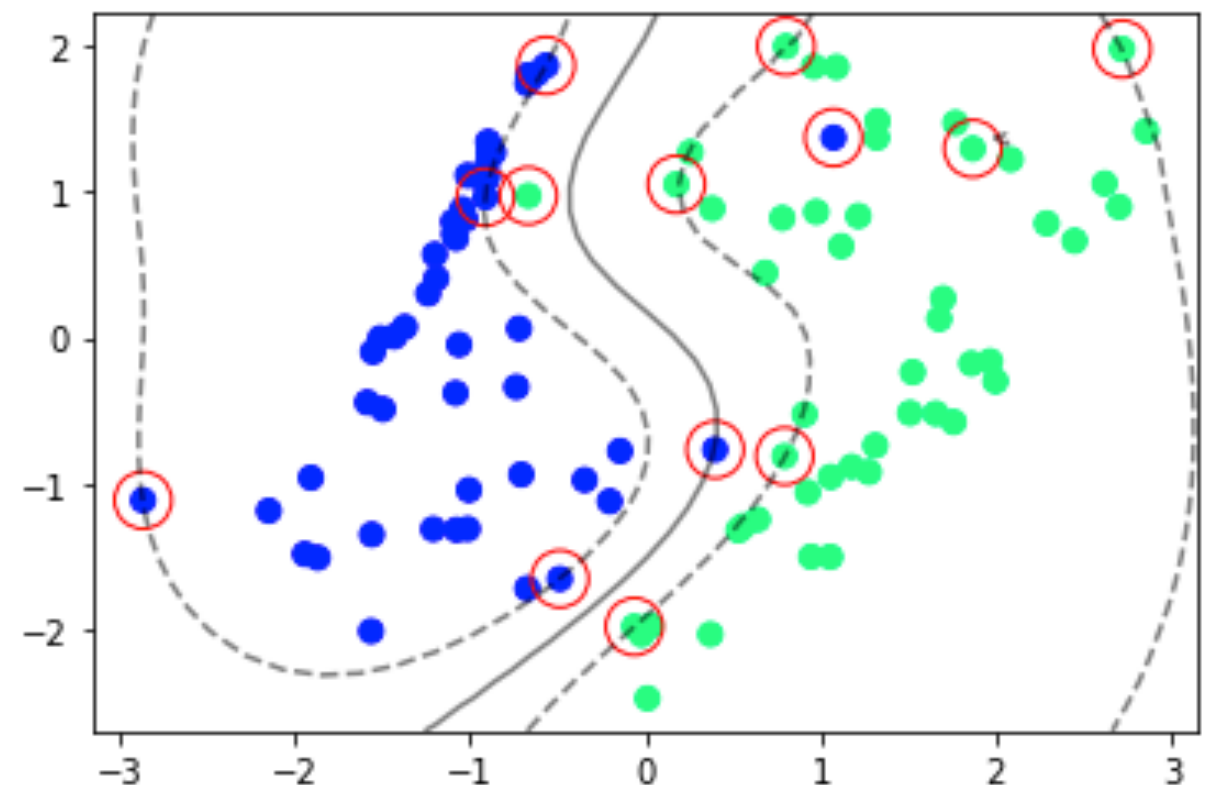
C is a hyperparameter that controls the trade-off between the amount of slack and the margin

Examples

Larger C



Smaller C



Code adapted from: <https://jakevdp.github.io/PythonDataScienceHandbook/05.07-support-vector-machines.html>

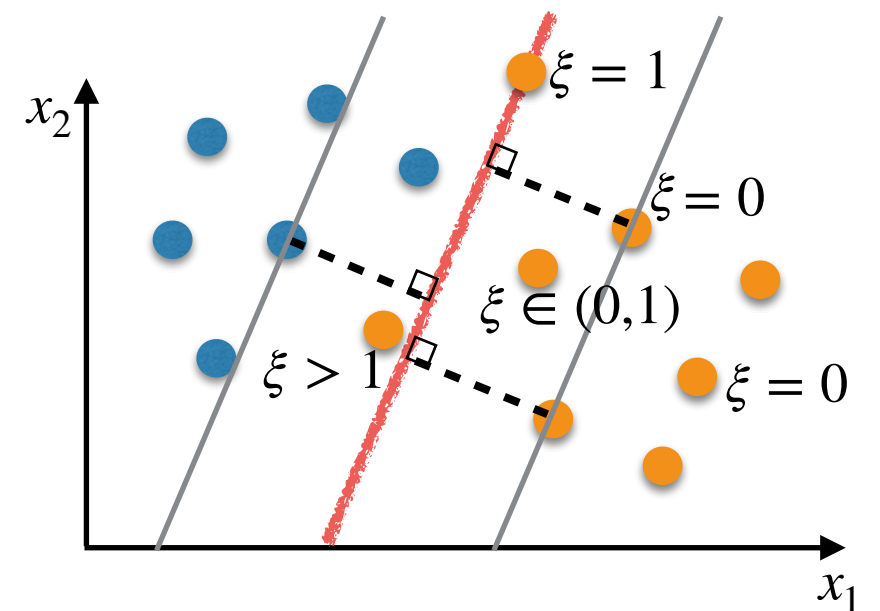
$$\operatorname{argmin}_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} \right\}$$

Soft Margin SVM

- Recap of our optimisation problem (primal representation):

$$\operatorname{argmin}_{\mathbf{w}, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\}$$

Subject to: $y^{(n)} h(\mathbf{x}^{(n)}) \geq 1, \forall n \in \{1, 2, \dots, N\}$



- Using Slack

$$\operatorname{argmin}_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} \right\}$$

Subject to: $y^{(n)} h(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)}, \forall n \in \{1, 2, \dots, N\}$
 $\xi^{(n)} \geq 0$

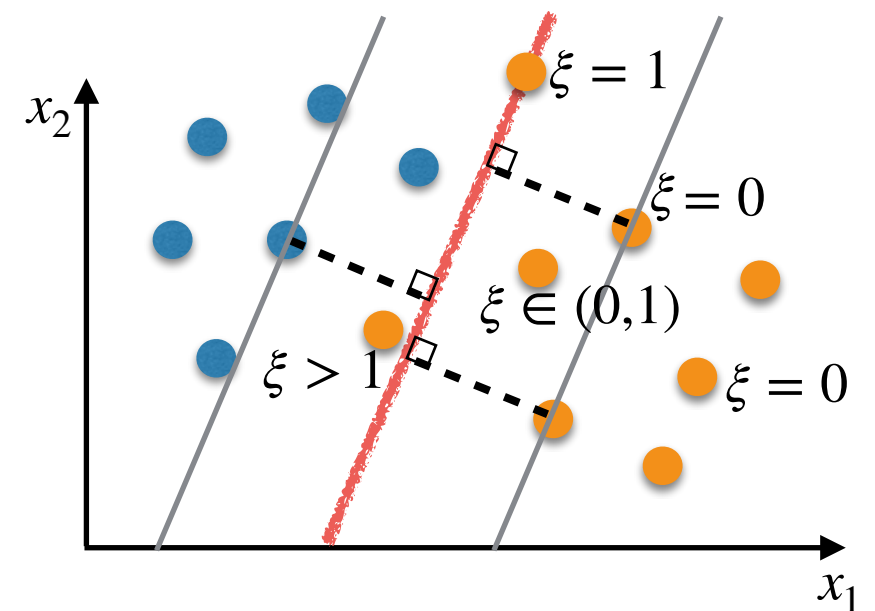
We will allow examples to be within the margin or in the wrong side of the decision boundary based on $\xi^{(n)}$

Soft Margin SVM

- Recap of our optimisation problem (primal representation):

$$\operatorname{argmin}_{\mathbf{w}, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\}$$

Subject to: $y^{(n)} h(\mathbf{x}^{(n)}) \geq 1, \forall n \in \{1, 2, \dots, N\}$



- Using Slack

$$\operatorname{argmin}_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} \right\}$$

Subject to: $y^{(n)} h(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)}, \forall n \in \{1, 2, \dots, N\}$
 $\xi^{(n)} \geq 0$

When we allow slacks > 0 , the margin is called a “**soft margin**”, as opposed to a “**hard margin**”.

Making Predictions

$$h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \begin{cases} \mathbf{w}^T \mathbf{x} + b > 0 \rightarrow \text{class } +1 \\ \mathbf{w}^T \mathbf{x} + b < 0 \rightarrow \text{class } -1 \end{cases}$$

$$h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \begin{cases} \mathbf{w}^T \mathbf{x} + b > 0 \rightarrow \text{class } +1 \\ \mathbf{w}^T \mathbf{x} + b < 0 \rightarrow \text{class } -1 \end{cases}$$

Soft Margin SVM: From Primal to Dual

Primal
Formulation

$$\operatorname{argmin}_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} \right\} \quad \text{Subject to: } \begin{aligned} & y^{(n)} h(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)} \\ & \xi^{(n)} \geq 0 \\ & \forall n \in \{1, 2, \dots, N\} \end{aligned}$$



Lagrange
Relaxation

$$\text{Subject to: } y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b) - 1 + \xi^{(n)} \geq 0$$

$$\text{Subject to: } 1 - \xi^{(n)} - y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b) \leq 0$$

$$-\xi^{(n)} \leq 0$$

Recap: Lagrange Relaxation

Primal
Formulation

$$\min_{\mathbf{x}} F(\mathbf{x})$$

Subject to: $f_i(\mathbf{x}) \leq 0, i \in \{1, \dots, n\}$

Lagrange
Relaxation

$$\min_{\mathbf{x}} F(\mathbf{x}) + \sum_{i=1}^n a_i f_i(\mathbf{x})$$

where $a_i \geq 0, i \in \{1, \dots, n\}$ are the Lagrange multipliers and

$L(\mathbf{x}, \mathbf{a}) = F(\mathbf{x}) + \sum_{i=1}^n a_i f_i(\mathbf{x})$ is the Lagrangian.

Soft Margin SVM: From Primal to Dual

Primal
Formulation

$$\operatorname{argmin}_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} \right\} \quad \text{Subject to:}$$

$$1 - \xi^{(n)} - y^{(n)}(\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b) \leq 0$$

$$-\xi^{(n)} \leq 0$$

$$\forall n \in \{1, 2, \dots, N\}$$

Lagrange
Relaxation

$$\min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} + \sum_{n=1}^N a^{(n)} (1 - \xi^{(n)} - y^{(n)}(\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) - \sum_{n=1}^N \beta^{(n)} \xi^{(n)} \right\}$$

Where N is the number of training examples.

Subject to: $a^{(n)} \geq 0, \beta^{(n)} \geq 0 \quad \forall (\mathbf{x}^{(n)}, y^{(n)}) \in \mathcal{T}$

Recap: Minimax Primal and Dual Formulation

Minimax Primal
Formulation

$$\min_{\mathbf{x}} \max_{\mathbf{a}} \left(F(\mathbf{x}) + \sum_{i=1}^N a_i f_i(\mathbf{x}) \right)$$

Dual
Formulation

$$\max_{\mathbf{a}} \min_{\mathbf{x}} \left(F(\mathbf{x}) + \sum_{i=1}^N a_i f_i(\mathbf{x}) \right)$$

where $a_i \geq 0$, $i \in \{1, \dots, N\}$ are the Lagrange multipliers and

$L(\mathbf{x}, \mathbf{a}) = F(\mathbf{x}) + \sum_{i=1}^N a_i f_i(\mathbf{x})$ is the Lagrangian.

Soft Margin SVM: From Primal to Dual

Lagrange
Relaxation

$$\min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} + \sum_{n=1}^N a^{(n)} (1 - \xi^{(n)} - y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) - \sum_{n=1}^N \beta^{(n)} \xi^{(n)} \right\}$$

Minimax Primal
Formulation

Find $\mathbf{w}, b, \xi, \mathbf{a}, \beta$ such that:

Subject to: $a^{(n)} \geq 0, \beta^{(n)} \geq 0, \forall (\mathbf{x}^{(n)}, y^{(n)}) \in \mathcal{T}$

$$\min_{\mathbf{w}, b, \xi} \max_{\mathbf{a}, \beta} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} + \sum_{n=1}^N a^{(n)} (1 - \xi^{(n)} - y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) - \sum_{n=1}^N \beta^{(n)} \xi^{(n)} \right\}$$

Dual
Formulation

$$\max_{\mathbf{a}, \beta} \min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} + \sum_{n=1}^N a^{(n)} (1 - \xi^{(n)} - y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) - \sum_{n=1}^N \beta^{(n)} \xi^{(n)} \right\}$$

Further Simplifying Equations

$$\max_{\mathbf{a}, \beta} \min_{\mathbf{w}, b, \xi} \left\{ L(\mathbf{a}, \beta, \mathbf{w}, b, \beta) = \frac{1}{2} \|\mathbf{w}\|^2 + \underbrace{C}_{\text{blue}} \sum_{n=1}^N \underbrace{\xi^{(n)}}_{\text{blue}} + \sum_{n=1}^N \underbrace{a^{(n)}}_{\text{red}} (1 - \underbrace{\xi^{(n)}}_{\text{red}} - y^{(n)} (\underbrace{\mathbf{w}^T \phi(\mathbf{x}^{(n)})}_{\text{red}} + b)) - \sum_{n=1}^N \underbrace{\beta^{(n)}}_{\text{purple}} \underbrace{\xi^{(n)}}_{\text{purple}} \right\}$$

Subject to: $\underbrace{a^{(n)}}_{\text{red}} \geq 0, \underbrace{\beta^{(n)}}_{\text{purple}} \geq 0, \forall (\mathbf{x}^{(n)}, y^{(n)}) \in \mathcal{T}$

Once \mathbf{a}, β are fixed, there are no constraints and, at the optimum, $\nabla_{\mathbf{w}} L$ equals to zero (KKT stationarity):

$$\mathbf{w} - \sum_{n=1}^N a^{(n)} y^{(n)} \phi(\mathbf{x}^{(n)}) = 0 \longrightarrow \mathbf{w} = \sum_{n=1}^N a^{(n)} y^{(n)} \phi(\mathbf{x}^{(n)})$$

$$\text{And so does } \frac{\partial L}{\partial b}: \sum_{n=1}^N a^{(n)} y^{(n)} = 0$$

** Same as in the hard margin SVM!

Further Simplifying Equations

$$\max_{\mathbf{a}, \beta} \min_{\mathbf{w}, b, \xi} \left\{ L(\mathbf{a}, \beta, \mathbf{w}, b, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} + \sum_{n=1}^N a^{(n)} (1 - \xi^{(n)} - y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) - \sum_{n=1}^N \beta^{(n)} \xi^{(n)} \right\}$$

Subject to: $a^{(n)} \geq 0, \beta^{(n)} \geq 0, \forall (\mathbf{x}^{(n)}, y^{(n)}) \in \mathcal{T}$

Consider the following as a constraint to eliminate b :

$$\sum_{n=1}^N a^{(n)} y^{(n)} = 0$$

Substituting

$$\mathbf{w} = \sum_{n=1}^N a^{(n)} y^{(n)} \phi(\mathbf{x}^{(n)})$$

Further Simplifying Equations

$$\max_{\mathbf{a}, \beta} \min_{\mathbf{w}, b, \xi} \left\{ L(\mathbf{a}, \beta, \mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi^{(n)} + \sum_{n=1}^N a^{(n)} (1 - \xi^{(n)} - y^{(n)} (\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) - \sum_{n=1}^N \beta^{(n)} \xi^{(n)} \right\}$$

Subject to: $a^{(n)} \geq 0, \beta^{(n)} \geq 0, \forall (\mathbf{x}^{(n)}, y^{(n)}) \in \mathcal{T}$

Once \mathbf{a}, β are fixed, there are no constraints and, at the optimum, the following is true (KKT stationarity):

$$\frac{\partial L}{\partial \xi^{(n)}} = C - a^{(n)} - \beta^{(n)} = 0, \quad \forall n \in \{1, \dots, N\}$$

So, substitute $\beta^{(n)} = C - a^{(n)}$

Dual Representation

$$\operatorname{argmax}_{\mathbf{a}} \tilde{L}(\mathbf{a})$$

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^N a^{(n)} - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a^{(n)} a^{(m)} y^{(n)} y^{(m)} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})$$

Subject to: $a^{(n)} \geq 0, C - a^{(n)} \geq 0, \forall n \in \{1, \dots, N\} \quad \sum_{n=1}^N a^{(n)} y^{(n)} = 0$

Subject to: $0 \leq a^{(n)} \leq C, \forall n \in \{1, \dots, N\} \quad \sum_{n=1}^N a^{(n)} y^{(n)} = 0$

Box constraints

Dual Representation

- Dual formulation for Soft Margin SVM:

$$\operatorname{argmax}_{\mathbf{a}} \tilde{L}(\mathbf{a})$$

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^N a^{(n)} - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a^{(n)} a^{(m)} y^{(n)} y^{(m)} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})$$

$$\text{Subject to: } 0 \leq a^{(n)} \leq C, \forall n \in \{1, \dots, N\} \quad \sum_{n=1}^N a^{(n)} y^{(n)} = 0$$

- Recap of dual formulation for hard margin SVM:

$$\operatorname{argmax}_{\mathbf{a}} \tilde{L}(\mathbf{a})$$

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^N a^{(n)} - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a^{(n)} a^{(m)} y^{(n)} y^{(m)} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})$$

$$\text{Subject to: } a^{(n)} \geq 0, \forall n \in \{1, \dots, N\} \quad \sum_{n=1}^N a^{(n)} y^{(n)} = 0$$

Making Predictions

$$h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \begin{cases} \mathbf{w}^T \mathbf{x} + b > 0 \rightarrow \text{class } +1 \\ \mathbf{w}^T \mathbf{x} + b < 0 \rightarrow \text{class } -1 \end{cases}$$
$$\mathbf{w} = \sum_{n=1}^N a^{(n)} y^{(n)} \phi(\mathbf{x}^{(n)})$$

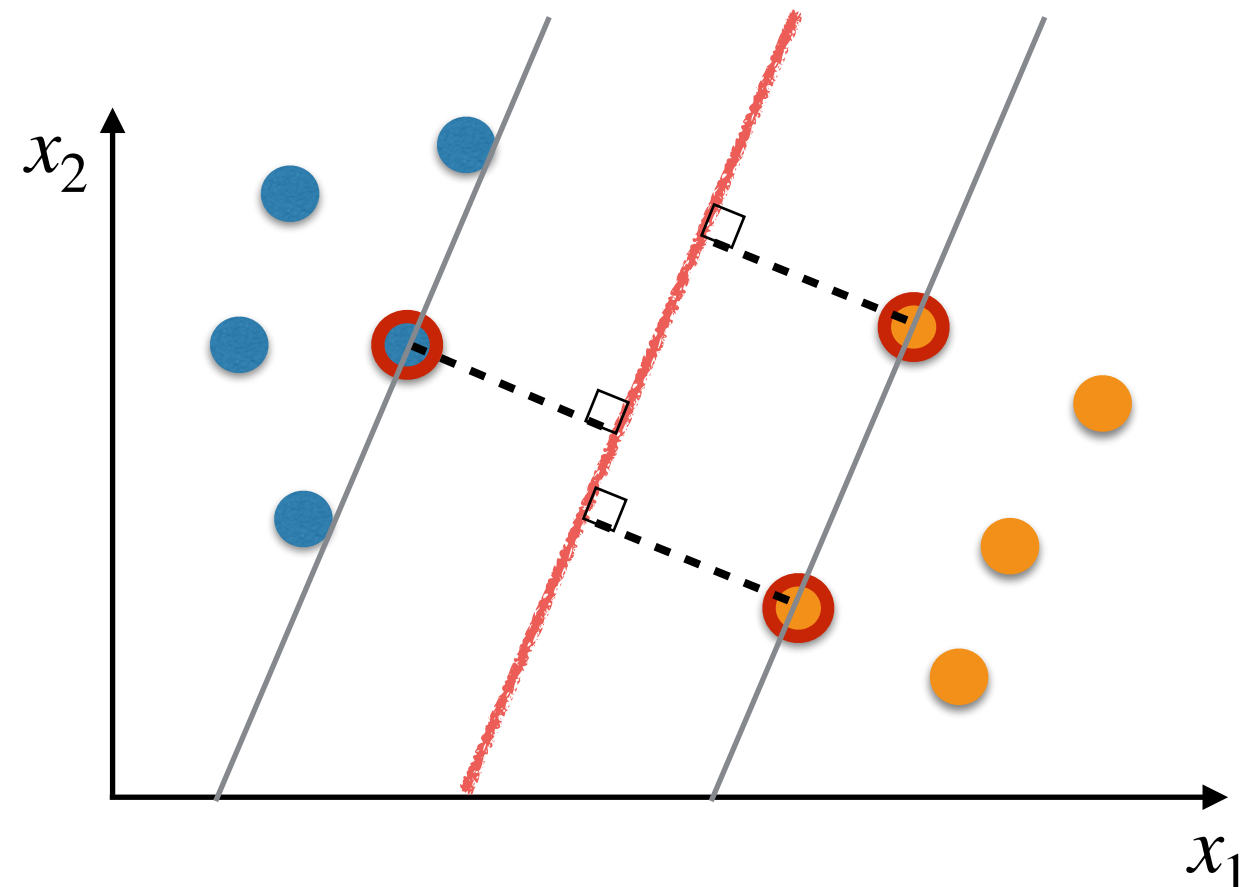
$$h(\mathbf{x}) = \sum_{n \in S} a^{(n)} y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)}) + b$$

Support Vectors and Making Predictions

- Before slack variables

$$h(\mathbf{x}) = \sum_{n \in S} a^{(n)} y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)}) + b$$

All examples for which $a^{(n)} \neq 0$ (and $y^{(n)} h(\mathbf{x}^{(n)}) = 1$, given KKT complementary slackness) are support vectors.

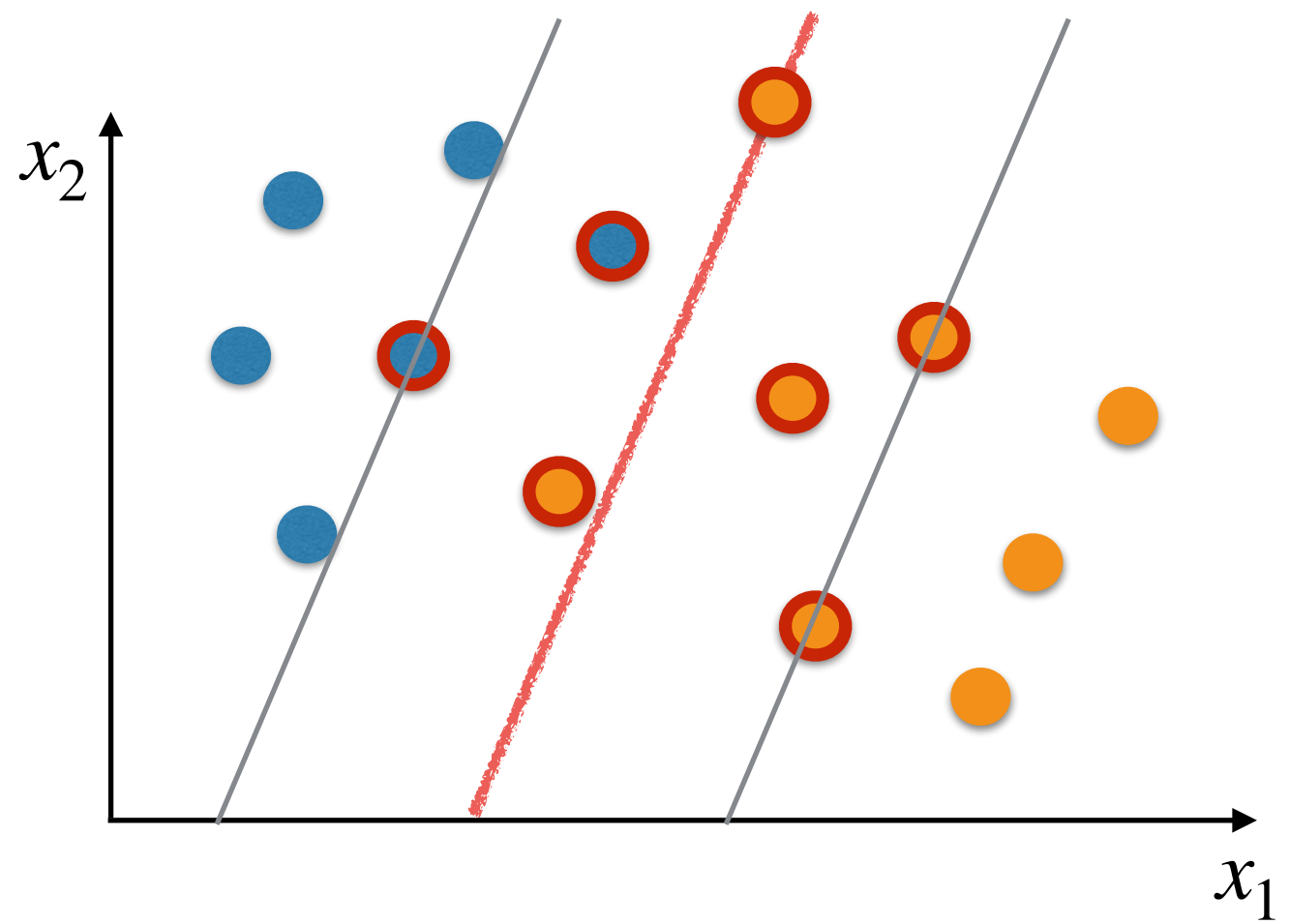


Support Vectors and Making Predictions

- With slack variables

$$h(\mathbf{x}) = \sum_{n \in S} a^{(n)} y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)}) + b$$

All examples for which $a^{(n)} \neq 0$ (and $y^{(n)} h(\mathbf{x}^{(n)}) = 1 - \xi^{(n)}$ due to KKT complementary slackness) are support vectors.

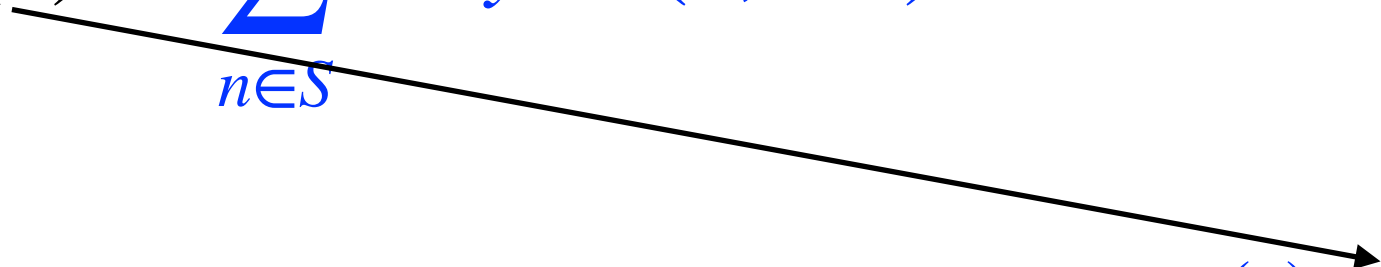


Support Vectors and Making Predictions

$$\text{Dual: } h(\mathbf{x}) = \sum_{n=1}^N a^{(n)} y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)}) + b \begin{cases} h(\mathbf{x}) > 0 \rightarrow \text{class } +1 \\ h(\mathbf{x}) < 0 \rightarrow \text{class } -1 \end{cases}$$

- Either: $a^{(n)} = 0$, so the value of $y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)})$ won't matter.
- Or: $a^{(n)} > 0$, so this is a support vector, i.e., an example that defines the position of the decision boundary.
 - From KKT complementary slackness,
 $a^{(n)}(1 - \xi^{(n)} - y^{(n)}(\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) = 0$.
 - So, $(1 - \xi^{(n)} - y^{(n)}(\mathbf{w}^T \phi(\mathbf{x}^{(n)}) + b)) = 0$ for these examples, i.e.,
 $y^{(n)} h(\mathbf{x}^{(n)}) = 1 - \xi^{(n)}$.
 - Therefore, support vectors are on the margin, within the margin, or on the wrong side of the decision boundary!

Calculation of b

$$h(\mathbf{x}) = \sum_{n \in S} a^{(n)} y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)}) + b$$


Our calculation of b was based on examples for which $y^{(n)} h(\mathbf{x}^{(n)}) = 1$.

$$b = \frac{1}{N_M} \sum_{n \in M} \left(y^{(n)} - \sum_{m \in S} a^{(m)} y^{(m)} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) \right)$$

where M is the set of indexes of the support vectors that are on the margin and N_M is the number of such support vectors, while S is the set of all support vectors.

Calculation of b

Our support vectors have $y^{(n)}h(x^{(n)}) = 1 - \xi^{(n)}$.
What if none of the support vectors has $y^{(n)}h(\mathbf{x}^{(n)}) = 1$?

We would get different values for b depending on $\xi^{(n)}$.

We can pick any of them to set b .

Summary

- We've seen how to make predictions when using the dual representation.
- Soft margin SVM allows some training examples to be within the margin or misclassified.
- This can improve generalisation.
- We can use soft margin SVM both in the primal and dual format.