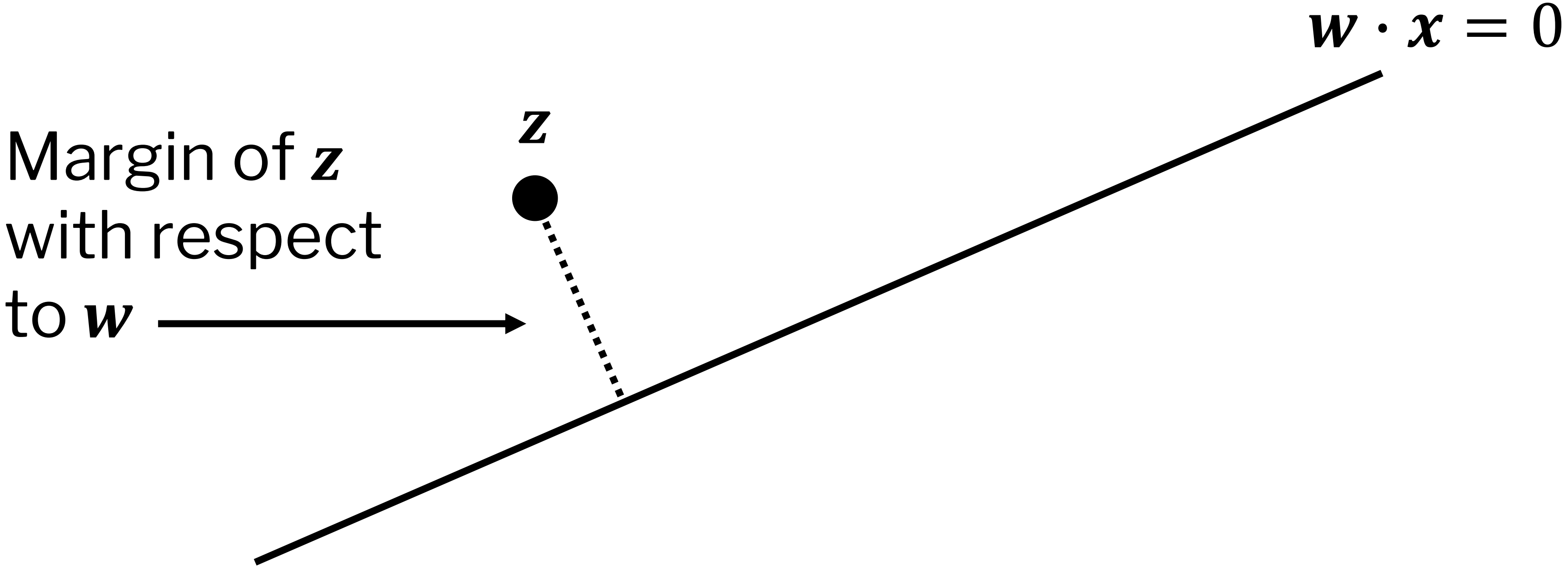
- Support vector machines (SVM)
  - Duality

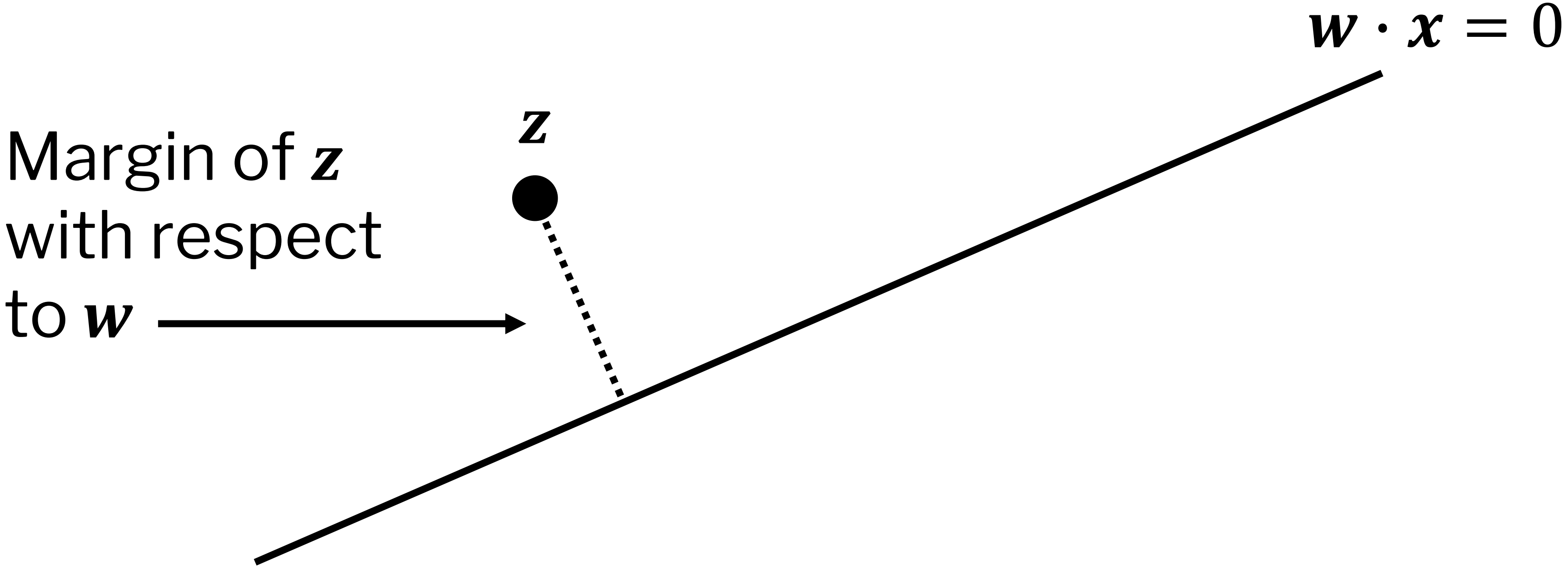
- Kernels
  - SVM with kernels

Definition: Geometric margin

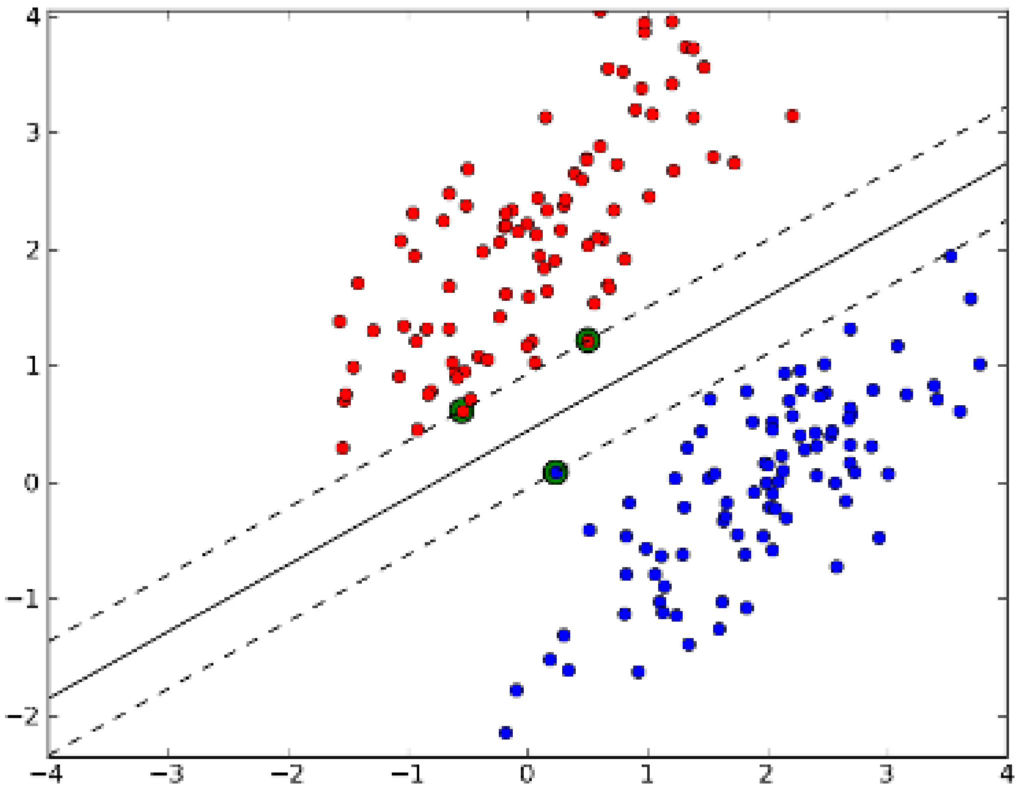
The **geometric margin** of a vector $\boldsymbol{z}$ with respect to a linear separator $\boldsymbol{w}$ is the distance from $\boldsymbol{z}$ to the plane $\boldsymbol{w} \cdot \boldsymbol{x} = 0$.



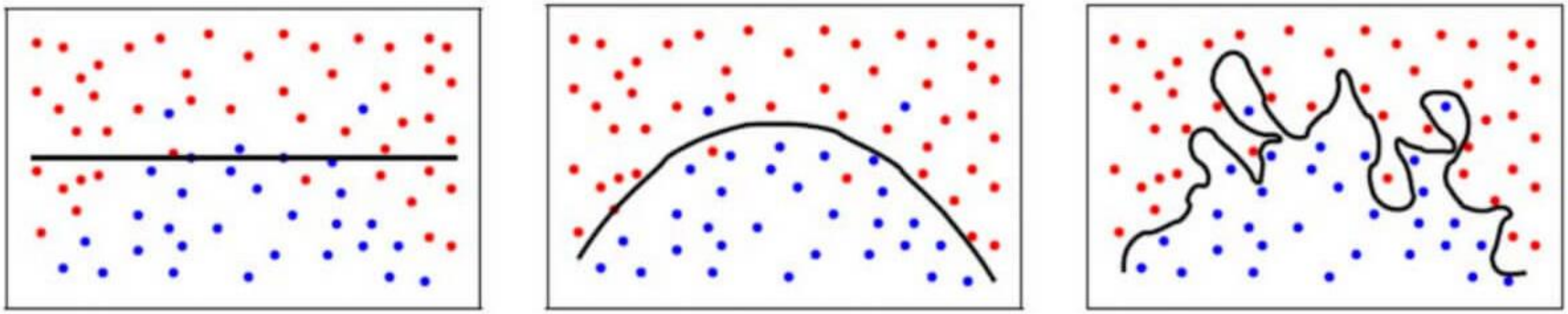
The margin equals $\frac{|\boldsymbol{w}\cdot\boldsymbol{z}|}{||\boldsymbol{w}||}$.



When the data is **linearly separable**, the “support vector machine” (SVM) algorithm finds the linear separator with maximum margin.

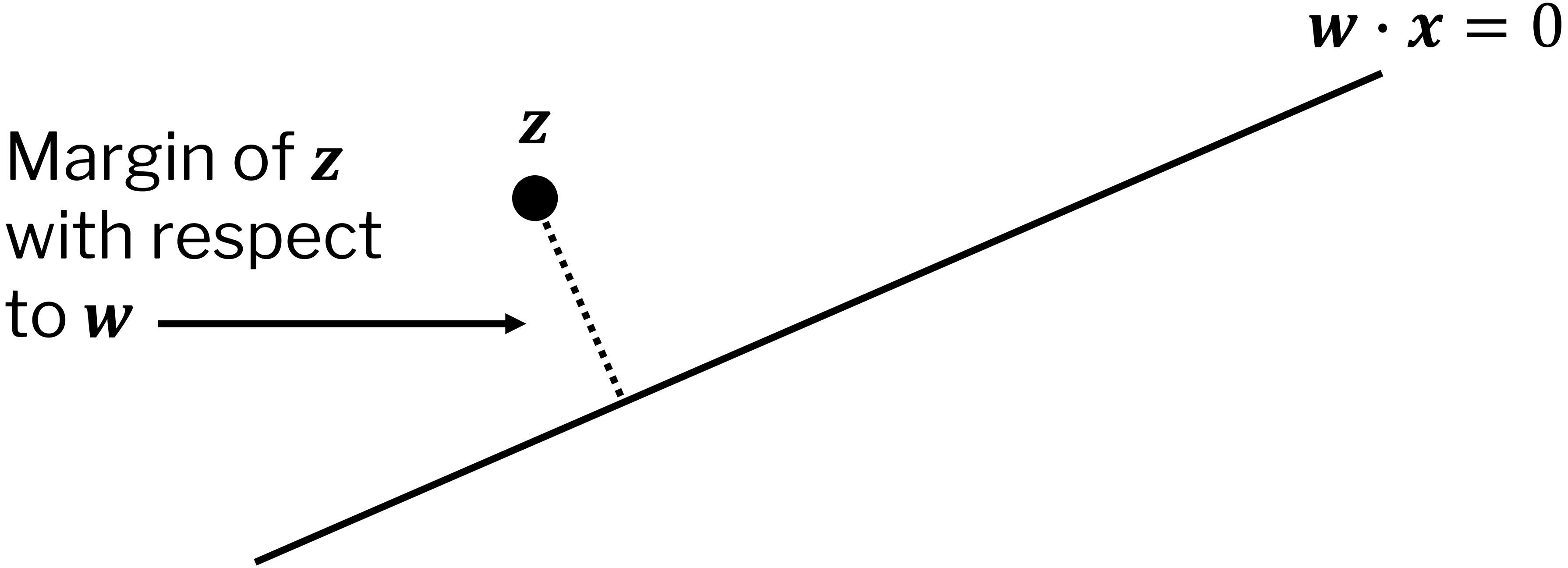


Underfitting ⟷ Overfitting



Theorem

Suppose the input data to the perceptron algorithm is linearly separable with a margin of γ. Also, suppose the data points lie in a ball of radius R. Then the Perceptron algorithm makes at most $(R/\gamma)^2$ mistakes before it converges.



Suppose $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ is a set of labeled vectors that are **linearly separable**.

The linear separator $\boldsymbol{w}$ with the largest margin is the solution to:

$$\text{maximize} \min_{i \in [m]} \frac{|\boldsymbol{x}_i \cdot \boldsymbol{w}|}{||\boldsymbol{w}||}$$

such that $\boldsymbol{x}_i \cdot \boldsymbol{w} > 0$ if $y_i = 1$ and $\boldsymbol{x}_i \cdot \boldsymbol{w} < 0$ if $y_i = -1$.

The linear separator $\boldsymbol{w}$ with the largest margin is the solution to:

$$\text{maximize} \min_{i \in [m]} \frac{|\boldsymbol{x}_i \cdot \boldsymbol{w}|}{||\boldsymbol{w}||}$$

such that $\boldsymbol{x}_i \cdot \boldsymbol{w} > 0$ if $y_i = 1$ and $\boldsymbol{x}_i \cdot \boldsymbol{w} < 0$ if $y_i = -1$.

In the homework, you'll show that this is equivalent to the quadratic program you saw in class:

$$\text{minimize} ||\boldsymbol{w}||^2$$

such that $y_i(\boldsymbol{x_i} \cdot \boldsymbol{w}) \geq 1$.

Support vector machines (SVM)

- Duality

Kernels

- SVM with kernels

Hard SVM

$$\text{minimize}\ \frac{\|\boldsymbol{w}\|^2}{2}$$

$$\text{such that } y_i(\boldsymbol{x_i} \cdot \boldsymbol{w}) \geq 1$$

Question:

Let $\boldsymbol{x}$ be a vector and let y be a label in $\{-1,1\}$.

If $\mathrm{y}(\boldsymbol{w} \cdot \boldsymbol{x}) \geq 1$, what is $\max_{\alpha \geq 0} \alpha \left(1 - \mathrm{y}(\boldsymbol{w} \cdot \boldsymbol{x})\right)$?

Answer:

0.

Question:

Let $\boldsymbol{x}$ be a vector and let y be a label in $\{-1,1\}$.

If $\mathrm{y}(\boldsymbol{w} \cdot \boldsymbol{x}) < 1$, what is $\max_{\alpha \geq 0} \alpha \left(1 - \mathrm{y}(\boldsymbol{w} \cdot \boldsymbol{x})\right)$?

Answer:

∞.

Question.

Suppose $\{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_m}, y_m)\}$ is a set of labeled vectors that are **linearly separable**.

$$g(\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \boldsymbol{0}} \sum_{i=1}^{m} \alpha_i \big(1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x_i})\big)$$

$$= ?$$

Answer.

Suppose $\{(\boldsymbol{x_1}, y_1), \dots, (\boldsymbol{x_m}, y_m)\}$ is a set of labeled vectors that are **linearly separable**.

$$g(\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \boldsymbol{0}} \sum_{i=1}^{m} \alpha_i \big(1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x_i})\big)$$

$$= \begin{cases} 0 & \text{if } \forall i, y_i(\boldsymbol{w} \cdot \boldsymbol{x_i}) \geq 1 \\ \infty & \text{otherwise} \end{cases}$$

Question.

Hard SVM
$$\text{minimize} \frac{\|\boldsymbol{w}\|^2}{2}$$ such that $y_i(\boldsymbol{x_i} \cdot \boldsymbol{w}) \geq 1$

How can we rewrite Hard SVM using $g(\boldsymbol{w})$?

Answer.

Hard SVM
$$\text{minimize}\ \frac{\|\boldsymbol{w}\|^2}{2}$$ such that $y_i(\boldsymbol{x_i} \cdot \boldsymbol{w}) \geq 1$

$$\text{minimize}\left(\frac{\|\boldsymbol{w}\|^2}{2} + g(\boldsymbol{w})\right)$$

$$= \text{minimize}\left(\frac{\|\boldsymbol{w}\|^2}{2} + \max_{\boldsymbol{\alpha}\in\mathbb{R}^m, \boldsymbol{\alpha}\geq\boldsymbol{0}} \sum_{i=1}^{m} \alpha_i\left(1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x_i})\right)\right)$$

“Lagrange multipliers”

Question.

Fact (Strong duality)
$$\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)$$ $$=\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)\right\}$$

Suppose we've figured out the $\boldsymbol{\alpha}$ that maximizes

$$\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)$$

What $\boldsymbol{w}$ minimizes this expression?

Suppose we've figured out the $\boldsymbol{\alpha}$ that maximizes

$$\min_{\boldsymbol{w}} \left(\frac{\|\boldsymbol{w}\|^2}{2} + \sum_{i=1}^{m} \alpha_i \big(1 - y_i (\boldsymbol{w} \cdot \boldsymbol{x_i})\big) \right)$$

What $\boldsymbol{w}$ minimizes this expression?

Answer.

Take the gradient and set it to $\mathbf{0}$:

$$\boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i} = \mathbf{0} \qquad \Longrightarrow \qquad \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i}$$

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)\right\},\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x_i}$$

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\frac{1}{2}\left\|\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x_i}\right\|^2+\sum_{i=1}^{m}\alpha_i\left(1-y_i\left(\sum_{j=1}^{m}\alpha_j y_j\boldsymbol{x_j}\right)\cdot\boldsymbol{x_i}\right)\right\}$$

$$=\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(\boldsymbol{x_i}\cdot\boldsymbol{x_j})\right\}$$

We only have to care about dot products!

Support vector machines (SVM)

Duality

➡ Kernels

SVM with kernels

Question:

Consider this set of points labeled points.
Are they **linearly separable**?

x	y
-3	-1
-2	-1
-1	1
0	1
1	1
2	1
3	-1

Answer:

No.

Question:

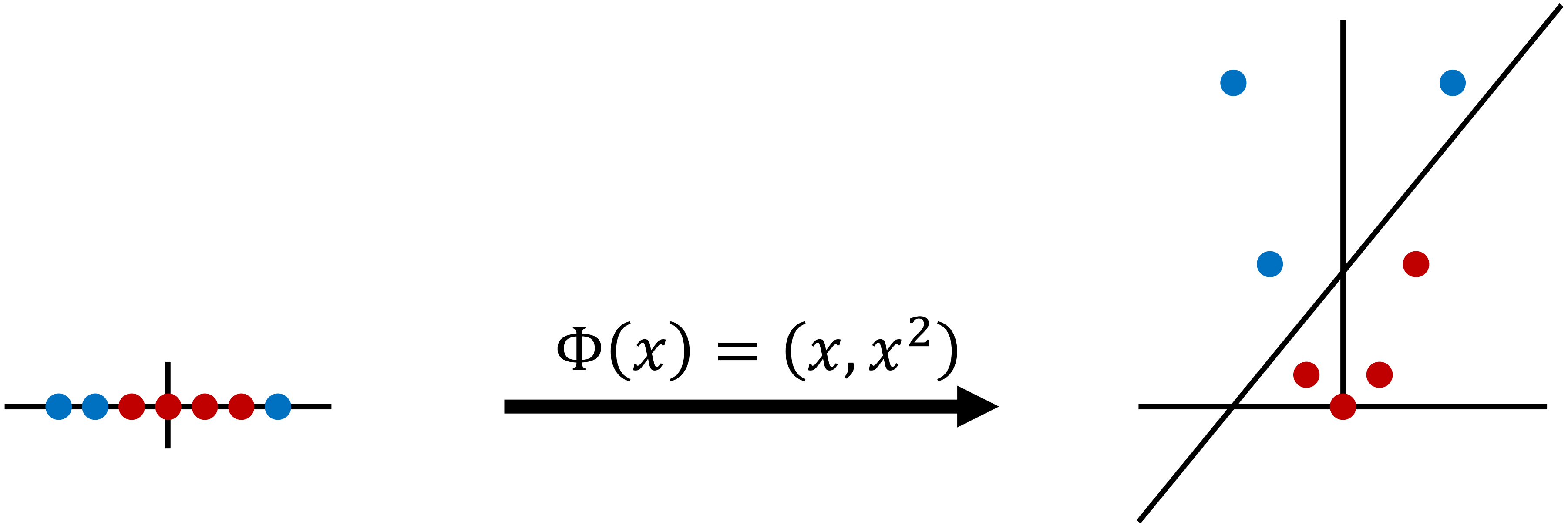
Consider this set of points labeled points.

Is there a way to map these points into $\mathbb{R}^2$ so that they become linearly separable?

x	y
-3	-1
-2	-1
-1	1
0	1
1	1
2	1
3	-1

Answer:

Yes. $\Phi(x) = (x, x^2)$.


$\Phi(x) = (x, x^2)$

Definition: Kernel

The function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if it can be written as an inner product:

- There exists a mapping $\Phi: \mathcal{X} \to \mathbb{R}^d$ such that
  $K(x, y) = \Phi(x) \cdot \Phi(y)$ for all $x, y \in \mathcal{X}$.

Question:

Let the original instance space be $\mathbb{R}$. Consider the mapping Φ where for each integer $n \geq 0$, there is a component

$$\Phi(x)_n = \frac{1}{\sqrt{n!}} e^{-x^2/2} x^n$$

What is the corresponding kernel $K: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$?

In other words, what is $K(x, y) = \Phi(x) \cdot \Phi(y)$ for any $x, y \in \mathbb{R}$?

You can use the fact that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.

Answer:

$$K(x, y) = \Phi(x) \cdot \Phi(y)$$

$$= \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \right) \left(\frac{1}{\sqrt{n!}} e^{-\frac{y^2}{2}} y^n \right)$$

$$= e^{-\frac{x^2+y^2}{2}} \sum_{n=0}^{\infty} \frac{(xy)^n}{n!}$$

$$= e^{-\frac{(x-y)^2}{2}}$$

More generally, given a scalar $\sigma > 0$, the **Gaussian kernel** (also known as the Radial Basis Function (RBF) kernel) is defined to be

$$K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma}}$$

Let Φ be the mapping such that $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$. Then Φ maps to an infinite-dimensional space but $\boldsymbol{K(x, y)}$ **requires very few computations.**

Support vector machines (SVM)

- Duality

Kernels

- SVM with kernels

Hard SVM

$$\text{minimize} \frac{\|\boldsymbol{w}\|^2}{2}$$
such that $y_i(\boldsymbol{x_i} \cdot \boldsymbol{w}) \geq 1$

Hard SVM is equivalent to:

$$\text{minimize} \left(\frac{\|\boldsymbol{w}\|^2}{2} + g(\boldsymbol{w}) \right)$$

$$= \text{minimize} \left(\frac{\|\boldsymbol{w}\|^2}{2} + \max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \boldsymbol{0}} \sum_{i=1}^{m} \alpha_i \big(1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x_i})\big) \right)$$

"Lagrange multipliers"

Question.

Fact (Strong duality)
$$\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\boldsymbol{0}}\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)$$ $$=\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\boldsymbol{0}}\left\{\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)\right\}$$

Suppose we've figured out the $\boldsymbol{\alpha}$ that maximizes

$$\min_{\boldsymbol{w}}\left(\frac{\|\boldsymbol{w}\|^2}{2}+\sum_{i=1}^{m}\alpha_i\big(1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i})\big)\right)$$

What $\boldsymbol{w}$ minimizes this expression?

Suppose we've figured out the $\boldsymbol{\alpha}$ that maximizes

$$\min_{\boldsymbol{w}} \left(\frac{\|\boldsymbol{w}\|^2}{2} + \sum_{i=1}^{m} \alpha_i \big(1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x_i})\big) \right)$$

What $\boldsymbol{w}$ minimizes this expression?

Answer.

Take the gradient and set it to $\mathbf{0}$:

$$\boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i} = \mathbf{0} \qquad \Longrightarrow \qquad \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i}$$

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \mathbf{0}} \left\{ \min_{\boldsymbol{w}} \left(\frac{\|\boldsymbol{w}\|^2}{2} + \sum_{i=1}^{m} \alpha_i \big(1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x_i})\big) \right) \right\}, \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i}$$

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \mathbf{0}} \left\{ \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i} \right\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(\sum_{j=1}^{m} \alpha_j y_j \boldsymbol{x_j} \right) \cdot \boldsymbol{x_i} \right) \right\}$$

$$= \max_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\alpha} \geq \mathbf{0}} \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x_i} \cdot \boldsymbol{x_j}) \right\}$$

We only have to care about dot products!

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right)\right\}$$

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\left(\Phi(\boldsymbol{x_i})\cdot\Phi(\boldsymbol{x_j})\right)\right\}$$

$$=\max_{\boldsymbol{\alpha}\in\mathbb{R}^m,\boldsymbol{\alpha}\geq\mathbf{0}}\left\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i},\boldsymbol{x_j})\right\}$$

Question:

Suppose we have learned a good linear separator

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

Suppose we want to calculate $w \cdot \Phi(x)$ for some new instance x. How can we write this in terms of the kernel function K?

Answer:

$$w \cdot \Phi(x)$$

$$= \left(\sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)\right) \cdot \Phi(x)$$

$$= \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i) \cdot \Phi(x)$$

$$= \sum_{i=1}^{m} \alpha_i y_i K(x_i, x)$$