Model

Two kinds of models: deterministic models and probabilistic models

Deterministic model: Given an input, produces an output

- Example: $\hat{y} = \vec{w}^{\top}\vec{x}$

Probabilistic model: Given an input, produces a distribution over possible outputs.

- Example: $y|\vec{x}, \vec{w}, \sigma \sim \mathcal{N}\left(\vec{w}^{\top}\vec{x}, \sigma^2\right)$

Loss Function

For a deterministic model: Can be any function that assigns high values to incorrect outputs and low values to correct outputs.

For a probabilistic model:

- Maximum likelihood (MLE): Loss function is the negative log-likelihood
$$\log p\left(\mathcal{D}\middle|\vec{\theta}\right)$$

- Maximum a posteriori (MAP): Loss function is the negative log-posterior
$$\log p\left(\vec{\theta}\middle|\mathcal{D}\right) = \log\left(\frac{p(\vec{\theta})p(\mathcal{D}|\vec{\theta})}{p(\mathcal{D})}\right)$$

Training

Training involves finding the model parameters that minimize the loss. Several approaches:

- Set the gradient to zero and solve for the optimal parameters analytically.
- If there is no closed-form solution for the optimal parameters:
  - If optimization problem is unconstrained: use iterative gradient-based optimization methods.
  - **Next: constrained optimization**

Machine Learning

CMPT 726

Mo Chen
SFU School of Computing Science

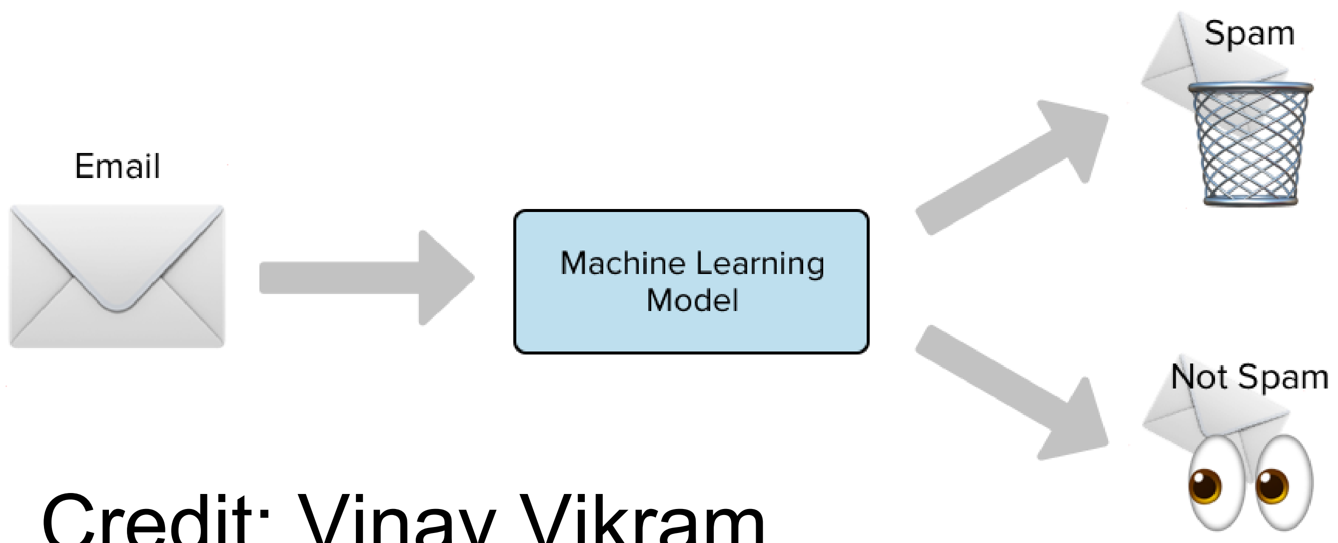
2025-10-29

Classification

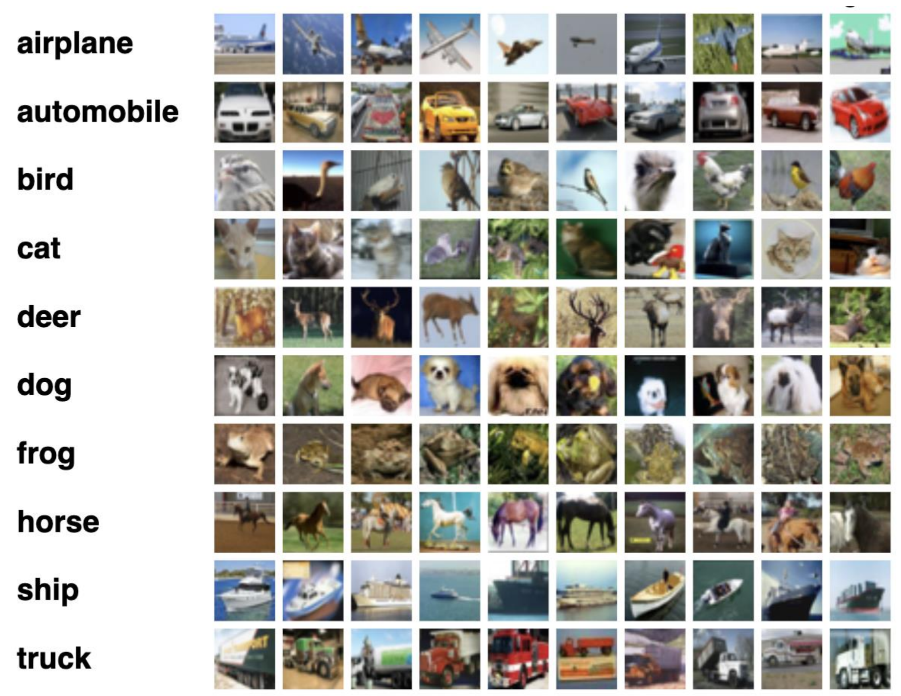
Motivation

In many problems, we would like to categorize an observation:

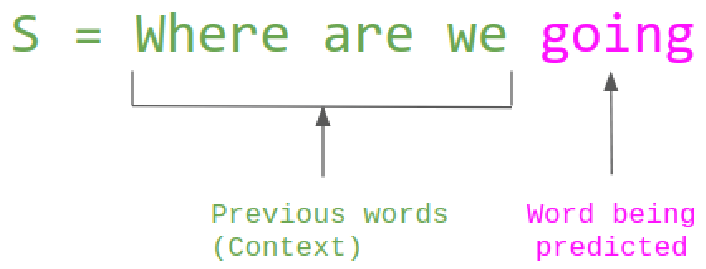
- Is an email spam?
- What object is depicted in an image?
- What will be the next word?



Credit: Vinay Vikram



Credit: Alex Krizhevsky



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

Credit: The Gradient

Binary Classification

Unlike regression, the goal of classification is to classify the input into one of multiple discrete classes.

A regression model produces a real number or in the case of multiple output regression, a real vector.

A classification model produces a class prediction.

Such a model is known as a **classifier**.

In binary classification, the goal is to classify into one of **two** discrete classes.

A binary classification model is known as a **binary classifier**.

Without loss of generality, we call one class the **positive class** and the other the **negative class**.

Data points whose labels are positive are known as **positive examples** and data points whose labels are negative are known as **negative examples**.

Support Vector Machines

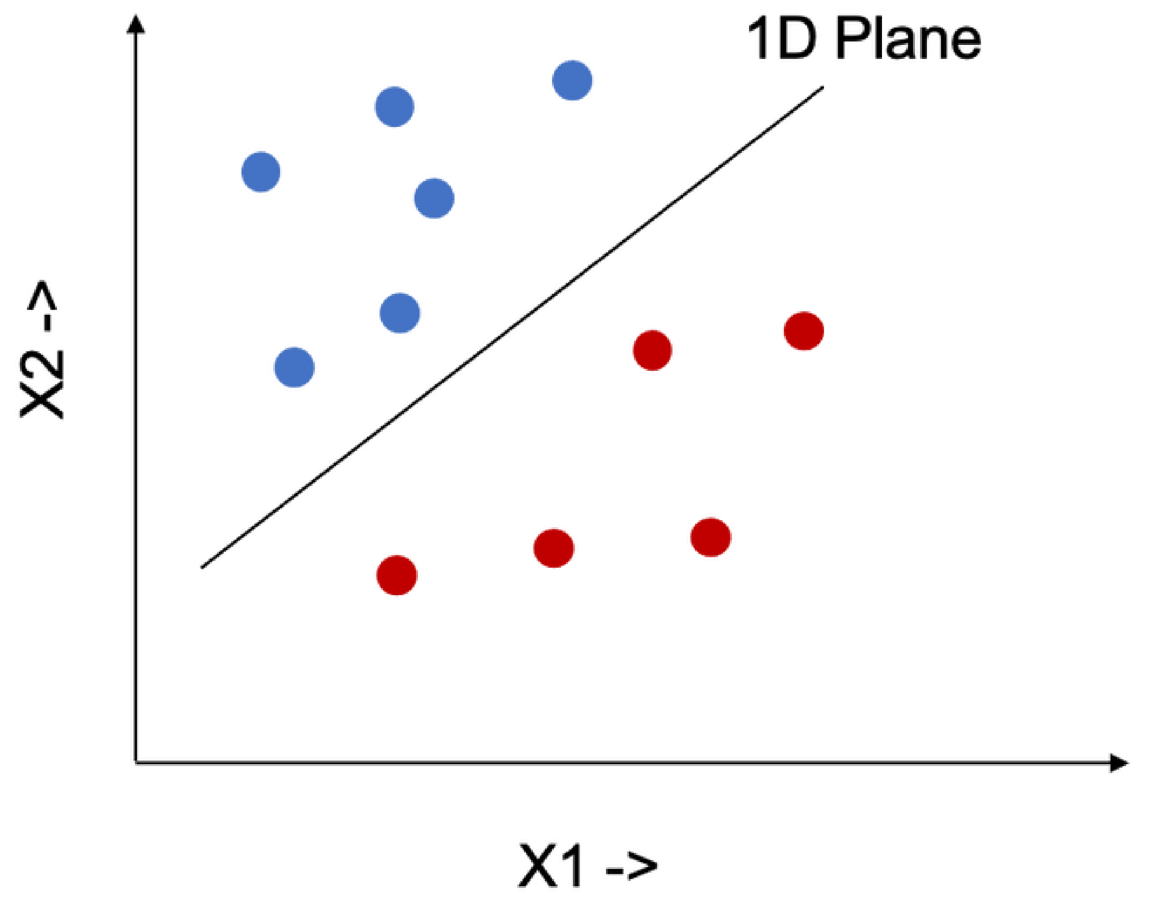
Given a dataset of input-output pairs (a.k.a. "observations"), $\{(\vec{x}_i, y_i)\}_{i=1}^{N}$, where $\vec{x}_i \in \mathbb{R}^{n-1}$ and $y_i \in \{-1,1\}$

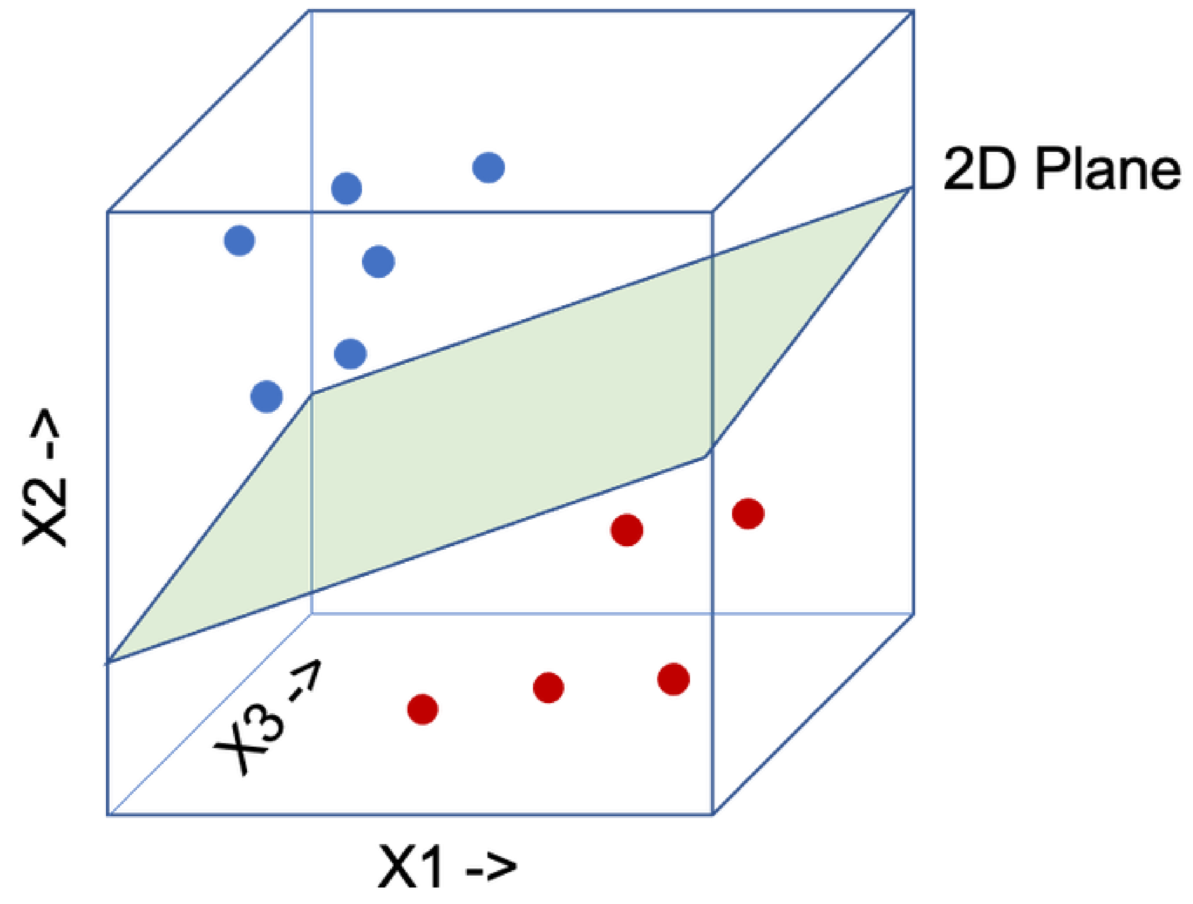
We will construct a model called the support vector machine (SVM) to predict the label y from the data point $\vec{x}$.

The model is simply a line (in the case of 2D data), a plane (in the case of 3D data) or more generally, a hyperplane (in the case of higher dimensional data) that separates the data points.

For a new data point on one side, we predict the positive label.

For a new data point on the other side, we predict the negative label.





Credit: Abhisek Jana

Support Vector Machines

The **decision boundary** is the boundary that separates the region where the model generates positive predictions from the region where the model generates negative predictions.

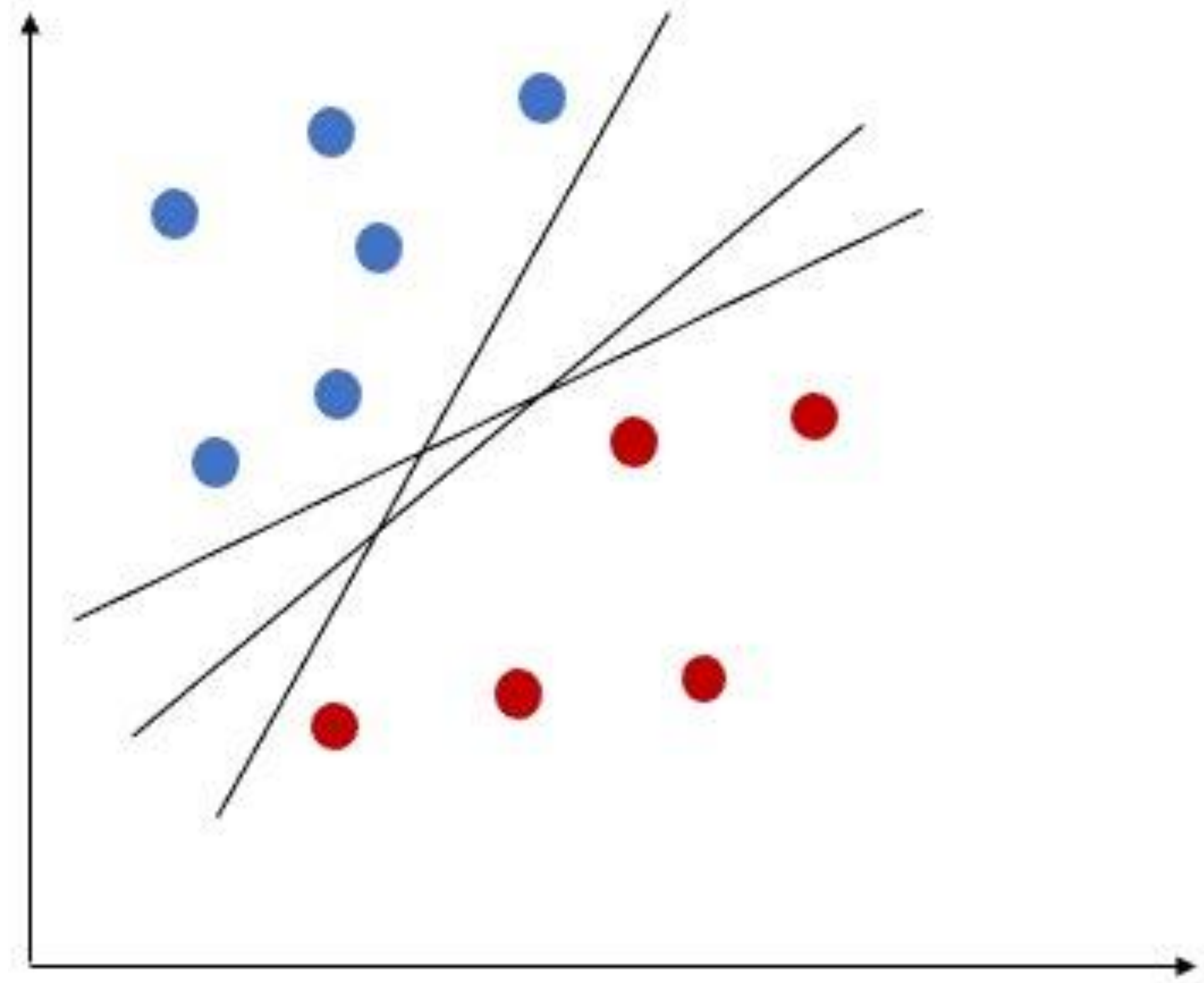
A **linear classifier** whose decision boundary is a hyperplane.

The support vector machine is an example of a linear classifier.

Support Vector Machines

There are many hyperplanes that would classify a training dataset perfectly.

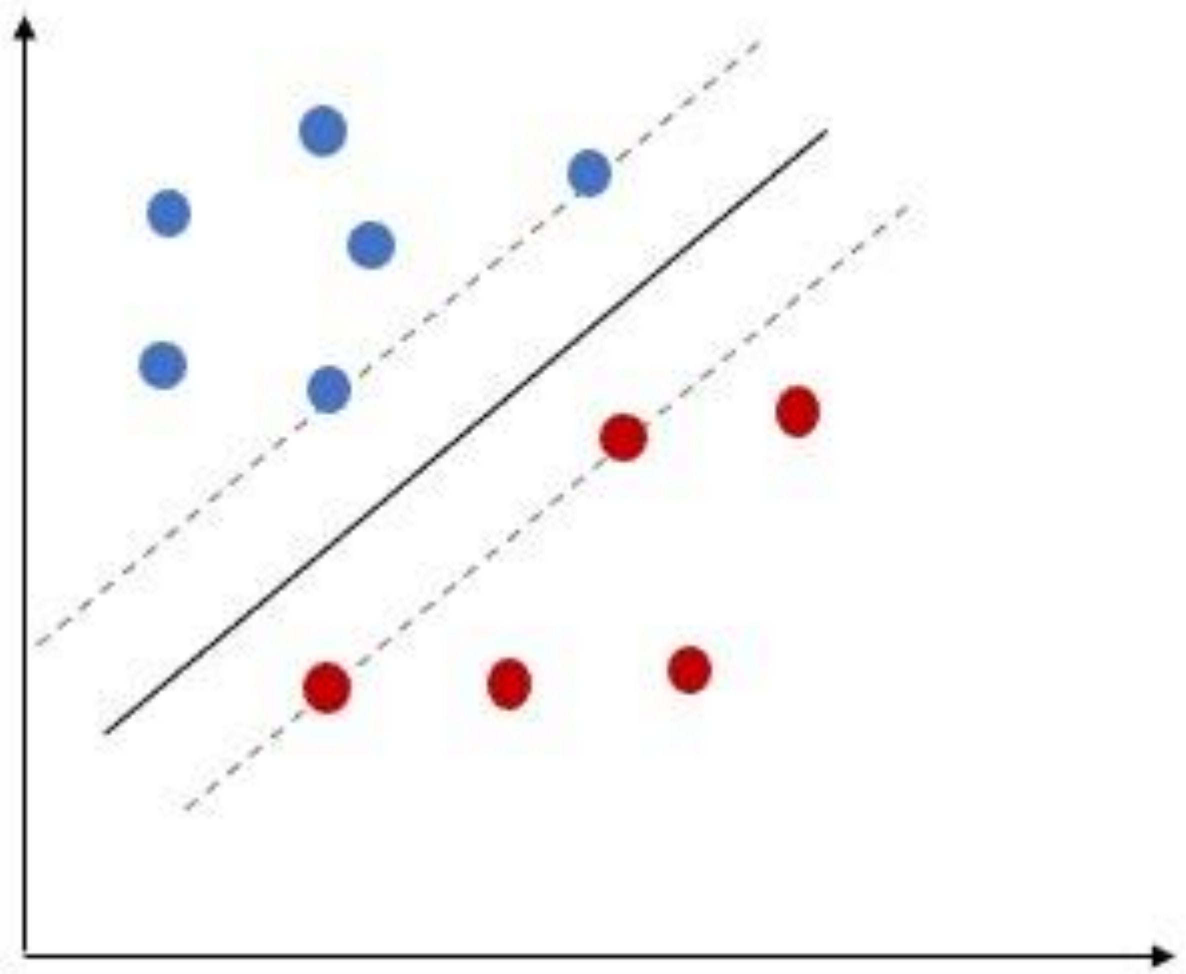
Which one should we choose?



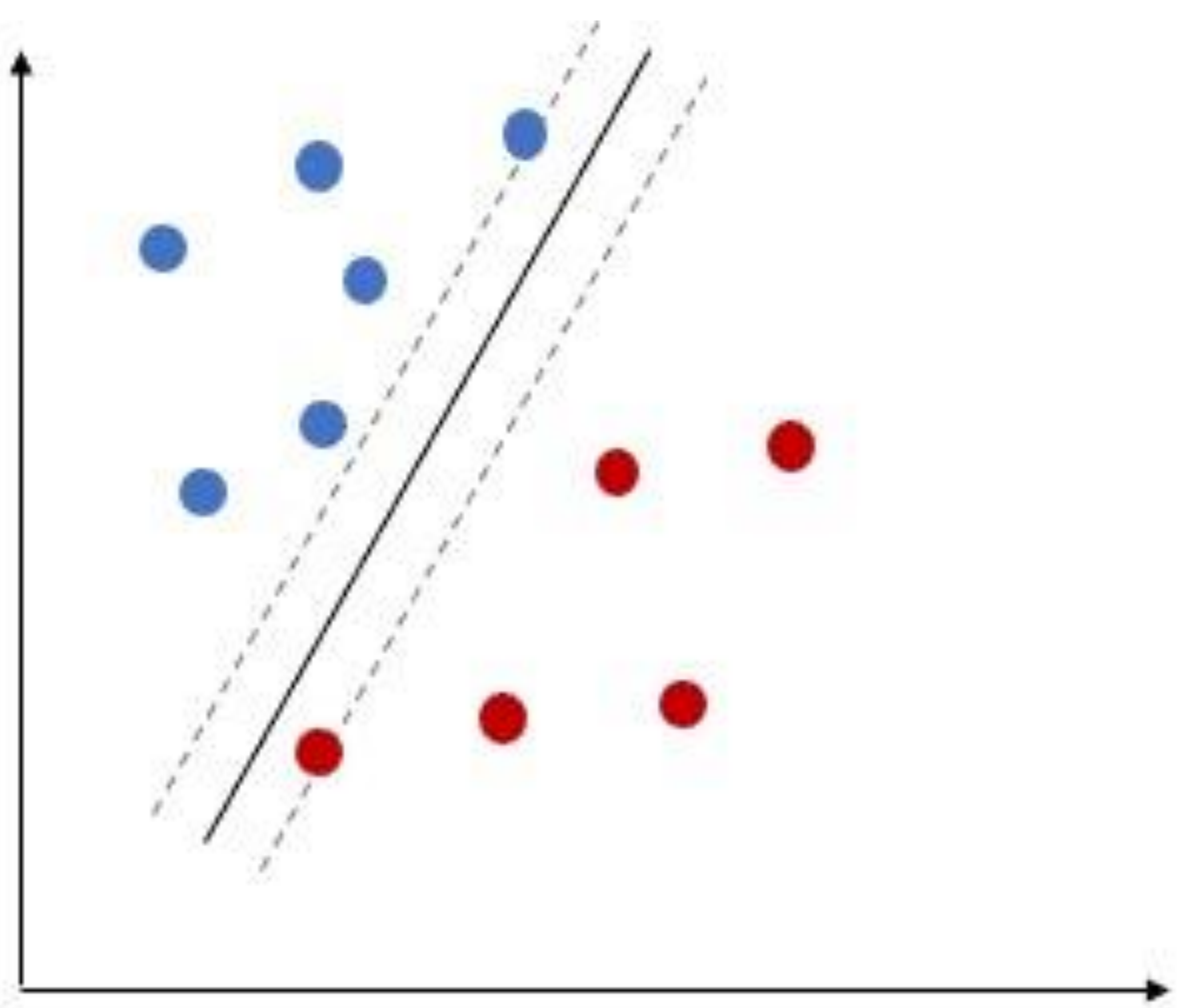
Support Vector Machines

A boundary that is as far away from data points as possible is more robust to perturbations to the data points.

Intuitively, such a boundary is less prone to overfitting.



Preferred



Less Preferred

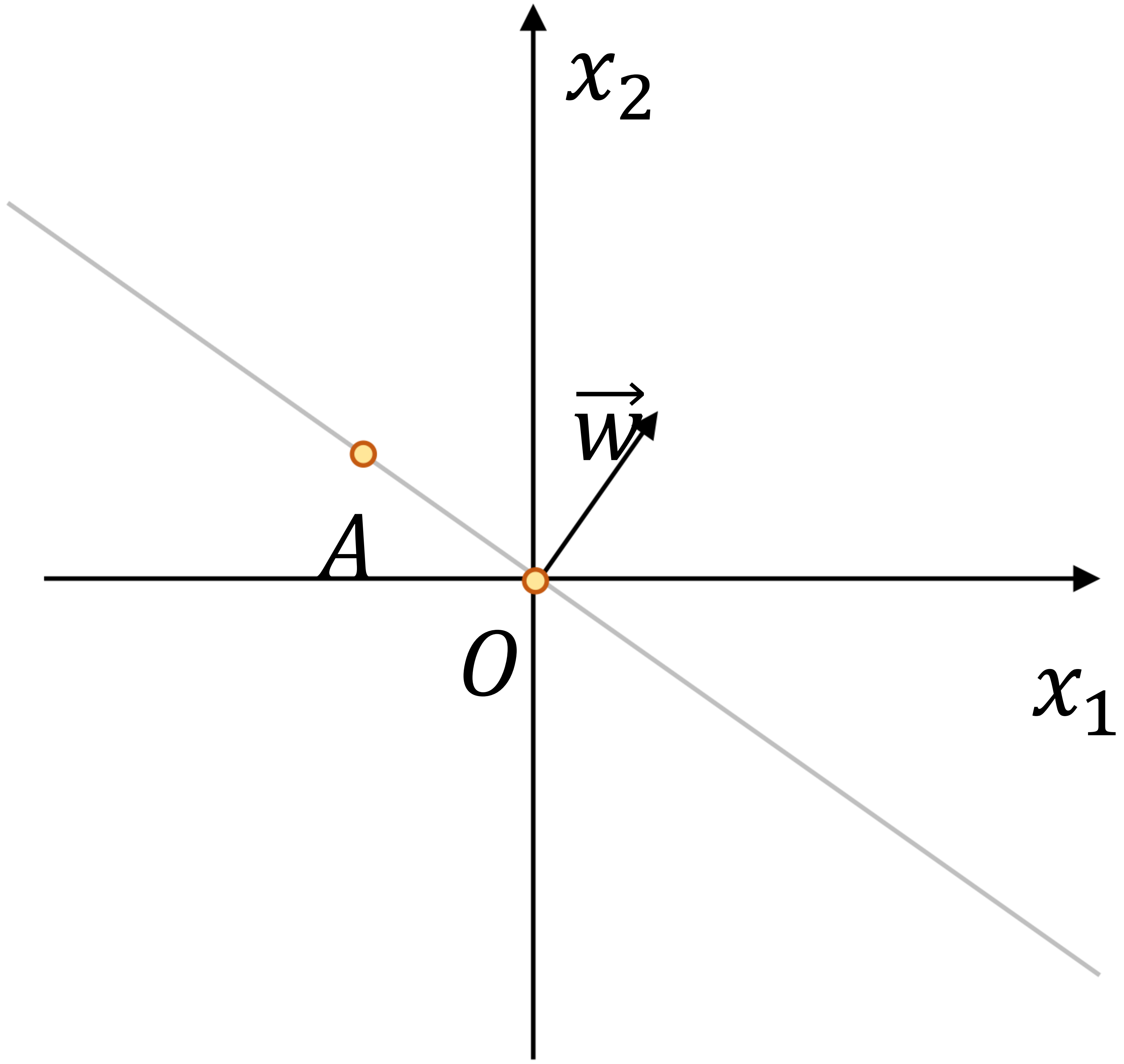
Let’s formulate an optimization problem to find such a boundary.

Hyperplanes

Consider a vector $\vec{w}$, and a hyperplane that is perpendicular to it (shown on the right).

For any A on the hyperplane, $\overrightarrow{OA}$ is perpendicular to $\vec{w}$.

Hence, $\vec{w}^{\top}(\overrightarrow{OA}) = 0$. So, this hyperplane corresponds to the set $\{\vec{x} | \vec{w}^{\top}\vec{x} = 0\}$.



Special case when $b = 0$

Hyperplanes

We shift the hyperplane in parallel (as shown on the right).

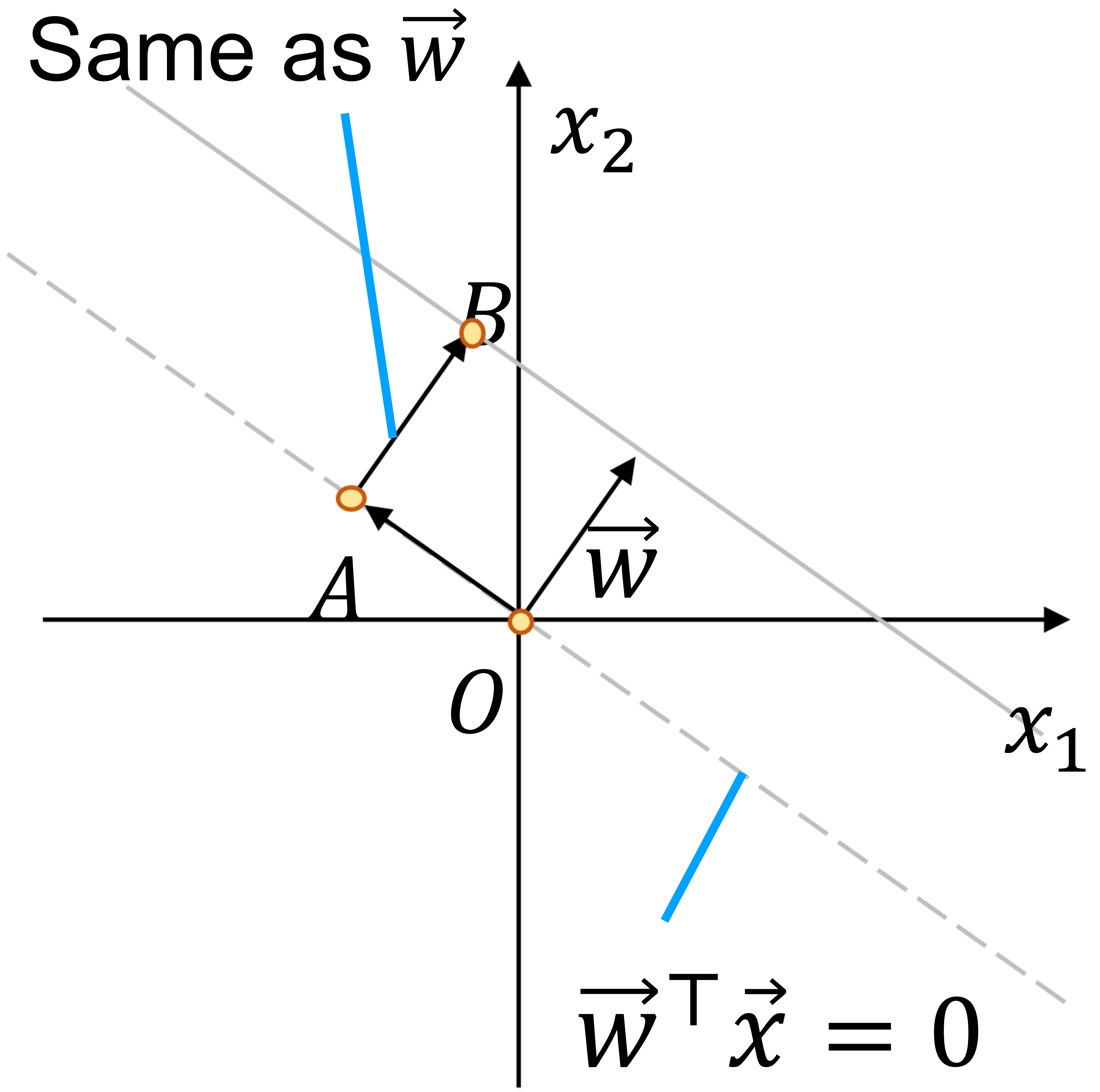
Consider any point B on the hyperplane and the vector $\overrightarrow{OB}$.

$$\vec{w}^\top\left(\overrightarrow{OB}\right) = \vec{w}^\top\left(\overrightarrow{OA} + \overrightarrow{AB}\right)$$
$$= \vec{w}^\top\left(\overrightarrow{OA}\right) + \vec{w}^\top\left(\overrightarrow{AB}\right)$$
$$= 0 + \vec{w}^\top\vec{w}$$
$$= \|\vec{w}\|_2^2$$

So, for any B on the hyperplane, $\vec{w}^\top(\overrightarrow{OB}) = \|\vec{w}\|_2^2$.

Hence, the hyperplane corresponds to the following set:

$$\{\vec{x} \mid \vec{w}^\top\vec{x} = \|\vec{w}\|_2^2\}$$

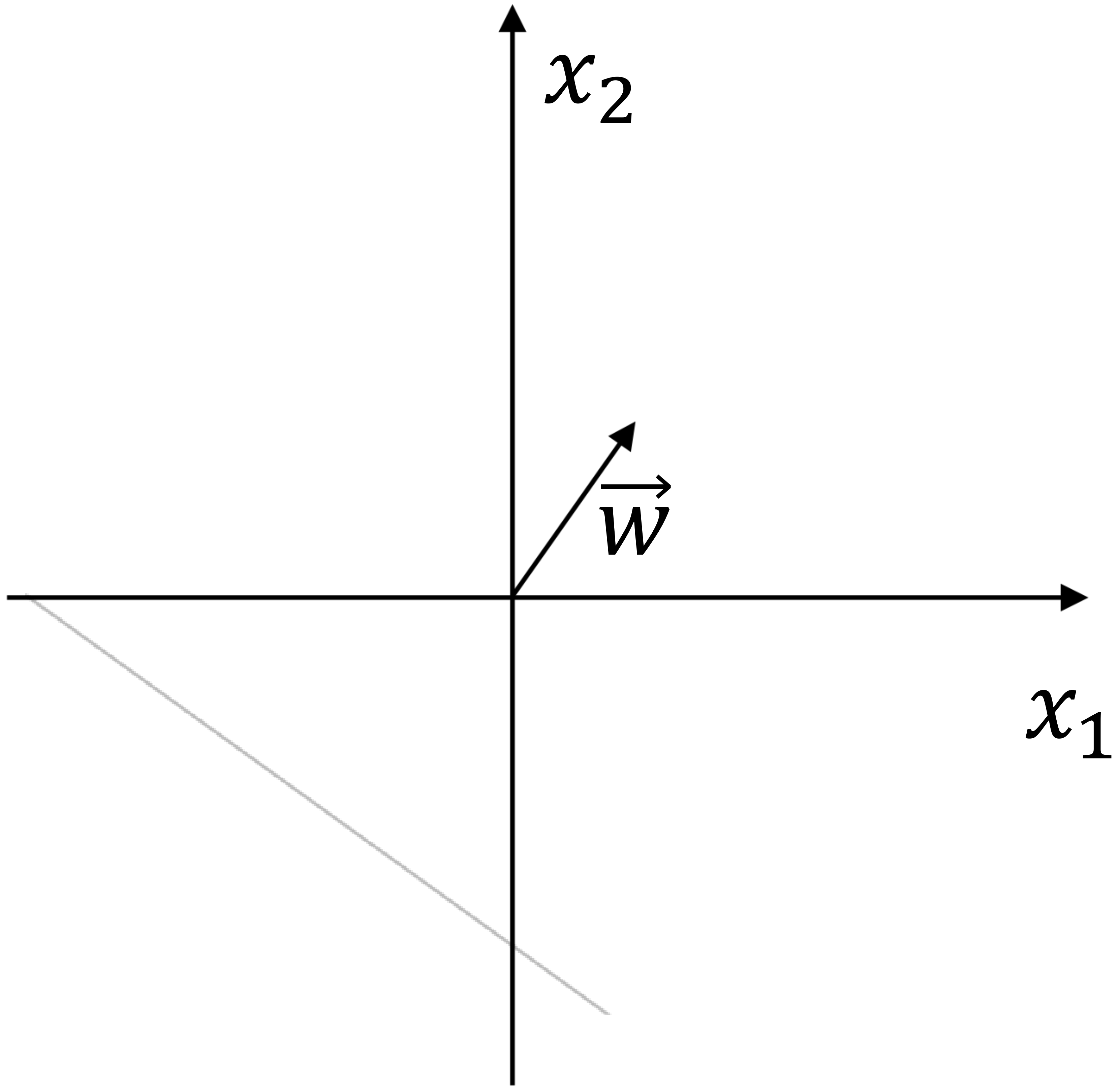


Special case when $b = \|\vec{w}\|_2^2$

Hyperplanes

So, in general:

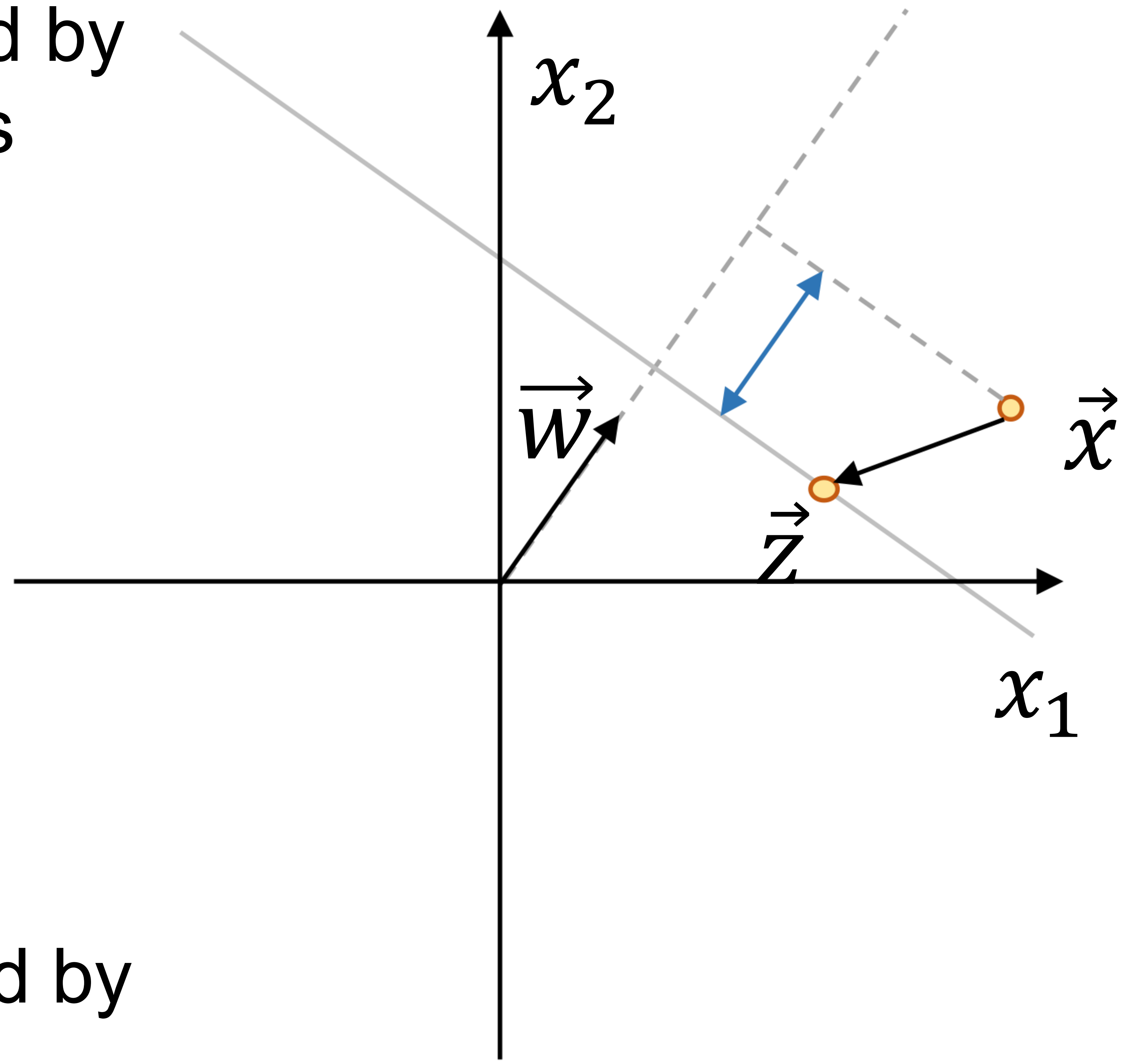
As b changes, $\{\vec{x} | \vec{w}^\top \vec{x} = b\}$ corresponds to shifting the hyperplane in parallel.



Distance to the Hyperplane

Consider an arbitrary points on the hyperplane, $\vec{z}$.

The distance to the hyperplane from $\vec{x}$ can be obtained by projecting the vector $\vec{x} - \vec{z}$ along the vector $\vec{w}$) which is orthogonal to the hyperplane).



The length of the projection is given by:

$$\left|\left\langle \vec{x} - \vec{z}, \frac{\vec{w}}{\|\vec{w}\|_2} \right\rangle\right| = \frac{1}{\|\vec{w}\|_2} |\langle \vec{x} - \vec{z}, \vec{w} \rangle|$$

$$= \frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x} - \vec{w}^\top \vec{z} \right|$$

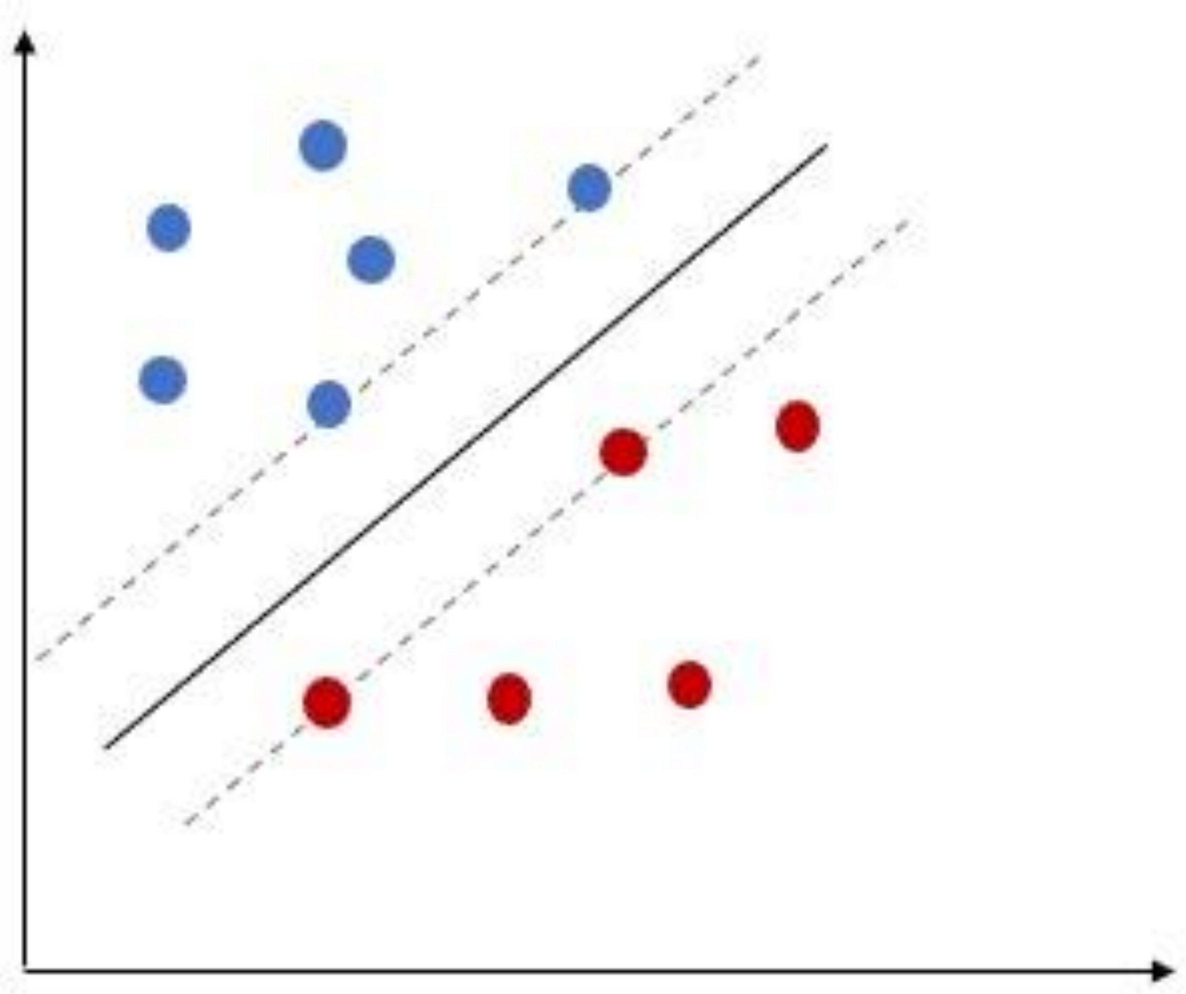
Because $\vec{z}$ is on the hyperplane, which is characterized by $\{\vec{x} | \vec{w}^\top \vec{x} = b\}$, $\vec{w}^\top \vec{z} = b$.

So, the distance to the hyperplane from $\vec{x}$ is $\frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x} - b \right|$.

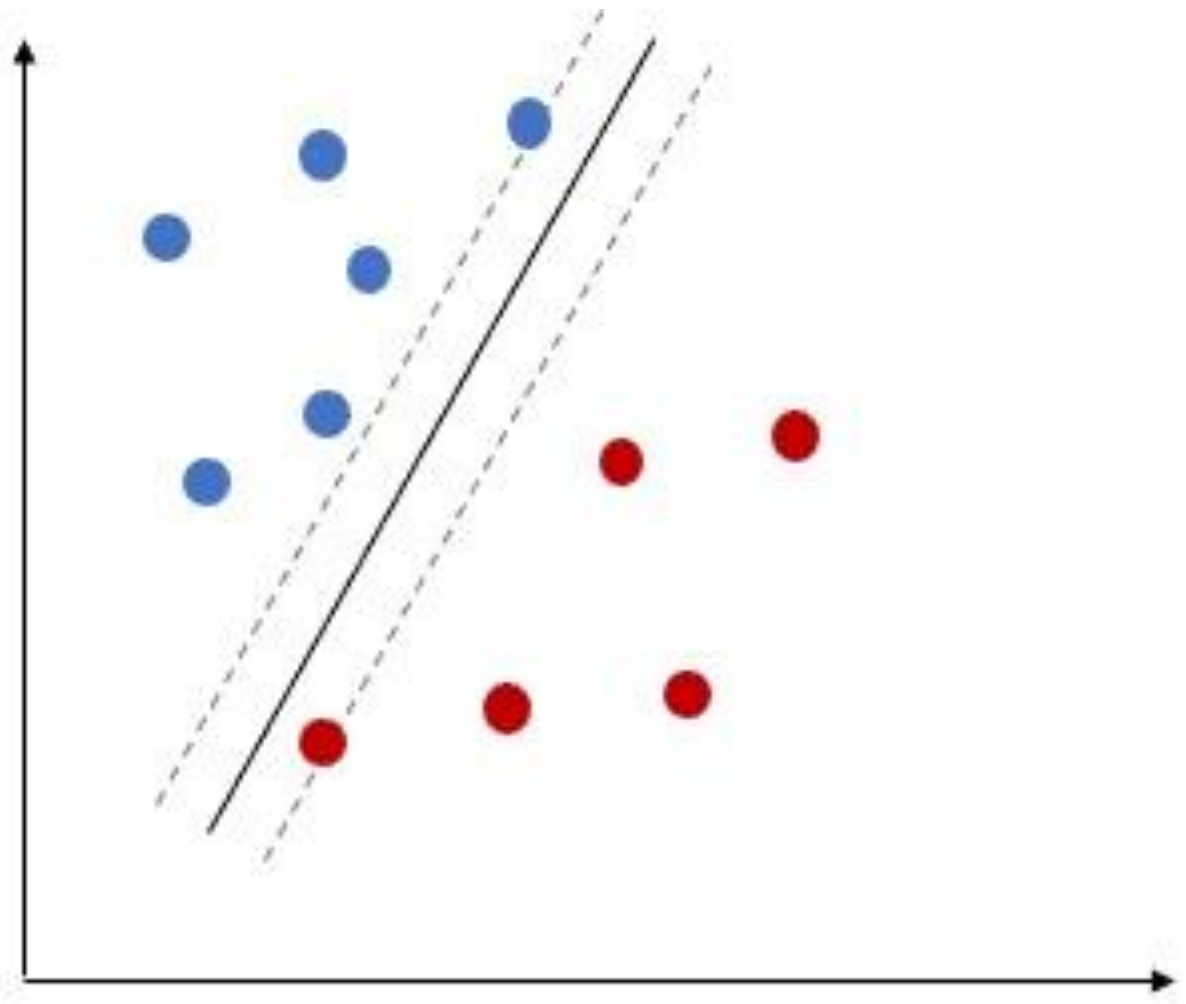
Goals of SVM

Recall: We would like to find a hyperplane that:

(1) Separates the positive data points from the negative data points
(2) Is as far away from the data points as possible



Preferred



Less Preferred

Margin

We define the width of the buffer on each side of the hyperplane as the margin.

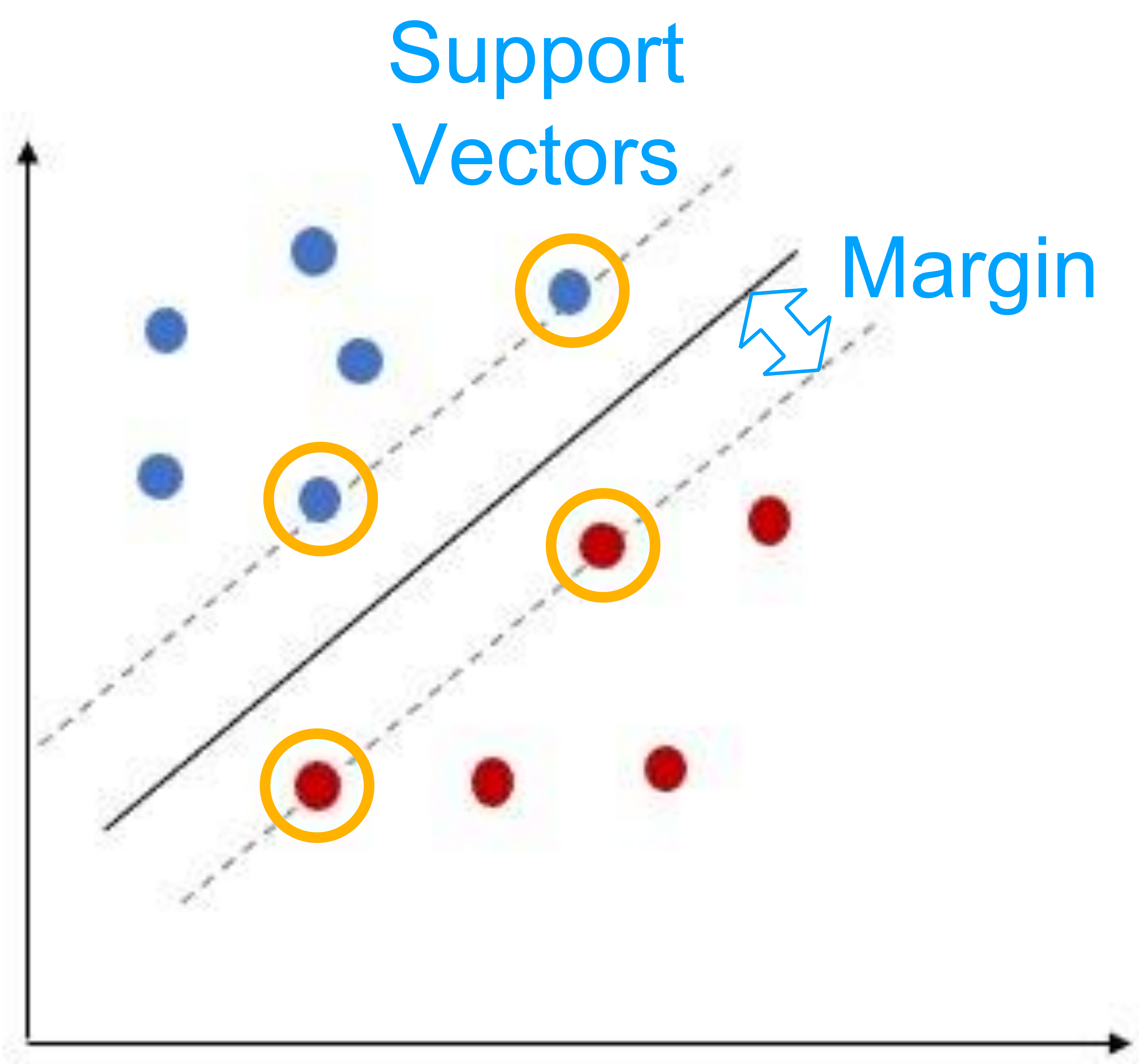
Let's derive a mathematical expression for the margin.

The margin is determined by the data points closest to the hyperplane.

These data points are known as **support vectors**.

The margin is the distance from support vectors to the hyperplane:

$$m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x}_i - b \right| \right\}$$



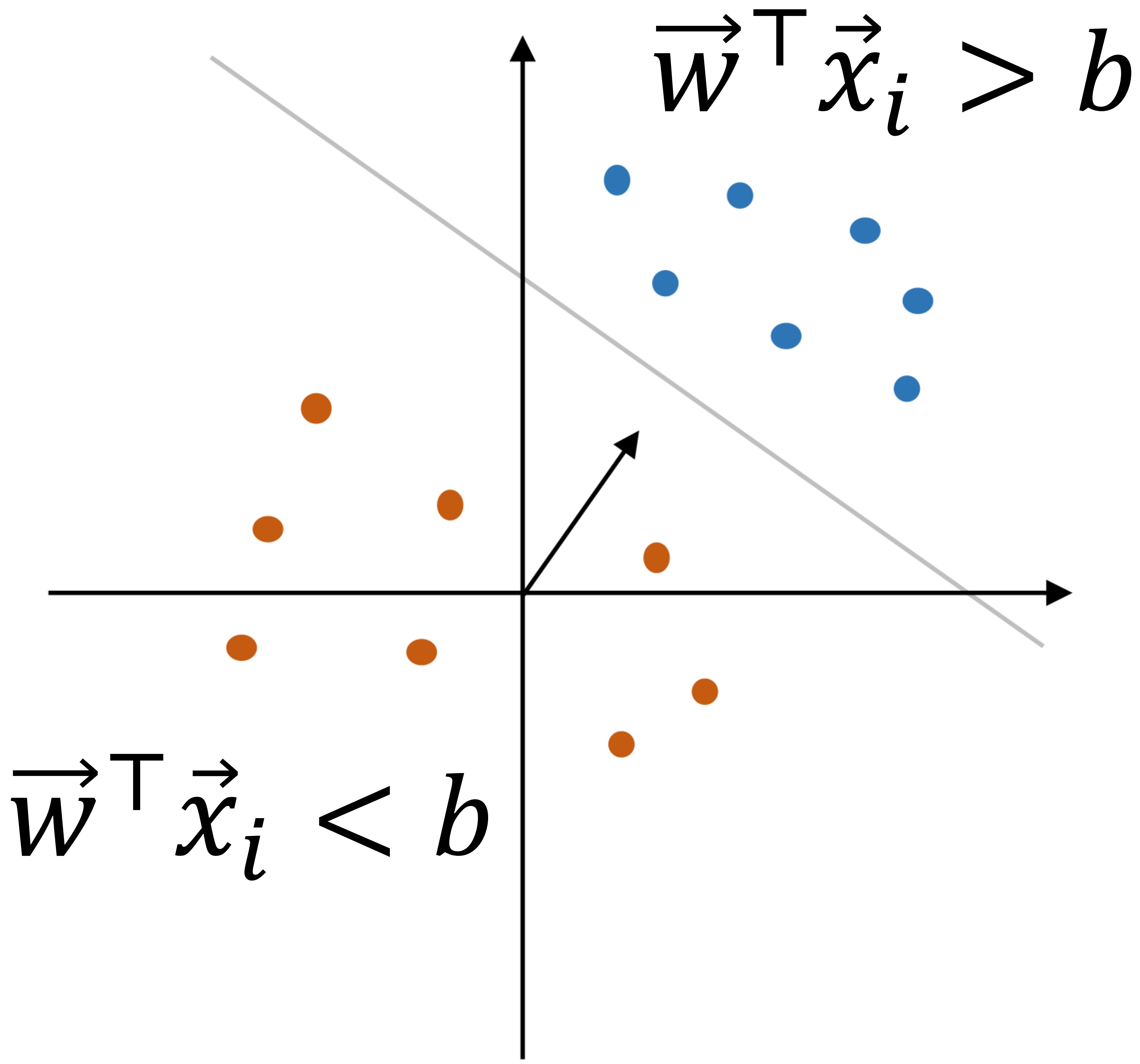
Formulation

Recall: We would like to find a hyperplane that:

(1) Separates the positive data points from the negative data points
$\vec{w}^\top \vec{x}_i > b$ for all i such that $y_i = 1$
$\vec{w}^\top \vec{x}_i < b$ for all i such that $y_i = -1$

(2) Is as far away from the data points as possible

Maximize the margin $m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x}_i - b \right| \right\}$



Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i > b \text{ for all } i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i < b \text{ for all } i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x}_i - b \right| \right\}$$

The definition of margin

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \text{ for all } i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \text{ for all } i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x}_i - b \right| \right\}$$

The definition of margin

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \ \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \ \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\frac{1}{\|\vec{w}\|_2} \left| \vec{w}^\top \vec{x}_i - b \right| \geq m$$

The margin is by definition less than or equal to the distance from any data point to the hyperplane

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \ \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \ \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\left|\vec{w}^\top \vec{x}_i - b\right| \geq m \|\vec{w}\|_2 \ \forall i$$

$$m \geq 0$$

The margin is by definition less than or equal to the distance from any data point to the hyperplane

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \ \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \ \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b \geq m\|\vec{w}\|_2 \text{ or } \vec{w}^\top \vec{x}_i - b \leq -m\|\vec{w}\|_2 \ \forall i$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \ \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \ \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b \geq m \|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$\vec{w}^\top \vec{x}_i - b \leq -m \|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

Maximize margin

$$\max_{m,\vec{w},b} m$$

subject to

$$\vec{w}^{\top}\vec{x}_i - b \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$\vec{w}^{\top}\vec{x}_i - b \leq -m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

Maximize margin

$$\max_{m,\vec{w},b} m$$

subject to

$$1 \cdot \left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$-1 \cdot \left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i\left((\alpha\vec{w})^\top \vec{x}_i - (\alpha b)\right) \geq m\|(\alpha\vec{w})\|_2$$

$$m \geq 0$$

It turns out that the scale of $\begin{pmatrix}\vec{w}\\ b\end{pmatrix}$ doesn't matter, in the sense that if $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ is a feasible solution, then $\begin{pmatrix}\alpha\vec{w}^*\\ \alpha b^*\\ m^*\end{pmatrix}$ is feasible and achieves the same objective value as $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ for any $\alpha > 0$.

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

$$\max_{m,\vec{w},b} m$$

subject to

$$\alpha y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m|\alpha|\|\vec{w}\|_2$$

$$m \geq 0$$

It turns out that the scale of $\begin{pmatrix}\vec{w}\\ b\end{pmatrix}$ doesn't matter, in the sense that if $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ is a feasible solution, then $\begin{pmatrix}\alpha\vec{w}^*\\ \alpha b^*\\ m^*\end{pmatrix}$ is feasible and achieves the same objective value as $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ for any $\alpha > 0$.

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

Therefore, without loss of generality, we can set the scale of $\vec{w}$.

We set $\|\vec{w}\|_2 = \frac{1}{m}$, so $m = \frac{1}{\|\vec{w}\|_2}$

Formulation

$$\max_{\vec{w},b} \frac{1}{\|\vec{w}\|_2}$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq 1$$

Since $x \mapsto \frac{1}{x}$ is strictly decreasing, we can apply the transformation to the objective and change the max to a min.

Formulation

$$\min_{\vec{w},b} ||\vec{w}||_2$$

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq 1$$

Want to make the objective function convex. Will see why this is useful later.

Since $x \mapsto \frac{1}{2}x^2$ is strictly increasing for $x \geq 0$, we can apply the transformation to the objective.

Formulation

$$\min_{\vec{w},b} \frac{1}{2} ||\vec{w}||_2^2$$

subject to

$$y_i\left(\vec{w}^\top \vec{x}_i - b\right) \geq 1$$

Since $x \mapsto \frac{1}{2}x^2$ is strictly increasing for $x \geq 0$, we can apply the transformation to the objective.

This is an example of a constrained optimization problem