Machine Learning

Lecture: Support Vector Machines

Hiroshi Shimodaira

2026

Ver. 1.1

Questions you should be able to answer after this week

- What is Support Vector Machine (SVM)?
- Training (optimisation problem) of linear SVM?
  What is maximum margin
- How to solve the optimisation problem?
- What are the support vectors?
- What is soft-margin SVM (SVM with slack variables)?
- How to make non-linear SVM?
- What is kernel and what is kernel trick?
- What are pros and cons with SVM?
- What applications are SVM successful for?

History of machine learning

18c	Naive Bayes classifier
1940s	Threshold logic - Warren McCulloch and Walter Pitts
	Logistic regression - Joseph Berkson
1951	k-NN - Evelyn Fix and Joseph Hodges
1957	Perceptron - Frank Rosenblatt
1959	Decision tree - William Belson (?)

History of machine learning

18c	Naive Bayes classifier
1940s	Threshold logic - Warren McCulloch and Walter Pitts
	Logistic regression - Joseph Berkson
1951	k-NN - Evelyn Fix and Joseph Hodges
1957	Perceptron - Frank Rosenblatt
1959	Decision tree - William Belson (?)
1986	ANN with EBP - D.Rumelhart, G.Hinton, and R.Williams

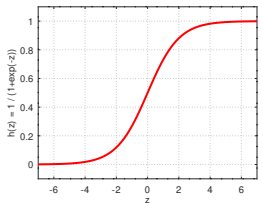
History of machine learning

18c	Naive Bayes classifier
1940s	Threshold logic - Warren McCulloch and Walter Pitts
	Logistic regression - Joseph Berkson
1951	k-NN - Evelyn Fix and Joseph Hodges
1957	Perceptron - Frank Rosenblatt
1959	Decision tree - William Belson (?)
1986	ANN with EBP - D.Rumelhart, G.Hinton, and R.Williams
1993-97	Support Vector Machine - Vladimir Vapnik

Recap – Logistic Regression

- $P(Y=1|\boldsymbol{x}) = \dfrac{1}{1+\exp\left(-(\boldsymbol{w}^T\boldsymbol{x}+w_0)\right)}$

  $\boldsymbol{x} = [x_1 \ \ldots \ x_d]^T$,
  $\boldsymbol{w} = [w_1 \ \ldots \ w_d]^T$, $\quad Y \in \{-1, +1\}$

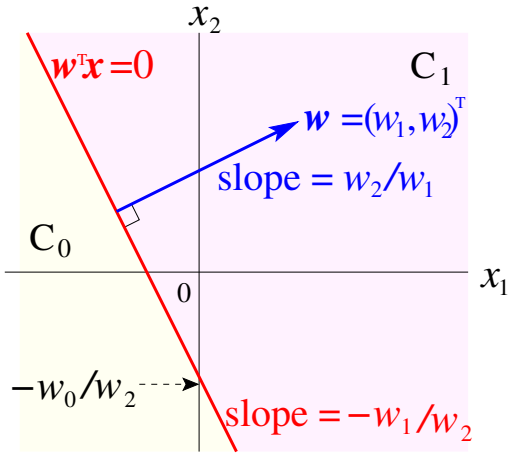


- Training on a data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ based on maximum likelihood estimation (MLE):

$$\max_{\boldsymbol{w}, w_0} \prod_{i=1}^{N} P(Y=y_i|\boldsymbol{x}_i)$$

Decision boundary and decision regions

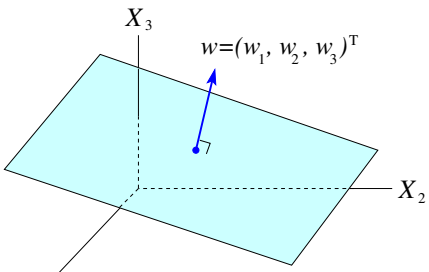
$$P(Y=1|\boldsymbol{x}) = \frac{1}{1+\exp\left(-(\boldsymbol{w}^T\boldsymbol{x}+w_0)\right)} \quad \rightarrow \quad \text{decision boundary: } \boldsymbol{w}^T\boldsymbol{x}+w_0=0$$



Decision boundary and decision regions (*cont.*)

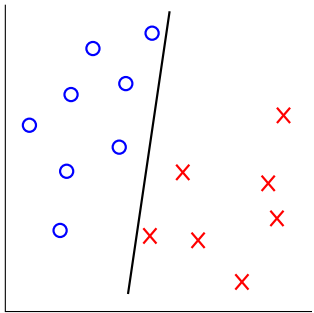
$$P(Y=1|\boldsymbol{x}) = \frac{1}{1+\exp\left(-(\boldsymbol{w}^T\boldsymbol{x}+w_0)\right)}$$

Dimension	Decision boundary	
2	line	$w_1x_1 + w_2x_2 + w_0 = 0$
3	plane	$w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0$
⋮		
d	hyperplane	$\left(\sum_{i=1}^{d} w_ix_i\right) + w_0 = 0$

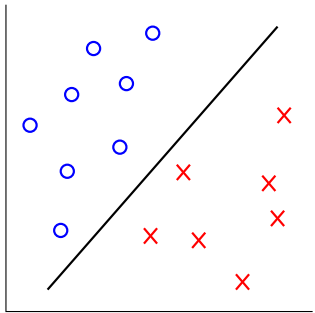


Large margin classifiers

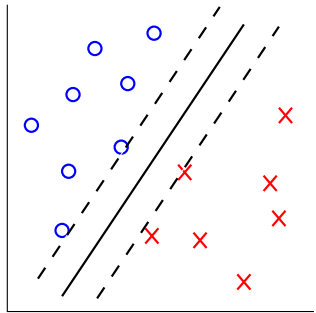
$\hat{y}(\boldsymbol{x}) = f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$



(a)



(b)



(c)

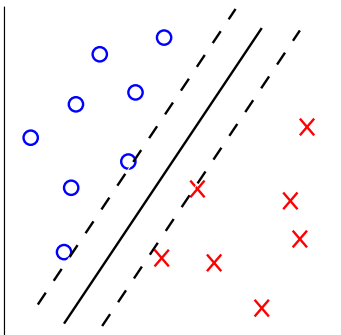
$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 = 0$

$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 = +1$

$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 = -1$

Large margin classifiers (*cont.*)

Proposed by several people, e.g. Vladimir Vapnik (1963, 1992)



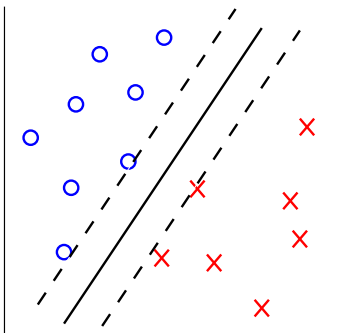
(c)

- Training: maximise the margin with these constraints:

$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq +1 \quad \forall i \text{ s.t. } y_i = +1$$
$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq -1 \quad \forall i \text{ s.t. } y_i = -1$$

Large margin classifiers (*cont.*)

Proposed by several people, e.g. Vladimir Vapnik (1963, 1992)



(c)

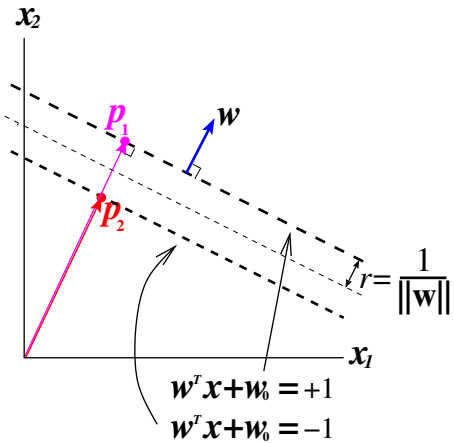
- Training: maximise the margin with these constraints:

$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq +1 \quad \forall i \text{ s.t. } y_i = +1$$
$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq -1 \quad \forall i \text{ s.t. } y_i = -1$$

- Classification based on

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

Margin



$$\begin{aligned}\|\mathbf{p}_1 - \mathbf{p}_2\| &= |\|\mathbf{p}_1\| - \|\mathbf{p}_2\|| \\ &= \left|\frac{-w_0+1}{\|\mathbf{w}\|} - \frac{-w_0-1}{\|\mathbf{w}\|}\right| \\ &= \frac{2}{\|\mathbf{w}\|} = 2r\end{aligned}$$

where

$$\begin{aligned}1 &= \mathbf{w}^T\mathbf{p}_1 + w_0 \\ &= \|\mathbf{w}\|\|\mathbf{p}_1\|\cos(\theta)|_{\theta=0} + w_0 \\ &= \|\mathbf{w}\|\|\mathbf{p}_1\| + w_0 \\ \rightarrow \quad \|\mathbf{p}_1\| &= \frac{-w_0+1}{\|\mathbf{w}\|}\end{aligned}$$

Support Vector Machine (SVM)

Training

$$\max_{\boldsymbol{w}} \quad \frac{1}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \quad \boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq +1 \quad \text{for all } i \text{ with } y_i = +1$$
$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq -1 \quad \text{for all } i \text{ with } y_i = -1$$

Support Vector Machine (SVM)

Training

$$\max_{\boldsymbol{w}} \quad \frac{1}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \quad \boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq +1 \;\text{ for all } i \text{ with } y_i = +1$$
$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq -1 \;\text{ for all } i \text{ with } y_i = -1$$

Equivalent to

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t.} \quad y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + w_0\right) \geq 1 \;\text{ for all } i$$

NB: $\boldsymbol{w}^T \boldsymbol{w} = \|\boldsymbol{w}\|^2$

NB: *constrained, quadratic and convex optimisation problem → no local-minima problem!*

Support Vector Machine (SVM)

Training $\max_{\boldsymbol{w}} \frac{1}{\|\boldsymbol{w}\|}$

s.t. $\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq +1$ for all i with $y_i = +1$
$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq -1$ for all i with $y_i = -1$

Equivalent to

$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2$ NB: $\boldsymbol{w}^T \boldsymbol{w} = \|\boldsymbol{w}\|^2$

s.t. $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + w_0\right) \geq 1$ for all i

NB: *constrained, quadratic and convex optimisation problem → no local-minima problem!*

Solution: $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i, \quad \alpha_i \geq 0 \quad \cdots$ most of α_i are zeros normally

Those $\{\boldsymbol{x}_i\}$ whose $\alpha_i > 0$ are called *support vectors*.

w_0: to be discussed later

Support Vector Machine (SVM)

Training

$$\max_{\boldsymbol{w}} \quad \frac{1}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \quad \boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq +1 \ \text{ for all } i \text{ with } y_i = +1$$
$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq -1 \ \text{ for all } i \text{ with } y_i = -1$$

Equivalent to

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t.} \quad y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + w_0\right) \geq 1 \ \text{ for all } i$$

NB: $\boldsymbol{w}^T \boldsymbol{w} = \|\boldsymbol{w}\|^2$

NB: *constrained, quadratic and convex optimisation problem → no local-minima problem!*

Solution: $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i, \quad \alpha_i \geq 0 \ \cdots$ most of α_i are zeros normally

Those $\{\boldsymbol{x}_i\}$ whose $\alpha_i > 0$ are called *support vectors*.

w_0: to be discussed later

Classification

$$g(\boldsymbol{x}) = \text{sgn}(\boldsymbol{w}^T \boldsymbol{x} + w_0) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i \, y_i \, \boldsymbol{x}_i^T \boldsymbol{x} + w_0\right)$$

Why $+1$ instead of $+\varepsilon$?

Assuming $\varepsilon > 0$,

$$\begin{aligned} \min_{\boldsymbol{w}, w_0} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + w_0\right) \geq \varepsilon \quad \text{for all } i \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} \min_{\boldsymbol{w}, w_o} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i \left(\frac{\boldsymbol{w}^T}{\varepsilon} \boldsymbol{x}_i + \frac{w_0}{\varepsilon}\right) \geq 1 \quad \text{for all } i \end{aligned}$$

Letting $\dot{w} = \frac{\mathbf{w}}{\varepsilon}$ and $\dot{w}_0 = \frac{w_0}{\varepsilon}$,

$$\begin{aligned} \min_{\boldsymbol{w}, w_0} \quad & \frac{\varepsilon^2}{2}\|\dot{\boldsymbol{w}}\|^2 \\ \text{s.t.} \quad & y_i \left(\dot{\boldsymbol{w}}^T \boldsymbol{x}_i + \dot{w}_0\right) \geq 1 \quad \text{for all } i \end{aligned}$$

Optimisation problems in SVM

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}, w_0} & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ \text{s.t.} & y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + w_0\right) \geq 1 \text{ for all } i \end{array}$$

Using the Lagrange multipliers $\alpha_i \geq 0$, the Lagrangian is given as:

$$L(\boldsymbol{\alpha}, \dot{\mathbf{w}}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{N} \alpha_i \left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1\right)$$

where $\boldsymbol{\alpha} = [\alpha_1 \ \ldots \ \alpha_n]$ and $\dot{\mathbf{w}} = [\mathbf{w}\ w_0]$.

Optimisation problems in SVM (*cont.*)

$$L(\boldsymbol{\alpha}, \dot{\mathbf{w}}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{N} \alpha_i \left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 \right)$$

$$\frac{\partial L(\boldsymbol{\alpha}, \dot{\mathbf{w}})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i \, y_i \, \boldsymbol{x}_i = \mathbf{0},$$
$$\frac{\partial L(\boldsymbol{\alpha}, \dot{\mathbf{w}})}{\partial w_0} = -\sum_{i=1}^{N} \alpha_i \, y_i = 0.$$

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

Optimisation problems in SVM (*cont.*)

Putting the results to the Lagrangian yields:

$$\begin{aligned} L(\boldsymbol{\alpha}, \dot{\mathbf{w}}) &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{N}\alpha_i \left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{N} y_i\, y_j\, \alpha_i\, \alpha_j\, \mathbf{x}_i^T\mathbf{x}_j - \sum_{i,j=1}^{N} y_i\, y_j\, \alpha_i\, \alpha_j\, \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i \\ &= -\frac{1}{2}\sum_{i,j=1}^{N} y_i\, y_j\, \alpha_i\, \alpha_j\, \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i \end{aligned}$$

The necessary and sufficient conditions for $\mathbf{w}^*$ to be an optimum are:

$$\frac{\partial L(\boldsymbol{\alpha}^*, \dot{\mathbf{w}}^*)}{\partial \mathbf{w}} = \mathbf{0}, \quad \frac{\partial L(\boldsymbol{\alpha}^*, \dot{\mathbf{w}}^*)}{\partial w_0} = 0, \quad \alpha_i^* \geq 0, \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 \geq 0,$$

$$\alpha_i^* \left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1\right) = 0, \quad \text{for all } i \quad \cdots \text{ Karush-Kuhn-Tuckert (KKT) conditions}$$

which means that either $\alpha_i^* = 0$ or $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 = 0$.

Optimisation problems in SVM (*cont.*)

$\alpha_i^* = 0$ or $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 = 0.$

Optimisation problems in SVM (*cont.*)

$\alpha_i^* = 0$ or $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 = 0$.

What does $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 = 0$ mean?

Optimisation problems in SVM (*cont.*)

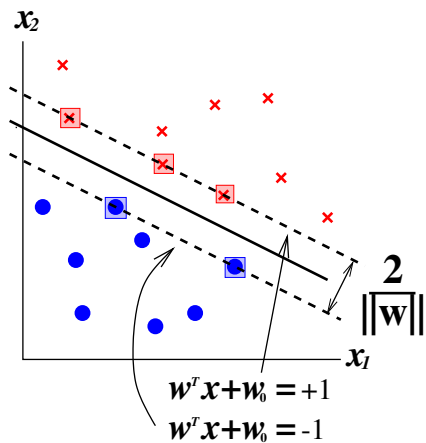
$\alpha_i^* = 0$ or $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 = 0$.

What does $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 = 0$ mean?

$\boldsymbol{x}_i$ with $\alpha_i^* \neq 0$ should satisfy:

$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 = +1 \quad \text{if } y_i = +1$$
$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 = -1 \quad \text{if } y_i = -1$$

See "complementary slackness" in the last lecture



How to find w_0?

Support vectors satisfy :

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + w_0) = 1 \tag{1}$$

$$y_i \left(\sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x}_j^\top \boldsymbol{x}_i + w_0 \right) = 1 \tag{2}$$

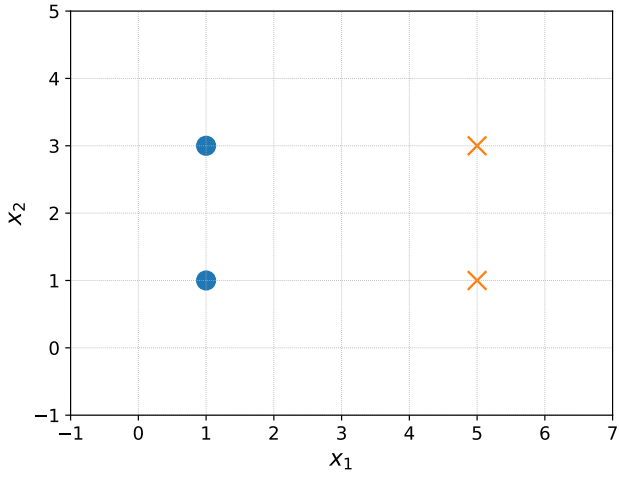
So,

$$w_0 = y_i - \boldsymbol{w}^\top \boldsymbol{x}_i \tag{3}$$

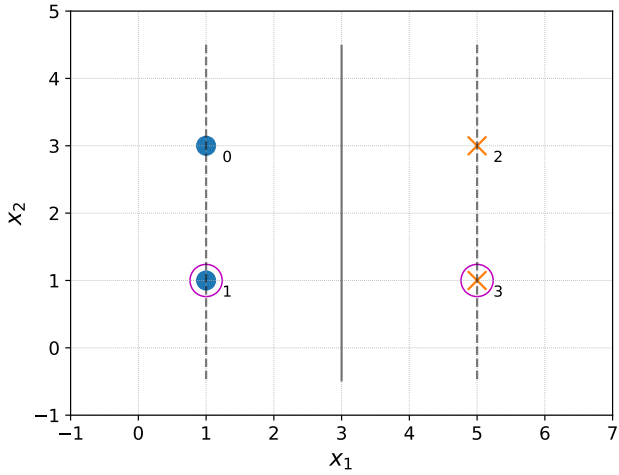
$$= y_i - \sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x}_j^\top \boldsymbol{x}_i \tag{4}$$

What if you get different w_0 for different support vectors?
See MML 12.3.1

Example 1



Example 1



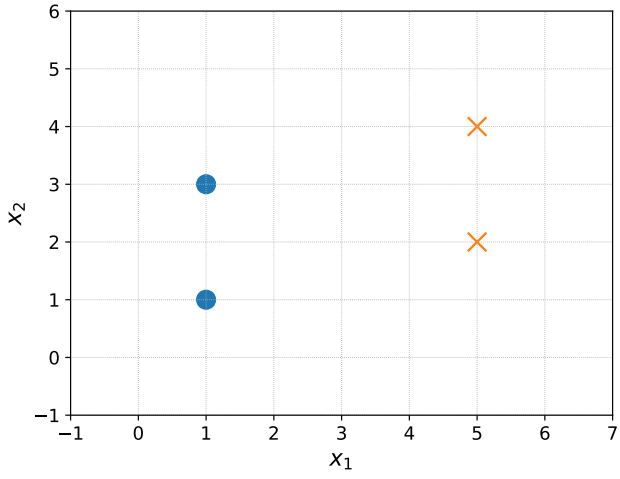
$\alpha_0 = 0.0$
$\alpha_1 = 0.125$
$\alpha_2 = 0.0$
$\alpha_3 = 0.125$

$\boldsymbol{w} = [0.5 \;\; 0.0]$

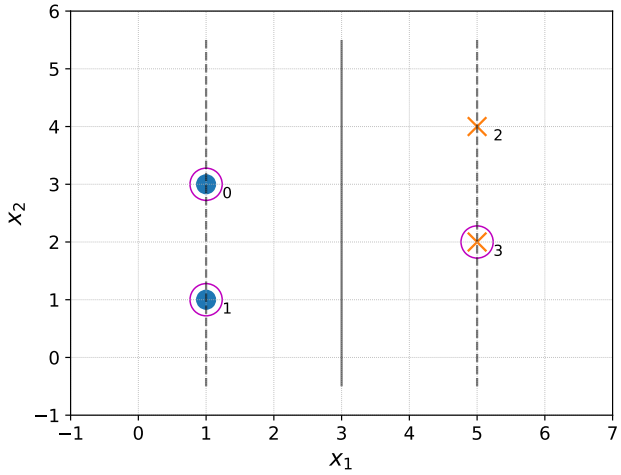
$\|\boldsymbol{w}\| = 0.5$

$\rightarrow \;\; r = 2$

Example 2



Example 2



$\alpha_0 = 0.0625$
$\alpha_1 = 0.0625$
$\alpha_2 = 0.0$
$\alpha_3 = 0.125$

$\boldsymbol{w} = [0.5 \;\; 0.0]$

$\|\boldsymbol{w}\| = 0.5$

$\rightarrow \;\; r = 2$

Example 3

- What if a dataset is linearly non-separable?

Example 3

- What if a dataset is linearly non-separable?
- In that case, does the optimisation problem have a solution?

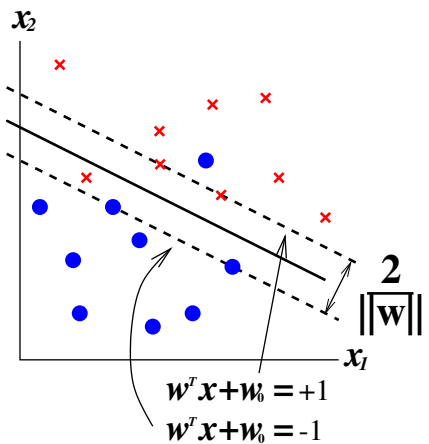
SVM with slack variables – soft margin SVM

Hard margin SVM

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 \text{ for all } i$$

Soft margin SVM

$$\min_{\boldsymbol{w},w_0} \quad \boldsymbol{w}^T\boldsymbol{w} + C\left(\sum_{i=1}^{N}\xi_i\right), \quad \text{where } C > 0,$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i \text{ for all } i,\ \xi_i \geq 0$$



SVM with slack variables – soft margin SVM

Hard margin SVM

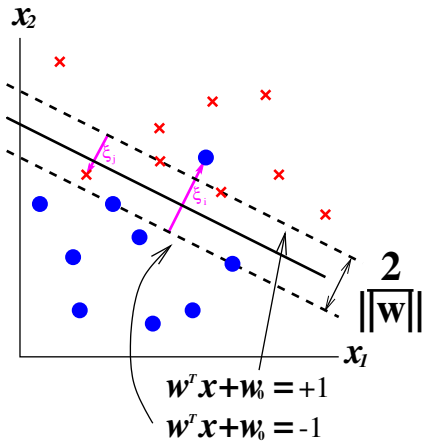
$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 \ \text{ for all } i$$

Soft margin SVM

$$\min_{\boldsymbol{w}, w_0} \quad \boldsymbol{w}^T\boldsymbol{w} + C\left(\sum_{i=1}^{N}\xi_i\right), \quad \text{where } C > 0,$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i \ \text{ for all } i, \ \xi_i \geq 0$$



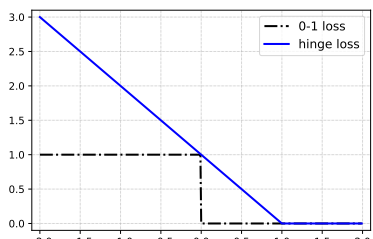
Loss function in soft-margin SVM

$$\begin{aligned} \min_{\boldsymbol{w}, w_0} \quad & \boldsymbol{w}^T \boldsymbol{w} + C \left(\sum_{i=1}^{N} \xi_i \right), \quad \text{where } C > 0, \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 1 - \xi_i \ \text{ for all } i, \ \xi_i \geq 0 \end{aligned}$$

The hinge loss:

$$\begin{aligned} \ell(t) &= \max(0, 1-t) \\ &= \begin{cases} 0, & \text{if } t \geq 1, \\ 1-t, & \text{otherwise,} \end{cases} \end{aligned}$$

where $t = y(\mathbf{w}^T \boldsymbol{x} + w_0)$.



Optimisation problem of soft-margin SVM

$$\begin{aligned} \min_{\boldsymbol{w}, w_0} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\left(\sum_{i=1}^{N}\xi_i\right), \quad \text{where } C > 0, \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i \text{ for all } i,\ \xi_i \geq 0 \end{aligned}$$

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - (1-\xi_i)\right) - \sum_{i=1}^{N}\beta_i\xi_i$$

Optimisation problem of soft-margin SVM

$$\begin{aligned} \min_{\boldsymbol{w}, w_0} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\left(\sum_{i=1}^{N}\xi_i\right), \quad \text{where } C > 0, \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i \text{ for all } i,\ \xi_i \geq 0 \end{aligned}$$

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - (1-\xi_i)\right) - \sum_{i=1}^{N}\beta_i\xi_i$$

$$= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \boldsymbol{w}^T\sum_{i=1}^{N}\alpha_i\, y_i\, \boldsymbol{x}_i - w_0\sum_{i=1}^{N}\alpha_i\, y_i + \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}(C - \alpha_i - \beta_i)\xi_i$$

Optimisation problem of soft-margin SVM (*cont.*)

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \boldsymbol{w}^T \sum_{i=1}^{N} \alpha_i\, y_i\, \boldsymbol{x}_i - w_0 \sum_{i=1}^{N} \alpha_i\, y_i + \sum_{i=1}^{N} \alpha_i + \sum_{i=1}^{N} (C - \alpha_i - \beta_i)\xi_i$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i\, y_i\, \boldsymbol{x}_i = \mathbf{0}, \tag{5}$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi})}{\partial w_0} = -\sum_{i=1}^{N} \alpha_i\, y_i = 0\,, \tag{6}$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi})}{\partial \xi_i} = C - \alpha_i - \beta_i = 0\,. \tag{7}$$

$\rightarrow$

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i\, y_i\, \boldsymbol{x}_i, \quad \sum_{i=1}^{N} \alpha_i\, y_i = 0, \quad C - \alpha_i - \beta_i = 0 \quad \forall i = 1, \ldots, N \tag{8}$$

Optimisation problem of soft-margin SVM (*cont.*)

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \boldsymbol{w}^T \sum_{i=1}^{N} \alpha_i\, y_i\, \boldsymbol{x}_i - w_0 \sum_{i=1}^{N} \alpha_i\, y_i + \sum_{i=1}^{N} \alpha_i + \sum_{i=1}^{N} (C - \alpha_i - \beta_i)\xi_i$$

$$= -\frac{1}{2}\sum_{i,j=1}^{N} y_i\, y_j\, \alpha_i\, \alpha_j\, \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{N} \alpha_i \qquad (9)$$

Dual problem:

$$\max_{\boldsymbol{\alpha}} \quad -\frac{1}{2}\sum_{i,j=1}^{N} y_i\, y_j\, \alpha_i\, \alpha_j\, \mathbf{x}_i^T \mathbf{x}_j + \sum_{i}^{N} \alpha_i$$

$$\text{s.t.} \quad \sum_{i}^{N} y_i\, \alpha_i = 0 \quad \text{and} \quad 0 \le \alpha_i \le C \quad \forall i = 1, \ldots, N$$

NB: $\alpha_i \le C$, because $C - \alpha_i - \beta_i = 0$ and $\alpha_i \ge 0,\ \beta_i \ge 0$.

$\{\beta_i\}$ do not appear!

Optimisation problem of soft-margin SVM (*cont.*)

The complementary slackness in the KKT conditions:

$$\alpha_i^* \left(y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \right) - \left(1 - \xi_i^* \right) \right) = 0 \quad \textit{and} \quad \beta_i^* \, \xi_i^* = 0 \quad \textit{for all } i = 1, \ldots, N$$

This means:

$$\alpha_i^* = 0 \quad \textit{or} \quad y_i (\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - (1 - \xi_i^*) = 0 \ \textit{and} \ 0 < \alpha_i^* \leq C$$

$$\textit{If} \ \alpha_i^* = C, \ \textit{then} \ \beta_i^* = 0, \ \textit{as} \ C - \alpha_i^* - \beta_i^* = 0.$$

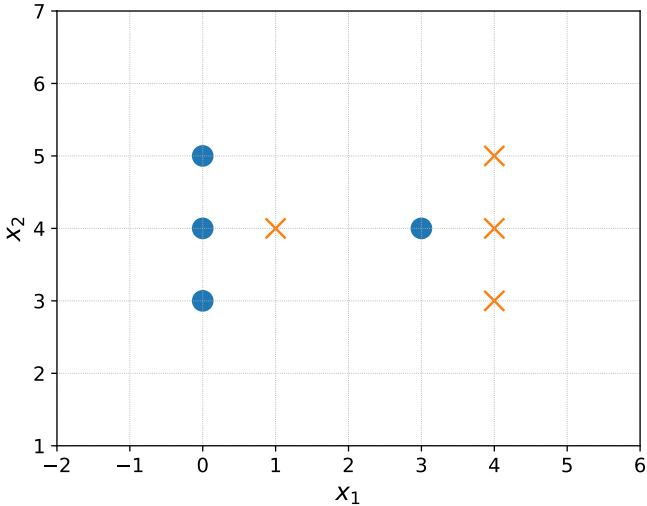
and

$$\beta_i^* = 0 \ \textit{if} \ \xi_i^* > 0 \quad \textit{or} \quad \xi_i^* = 0 \ \textit{and} \ 0 < \beta_i^* \leq C.$$

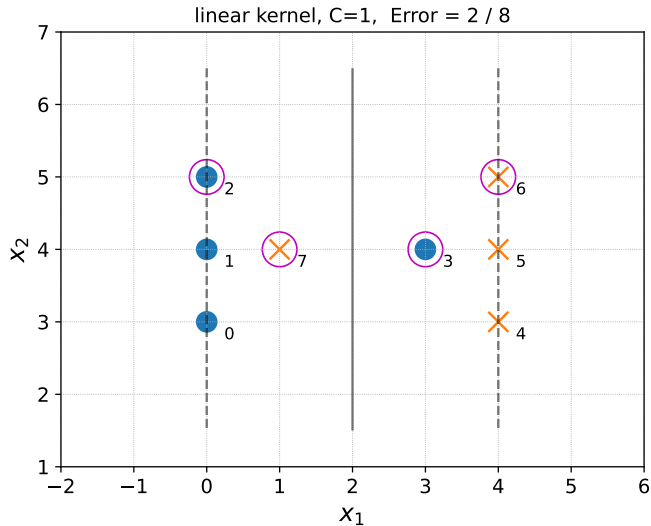
So,

$$\begin{array}{rcll} \alpha_i^* = 0 & : & \xi_i^* = 0, & \boldsymbol{x}_i \textit{ is not a support vector} \\ 0 < \alpha_i^* < C & : & \xi_i^* = 0, & \boldsymbol{x}_i \textit{ is a support vector on the corresponding hyperplane} \\ \alpha_i^* = C & : & \xi_i^* > 0, & \boldsymbol{x}_i \textit{ is a support vector not on the corresponding hyperplane} \end{array}$$

Soft-margin SVM – Example when $C = 1$



Soft-margin SVM – Example when $C = 1$



$\alpha_0 = 0.0$
$\alpha_1 = 0.0$
$\alpha_2 = 0.625$
$\alpha_3 = 1.0$
$\alpha_4 = 0.0$
$\alpha_5 = 0.0$
$\alpha_6 = 0.625$
$\alpha_7 = 1.0$

$\boldsymbol{w} = [0.5 \;\; 0.0], \; w_0 = -1$

$\|\boldsymbol{w}\| = 0.5$

$\rightarrow \; r = 2$

Generalisation error of SVM *(NE)*

Assuming the class $\mathcal{F}$ of real-valued functions on the ball of radius R in $\mathbb{R}^n$ as

$$\mathcal{F} = \{\boldsymbol{x} \mapsto \mathbf{w} \cdot \boldsymbol{x} : \|\mathbf{w}\| \leq 1, \|\boldsymbol{x}\| \leq R\}.$$

If a classifier $\operatorname{sgn}(f) \in \operatorname{sgn}(\mathcal{F})$ has margin at least γ on all the training examples, with probability at least $1 - \delta$ over n random examples, f has error no more than

$$L_D(f) \leq \frac{k}{N} + \sqrt{\frac{c}{N}\left(\frac{R^2}{\gamma^2}\log^2 n + \log\left(\frac{1}{\delta}\right)\right)}$$

where k is the number of labelled training examples with margin less than γ, c is a constant,

$$\text{VC-dim}(f) \leq \min(\frac{R^2}{\gamma^2}, n) + 1$$

Experiments on US Postal Service Database

C. Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

US Postal Service Database (handwritten digits):

Training samples	*7300*
Test samples	*2000*
Image resolution	16×16 *pixels*

Classifier	Err. [%]
Human performance	2.5
Decision tree, CART	17.0
Decision tree, V4.5	16.0
Best 2 layer NN	6.6
LeNet1 (5 layers)	5.1

d	Err. [%]	Support vectors	Dimensionality of feature space
1	12.0	200	256
2	4.7	127	~ 33000
3	4.4	148	$\sim 1 \times 10^6$
4	4.3	165	$\sim 1 \times 10^9$
5	4.3	175	$\sim 1 \times 10^{12}$
6	4.2	185	$\sim 1 \times 10^{14}$
7	4.3	190	$\sim 1 \times 10^{16}$

d: degree of polynomial kernel

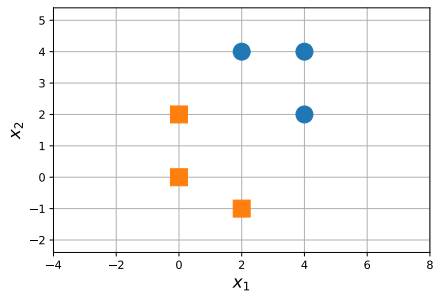
Some notes on SVMs

- How to choose the regulariser C? $\cdots$ use a validation set
- How to solve the constrained quadratic optimisation problem in SVM practically? It requires a kernel matrix of N-by-N.
  - Gradient, sub-gradient, coordinate ascent/descent
  - Sequential Minimal Optimisation (SMO) [John Platt, 1998]
  - LIBSVM [Chih-Chung Chang and Chih-Jen Lin]: an SVM software tool with SMO
- How to apply SVMs to multi-class classification problems?
- Performance deterioration (NB: not very specific to SVMs)
  - Heavily-overlapped data sets
  - Imbalanced data sets
  - (Too many support vectors)
  - (Large data sets)
- Output interpretability

Quizzes

Consider an SVM on the following data set.

x_1	x_2	y
2.0	4.0	1
4.0	2.0	1
4.0	4.0	1
0.0	2.0	2
2.0	-1.0	2
0.0	0.0	2



1. Using your intuition, what weight vector do you think will result from training an SVM on this data set?
2. Plot the data and the decision boundary of the weight vector you have chosen.
3. Which are the support vectors? What is the margin of this classifier?