Support Vector Machines

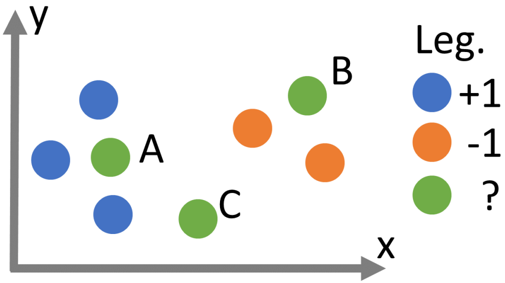
Machine Learning - ENS 2022

Thomas Romary

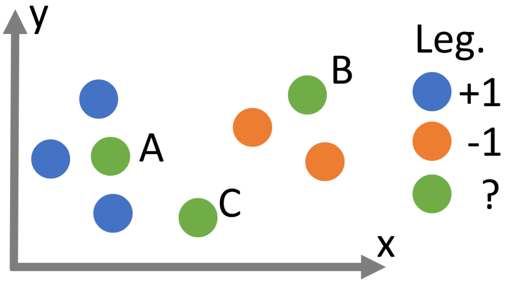
Mines Paris
(from a course of Tim Schlottmann, Hendrik Sieck, Jonas Kru)

10.26.2022

Motivation



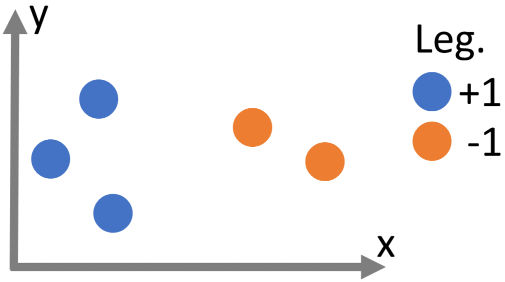
- Object classification
- Simple and efficient
- As accurate as possible



- Support Vector Machine
- Binary classification

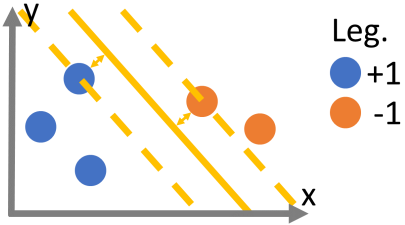
Contents

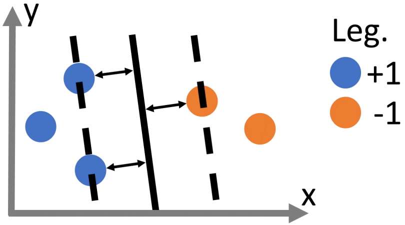
Problem



- How do I separate the two classes?

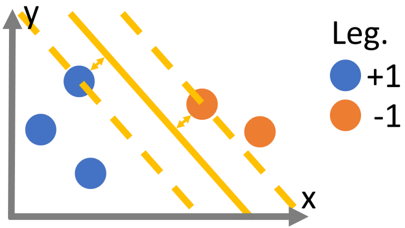
Problem

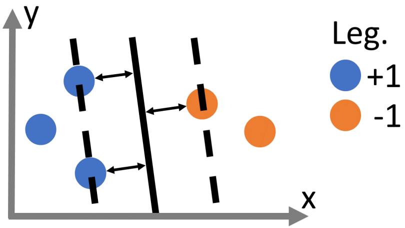




- What is the best way to set the hyperplane?
- Other name of Support Vector Machines: Large Margin Classifier

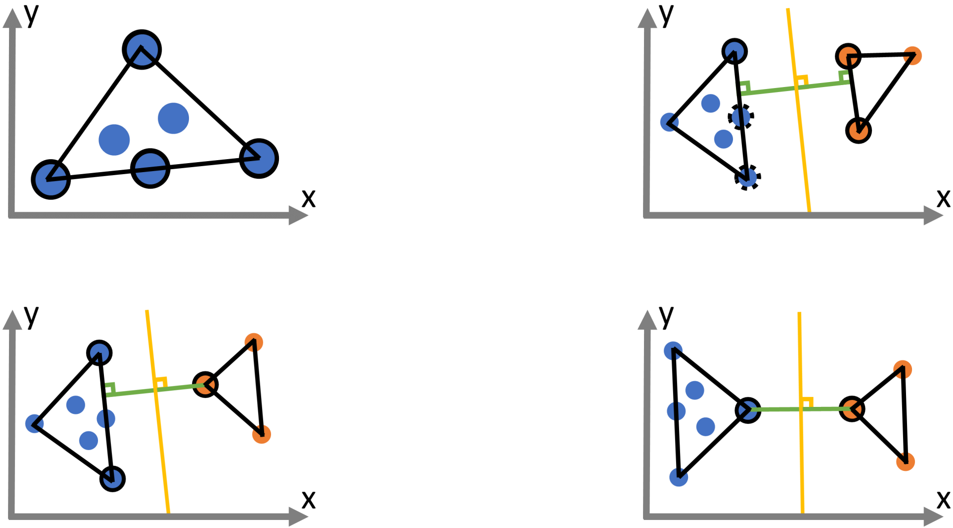
Problem





- What is the best way to set the hyperplane?
- Other name of Support Vector Machines: Large Margin Classifier

Support vectors (s.v.)

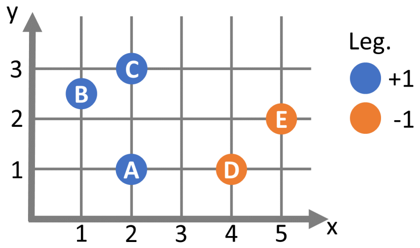


Definition

- $m \in \mathbb{R}$ data points
- Input $\boldsymbol{x} \in \mathbb{R}^N$
- Output $y \in \{-1, +1\}$
- Training set $S \in (\mathbb{R}^N \times \{-1, +1\})^m$

- Hypothesis

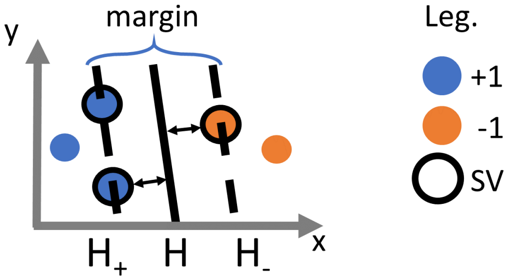
$$h : \mathbb{R}^N \longrightarrow \{+1, -1\}$$
$$\boldsymbol{x} \longmapsto y$$



- $m = 5$
- Training set S:

$$S = \begin{pmatrix} 1 & 2.5 & +1 \\ 5 & 2 & -1 \\ & \vdots & \end{pmatrix}$$

Hyperplane



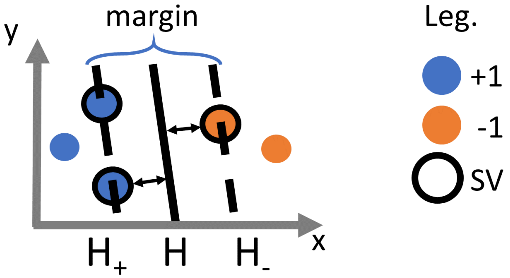
- Hyperplane: $H = \boldsymbol{w}^T \boldsymbol{x} + b = 0$
- Hyperplanes H_+ and H_-:

$$H_+ := \boldsymbol{w}^T \boldsymbol{x}_p + b = +1, \quad \forall \text{ s.v., lying on } H_+$$

$$H_- := \boldsymbol{w}^T \boldsymbol{x}_n + b = -1, \quad \forall \text{ s.v., lying on } H_-$$

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1 \quad \forall \text{ s.v.}$$

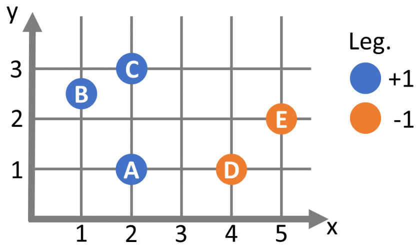
Hyperplane



- Hyperplane: $H = \boldsymbol{w}^T\boldsymbol{x} + b = 0$
- Classification:

$$h(x_i) = \begin{cases} +1 & \text{when } \boldsymbol{w}^T\boldsymbol{x}_i + b \geq 0 \\ -1 & \text{when } \boldsymbol{w}^T\boldsymbol{x}_i + b \leq 0 \end{cases}$$

Example I



- Hyperplane: $H = \boldsymbol{w}^T \boldsymbol{x} + b = 0$
- Constraints: $\boldsymbol{w}^T \boldsymbol{x}_i + b = y_i, \quad \forall$ s. v.

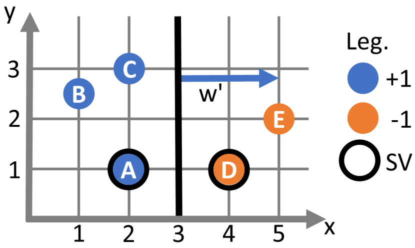
- Graphical determination of the hyperplane parameters:

$$x = 3$$

$$\boldsymbol{w} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 0 \end{pmatrix}^T \begin{pmatrix} 3 \\ 0 \end{pmatrix} + b = 0 \Rightarrow b = -6$$

Example I



- Hyperplane: $H = \boldsymbol{w}^T \boldsymbol{x} + b = 0$
- Constraints: $\boldsymbol{w}^T \boldsymbol{x}_i + b = y_i, \quad \forall$ s. v.

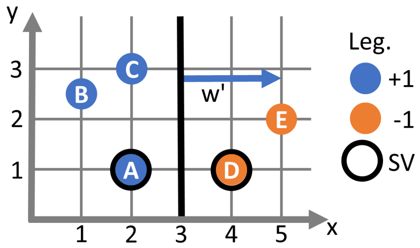
- Graphical determination of the hyperplane parameters:

$$x = 3$$

$$\boldsymbol{w} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 0 \end{pmatrix}^T \begin{pmatrix} 3 \\ 0 \end{pmatrix} + b = 0 \Rightarrow b = -6$$

Example II



- Hyperplane: $H = \boldsymbol{w}^T\boldsymbol{x} + b = 0$
- Constraints: $\boldsymbol{w}^T\boldsymbol{x}_i + b = y_i, \quad \forall$ s. v.

- Consider the constraint using the example of point A:

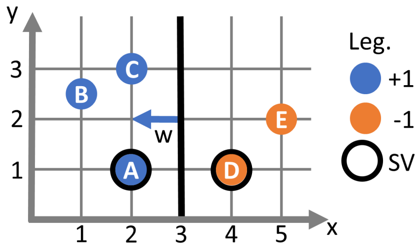
$$c\left(\begin{pmatrix}2 & 0\end{pmatrix}^T\begin{pmatrix}2\\1\end{pmatrix} - 6\right) \overset{!}{=} +1 \Rightarrow c = -0.5$$

- Thus for the canonical hyperplane:

$$\boldsymbol{w} = \begin{pmatrix}-1\\0\end{pmatrix} \qquad b = 3$$

- Control with point D:

Example II



- Hyperplane: $H = \boldsymbol{w}^T \boldsymbol{x} + b = 0$
- Constraints: $\boldsymbol{w}^T \boldsymbol{x}_i + b = y_i, \quad \forall$ s. v.

- Consider the constraint using the example of point A:

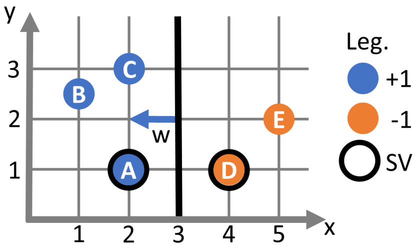
$$c\left(\begin{pmatrix}2 & 0\end{pmatrix}^T \begin{pmatrix}2\\1\end{pmatrix} - 6\right) \stackrel{!}{=} +1 \Rightarrow c = -0.5$$

- Thus for the canonical hyperplane:

$$\boldsymbol{w} = \begin{pmatrix}-1\\0\end{pmatrix} \quad b = 3$$

- Control with point D:

Example II



- Hyperplane: $H = \boldsymbol{w}^T \boldsymbol{x} + b = 0$
- Constraints: $\boldsymbol{w}^T \boldsymbol{x}_i + b = y_i, \quad \forall$ s. v.

- Consider the constraint using the example of point A:

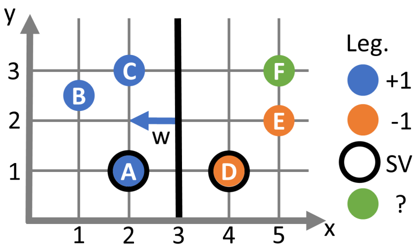
$$c \left(\begin{pmatrix} 2 & 0 \end{pmatrix}^T \begin{pmatrix} 2 \\ 1 \end{pmatrix} - 6 \right) \stackrel{!}{=} +1 \Rightarrow c = -0.5$$

- Thus for the canonical hyperplane:

$$\boldsymbol{w} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \quad b = 3$$

- Control with point D:

Example III



- Classification:

$$h(x_i) = \begin{cases} +1 & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i + b \geq 0 \\ -1 & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i + b \leq 0 \end{cases}$$

- Parameters of the canonical hyperplane

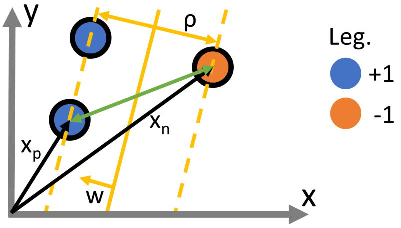
$$\boldsymbol{w} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \quad b = 3$$

- Classification of point F:

$$\begin{pmatrix} -1 & 0 \end{pmatrix}^T \begin{pmatrix} 5 \\ 3 \end{pmatrix} + 3 = -2$$

- So point F belongs to the class -1.

Minimization problem



- Projection property of the scalar product
- Width of the margin ρ:

$$\rho = (\boldsymbol{x}_p - \boldsymbol{x}_n)^T \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \Leftrightarrow \rho = (\boldsymbol{x}_p^T \boldsymbol{w} - \boldsymbol{x}_n^T \boldsymbol{w}) \frac{1}{\|\boldsymbol{w}\|}$$

- Constraint: $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1 \quad \forall$ s.v.
- $\rho = \frac{2}{\|\boldsymbol{w}\|}$
- Goal of an SVM: Maximize the margin

$$\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \Leftrightarrow \min_{\boldsymbol{w},b} \|\boldsymbol{w}\| \Leftrightarrow \min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|^2$$

with the constraint $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 \geq 0$

Dual problem: Lagrange Multipliers

- Lagrange Multipliers:

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum \alpha_i \left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1\right]$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum \alpha_i y_i \boldsymbol{x}_i = 0 \qquad \Rightarrow \boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0 \qquad \Rightarrow \sum_i \alpha_i y_i = 0$$

- Replace $\boldsymbol{w}$ in L:

$$L = \frac{1}{2}(\sum_i \alpha_i y_i \boldsymbol{x}_i)^T(\sum_j \alpha_j y_j \boldsymbol{x}_j) - (\sum_i \alpha_i y_i \boldsymbol{x}_i)^T(\sum_j \alpha_j y_j \boldsymbol{x}_j) - \sum_i \alpha_i y_i b + \sum_i \alpha_i$$

$$L = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

- Target: $\max_{\boldsymbol{\alpha}} L$
- Decision function:

$$h(\boldsymbol{x}) = \begin{cases} +1 & \sum_i \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b \geq 0 \\ -1 & \sum_i \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b \leq 0 \end{cases}$$

Support variables α

- Characteristics:

$$\boldsymbol{\alpha} \geq 0$$

$$\alpha_i \begin{cases} > 0 & \text{when } x_i \text{ is a support vector} \\ = 0 & \text{otherwise} \end{cases}$$

$$\sum_{\substack{\text{pos. s.v.} \\ p}} \alpha_p = \sum_{\substack{\text{neg. s.v.} \\ n}} \alpha_n$$

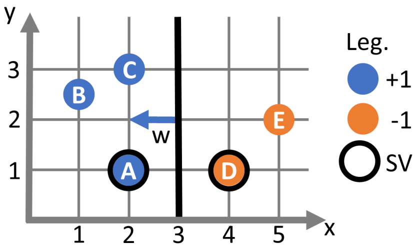
Support variables α

- Characteristics:

$$\boldsymbol{\alpha} \geq 0$$

$$\alpha_i \begin{cases} > 0 & \text{when } x_i \text{ is a support vector} \\ = 0 & \text{otherwise} \end{cases}$$

$$\sum_{\substack{\text{pos. s.v.} \\ p}} \alpha_p = \sum_{\substack{\text{neg. s.v.} \\ n}} \alpha_n$$



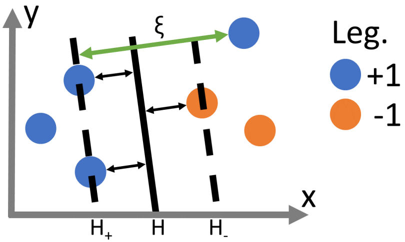
- Computation of α_A and α_D results:

$$\alpha_A = 0.5$$
$$\alpha_D = 0.5$$

- Computation of α_C result:

$$\alpha_C = 0$$

Compensation for a faulty training set



- Relaxation variable ξ to compensate for false classification
- It tolerates some outliers in the classification

- Constraints:

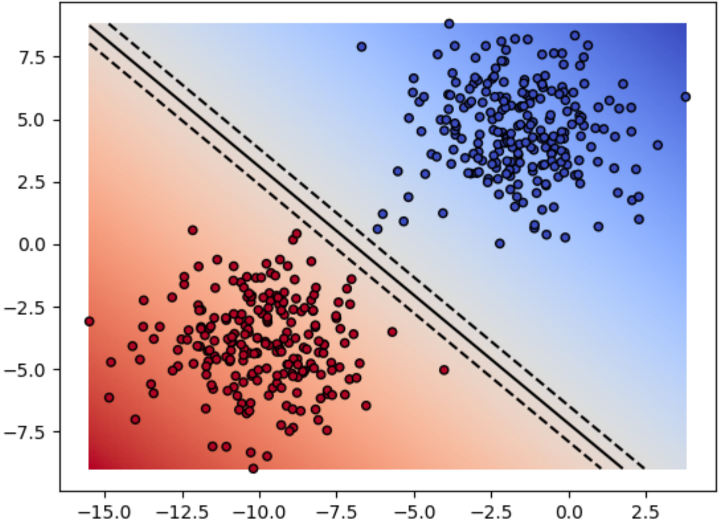
$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall \boldsymbol{x}_i \in S \qquad \xi_i \geq 0$$

- Minimization problem:

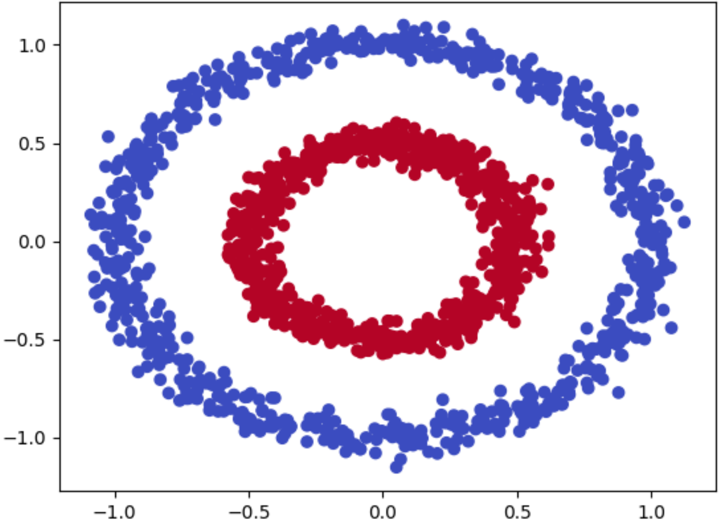
$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i$$

with the constraint $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$

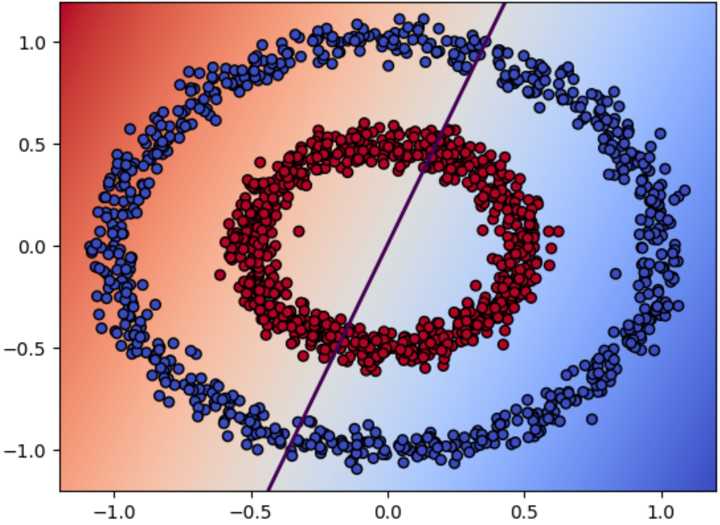
Introduction: Linear separable data



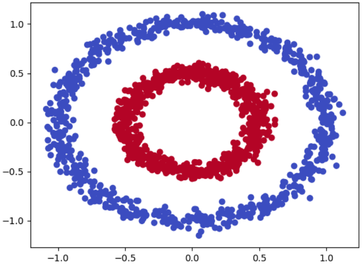
Introduction: Linearly non-separable data 1



Introduction: Linearly non-separable data 2

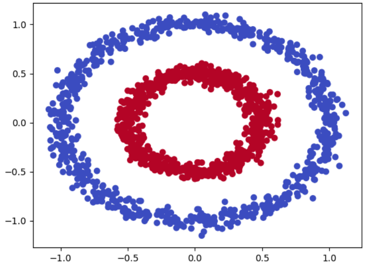


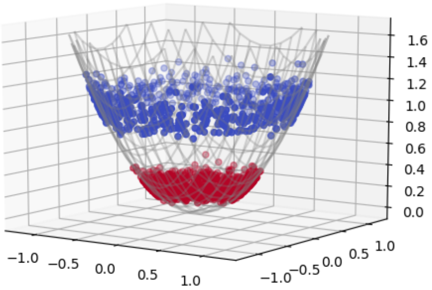
Approach: Introduction



$$\phi(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \\ {x_1}^2 + {x_2}^2 \end{bmatrix}$$

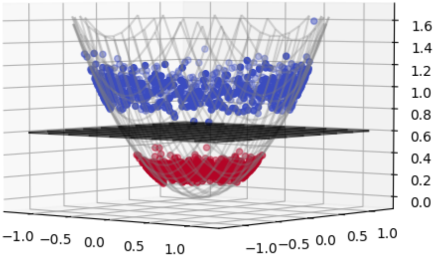
Approach: Introduction

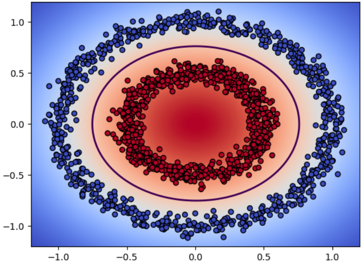




$$\phi(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \\ {x_1}^2 + {x_2}^2 \end{bmatrix}$$

Approach: image function $\phi(x)$





Change in dual problem

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$$

$$h(\boldsymbol{x}) = \sum_i \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b$$

$$h(\boldsymbol{x}) = \sum_i \alpha_i y_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}) + b$$

Change in dual problem

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

$$h(\mathbf{x}) = \sum_i \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$$

$$h(\mathbf{x}) = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + b$$

Definition: ϕ

Image function

$$\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$
$$\boldsymbol{x} \longmapsto \boldsymbol{f}$$

After transformation, the data are linearly separable
Problem:

- $m > n$ high computational effort
  Above a certain size, it's difficult to use
- Moreover, note that ϕ is only needed for scalar product

Definition: ϕ

Image function

$$\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$
$$\boldsymbol{x} \longmapsto \boldsymbol{f}$$

After transformation, the data are linearly separable
Problem:

- $m > n$ high computational effort
  Above a certain size, it's difficult to use
- Moreover, note that ϕ is only needed for scalar product

Kernel example

$$\phi(\boldsymbol{x}) = (1 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2 \;\; \dots \;\; x_1^2 \;\; x_2^2 \;\; \dots \;\; \sqrt{2}x_1x_2 \;\; \sqrt{2}x_1x_3 \;\; \dots)$$

$$\begin{aligned}
\phi(\boldsymbol{v})^T\phi(\boldsymbol{w}) &= \sum_j 2v_jw_j + \sum_j v_j^2w_j^2 + \sum_j \sum_{k>j} 2v_jv_kw_jw_k + \dots \\
&= (1 + \sum_j v_jw_j)^2 \\
&= (1 + \boldsymbol{v}^T\boldsymbol{w})^2 \\
&= K(\boldsymbol{v}, \boldsymbol{w})
\end{aligned}$$

Introduction of kernels

$$K(\boldsymbol{v}, \boldsymbol{w}) = \phi(\boldsymbol{v})^T \phi(\boldsymbol{w})$$

$$K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$$

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$h(\boldsymbol{x}) = \sum_i \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

Introduction of kernels

$$K(\boldsymbol{v}, \boldsymbol{w}) = \phi(\boldsymbol{v})^T \phi(\boldsymbol{w})$$

$$K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$$

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$h(\boldsymbol{x}) = \sum_i \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

Mercer's Theorem: 1

$\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$
$\mathcal{K}_{i,j} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$

$$
\begin{aligned}
\mathcal{K}_{i,j} &= K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \\
&= \phi(\boldsymbol{x}^{(i)})^T \phi(\boldsymbol{x}^{(j)}) \\
&= \phi(\boldsymbol{x}^{(j)})^T \phi(\boldsymbol{x}^{(i)}) \\
&= K(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(i)}) \\
&= \mathcal{K}_{j,i}
\end{aligned}
$$

Mercer's Theorem: 1

$\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$
$\mathcal{K}_{i,j} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$

$$\begin{aligned}
\mathcal{K}_{i,j} &= K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \\
&= \phi(\boldsymbol{x}^{(i)})^T \phi(\boldsymbol{x}^{(j)}) \\
&= \phi(\boldsymbol{x}^{(j)})^T \phi(\boldsymbol{x}^{(i)}) \\
&= K(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(i)}) \\
&= \mathcal{K}_{j,i}
\end{aligned}$$

Mercer's Theorem: 2

$\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$
$\mathcal{K}_{i,j} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$

Choose any $\boldsymbol{z}$:

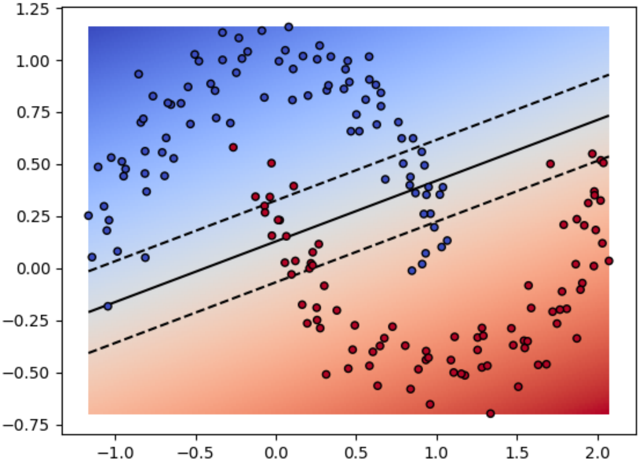
$$
\begin{aligned}
\boldsymbol{z}^T \mathcal{K} \boldsymbol{z} &= \sum_i \sum_j \boldsymbol{z}_i \mathcal{K}_{i,j} \boldsymbol{z}_j \\
&= \sum_i \sum_j \boldsymbol{z}_i \phi(\boldsymbol{x}^{(i)})^T \phi(\boldsymbol{x}^{(j)}) \boldsymbol{z}_j \\
&= \sum_i \sum_j \boldsymbol{z}_i \sum_k \phi_k(\boldsymbol{x}^{(i)}) \phi_k(\boldsymbol{x}^{(j)}) \boldsymbol{z}_j \\
&= \sum_k \sum_i \sum_j \boldsymbol{z}_i \phi_k(\boldsymbol{x}^{(i)}) \phi_k(\boldsymbol{x}^{(j)}) \boldsymbol{z}_j \\
&= \sum_k \left(\sum_i \boldsymbol{z}_i \phi_k(\boldsymbol{x}^{(i)}) \right)^2 \\
&\geq 0
\end{aligned}
$$

Different kernels in practice

- Linear Kernel
- Polynomial Kernel
- Gaussian Kernel

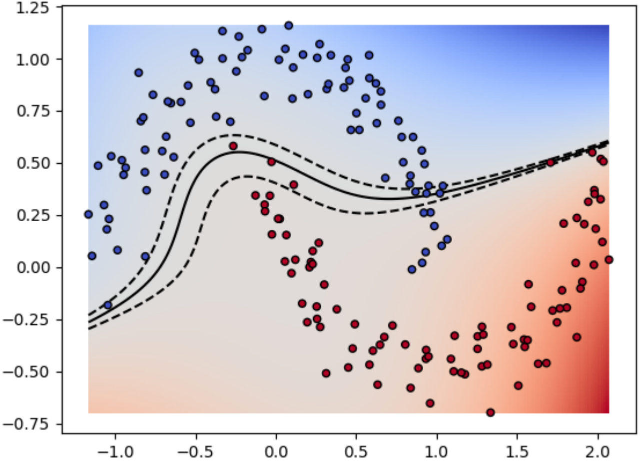
Different kernels in practice: Linear

$$K(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{v}^T \boldsymbol{w}$$



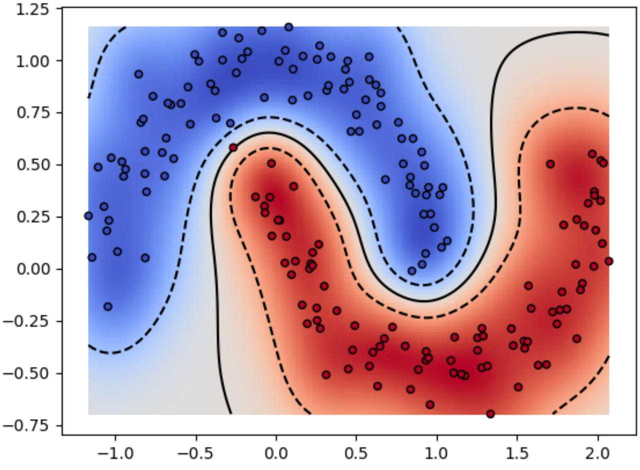
Different kernels in practice: Polynomial

$$K(\boldsymbol{v}, \boldsymbol{w}) = (\boldsymbol{v}^T \boldsymbol{w} + c)^d$$



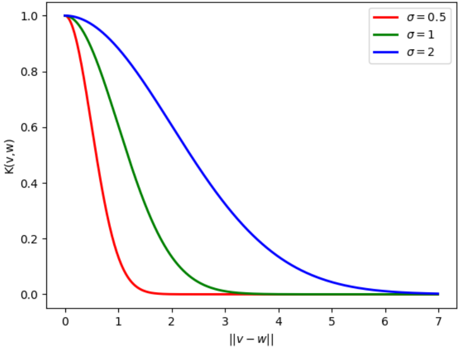
Different kernels in practice: Gaussian

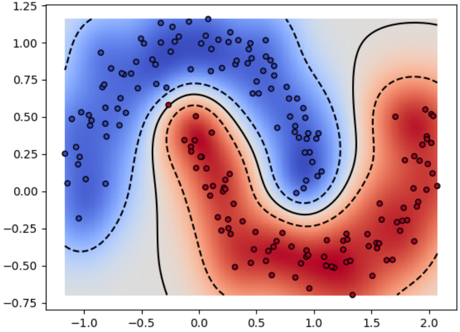
$$K(\boldsymbol{v}, \boldsymbol{w}) = \exp\left(-\frac{||\boldsymbol{v}-\boldsymbol{w}||^2}{2\sigma^2}\right)$$



Different kernels in practice: Gaussian

$$K(\boldsymbol{v}, \boldsymbol{w}) = \exp\left(-\frac{||\boldsymbol{v}-\boldsymbol{w}||^2}{2\sigma^2}\right)$$





Summary

- SVMs in their standard form have problems classifying non-linearly separable datasets
- Use $\phi(x)$ to map data into a space where this is possible
- Use the kernel to simplify the resulting calculation