

# Deep Learning - Theory and Practice

*IE 643  
Lecture 3*

August 6, 2024.

## 1 Perceptron

## 2 Nature of Machine Learning Tasks

- Supervised Machine Learning

## 3 Perceptron and Learning

# Perceptron

# Perceptron

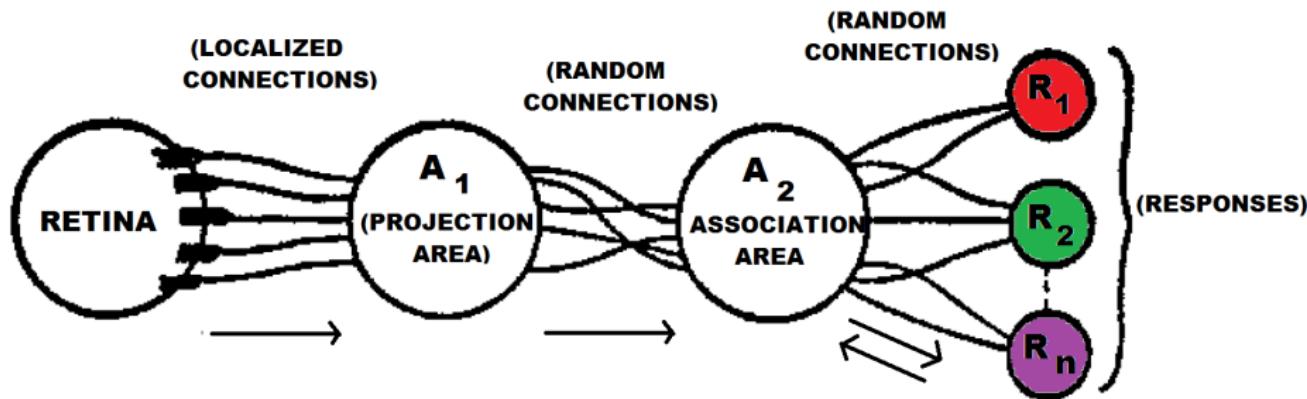
*Psychological Review*  
Vol. 65, No. 6, 1958

## THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN

F. ROSENBLATT

*Cornell Aeronautical Laboratory*

# Perceptron

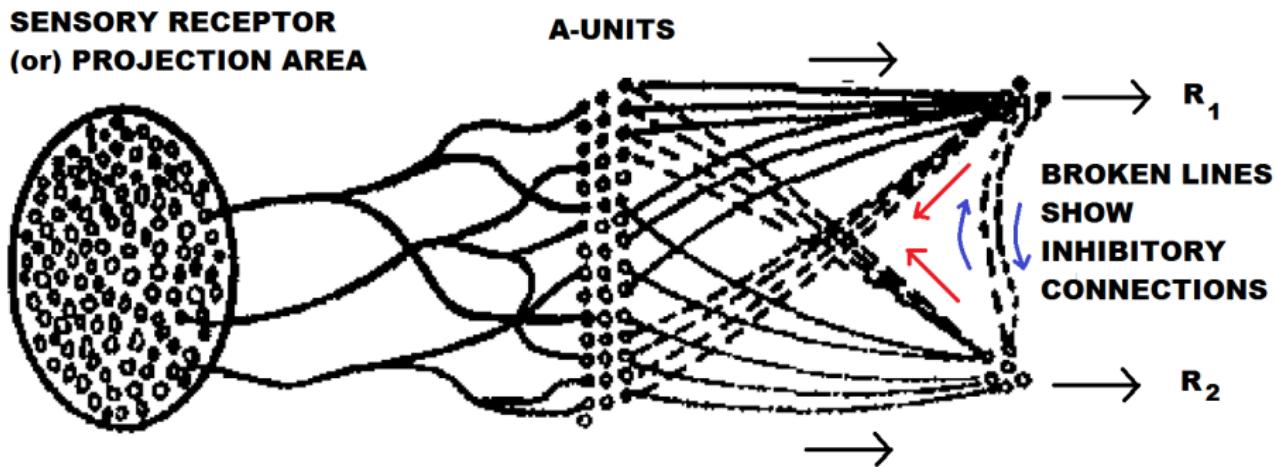


# Perceptron

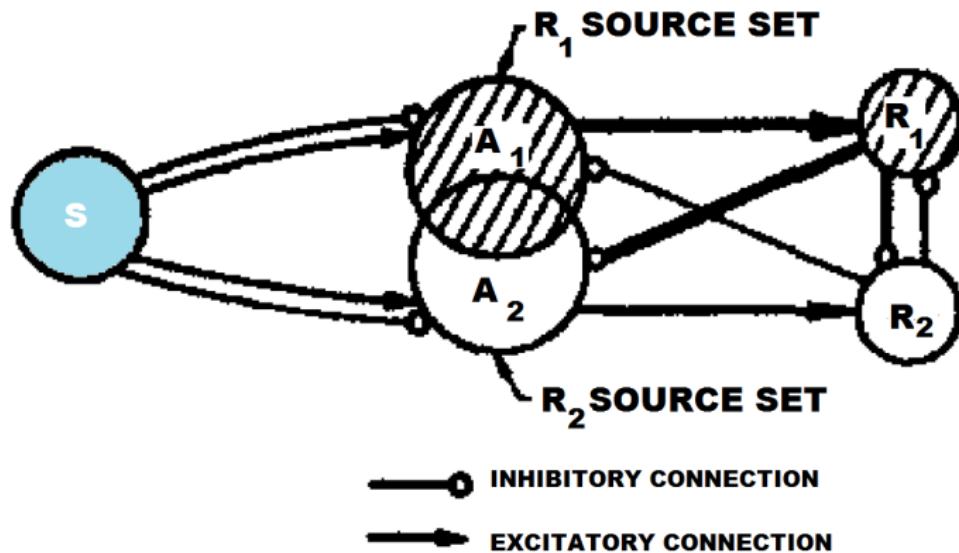
## Key Assumptions

- Stimuli which are **similar** will tend to form pathways to same sets of response cells.
- Stimuli which are **dissimilar** will tend to form pathways to different sets of response cells.
- Application of positive or negative reinforcements may facilitate or hinder the formation of connections.
- **Similarity** of stimuli is a dynamically evolving attribute.

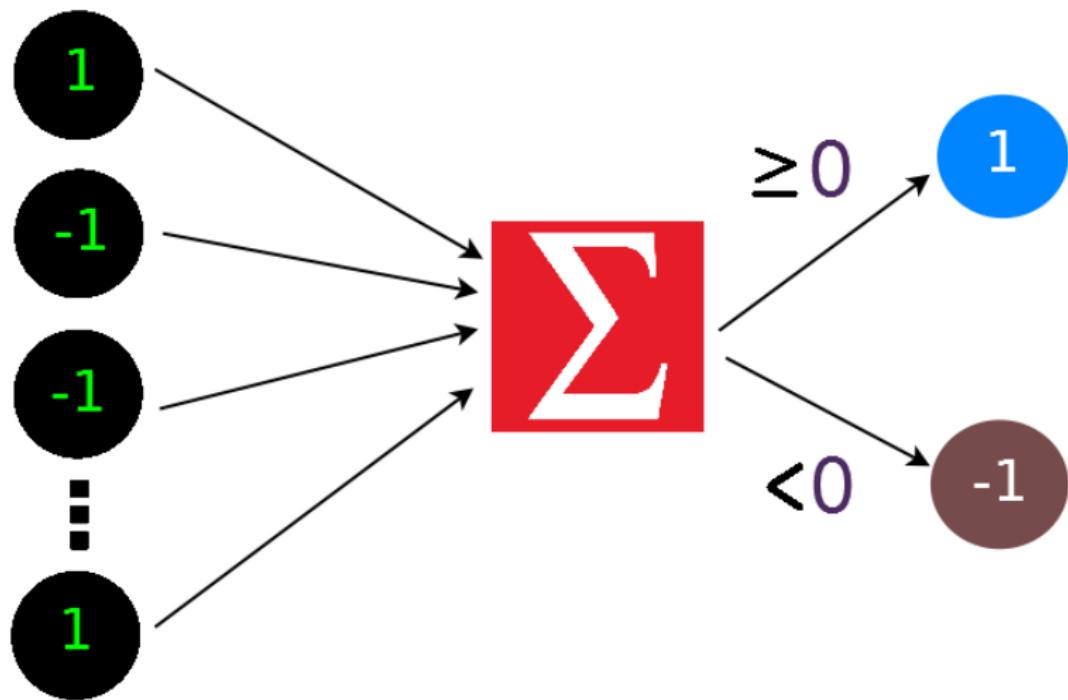
# Perceptron



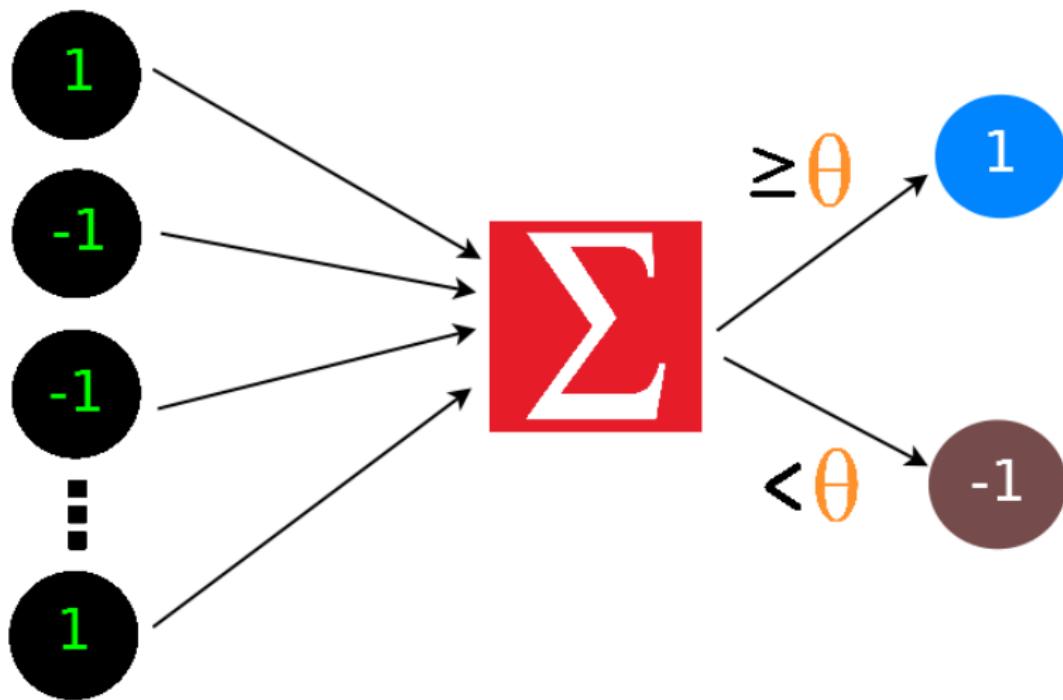
# Perceptron



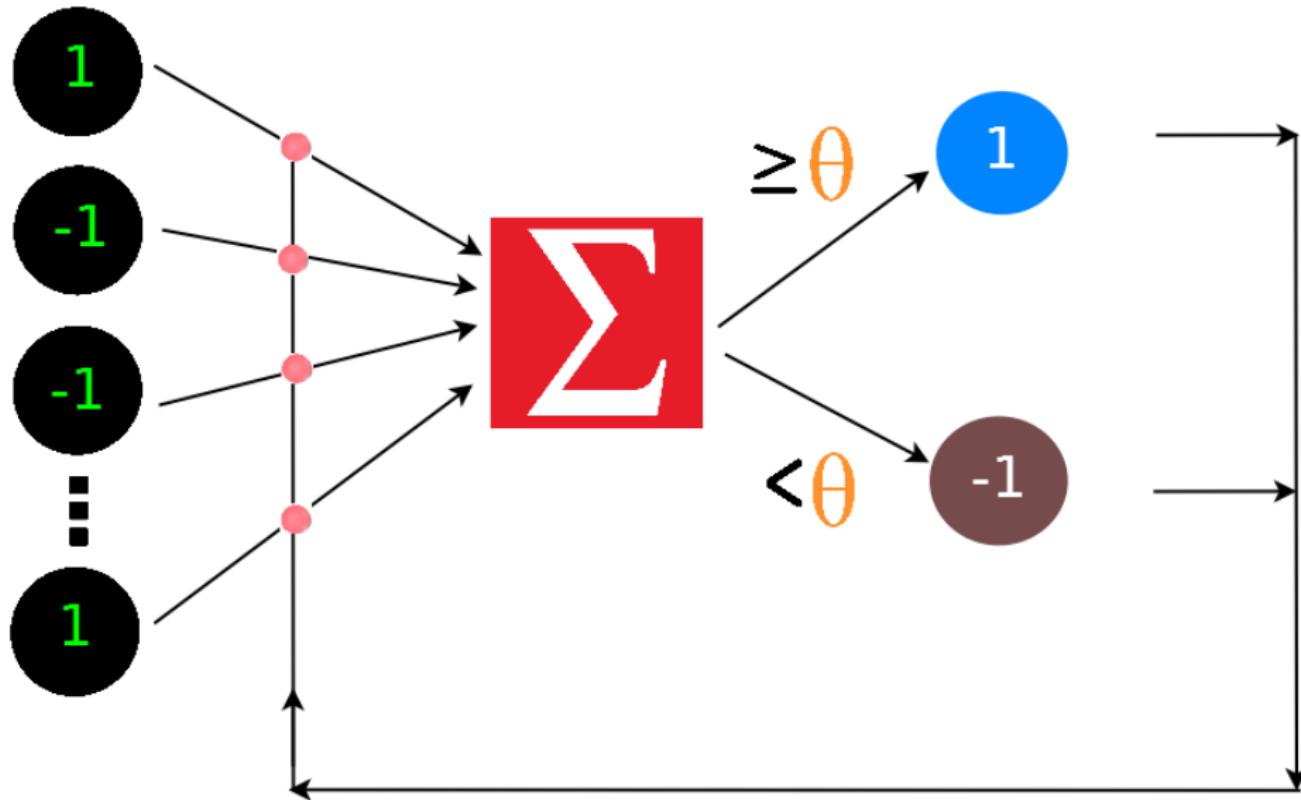
# Perceptron



# Perceptron



# Perceptron



# Nature of Machine Learning Tasks

# Nature of Machine Learning Tasks

- Supervised Learning
  - ▶ Inputs and corresponding outputs are known during learning
  - ▶ e.g. Classification (Binary, Multi-class, Multi-label)
- Unsupervised Learning
  - ▶ Input objects are not generally labeled
  - ▶ e.g. Clustering, Principal-component Analysis
- Semi-supervised Learning
  - ▶ learning from a few labeled data
- Transfer Learning
  - ▶ transferring a learned model from task  $T_1$  to  $T_2$
  - ▶ e.g. transfer from image captioning to video captioning

# Supervised Machine Learning

# Binary Classification

## Recall: e-mail Spam Classification

Genuine

Spam

The image shows an email inbox with two messages. The top message is from 'lbf.sue.com@postmaster.lbf.com' with the subject 'Hello'. A green arrow points to it with the word 'Genuine'. The bottom message is from 'Offers.genuine@replay-to-me.com' with the subject 'Can't see the picture? Select "Always deliver images" or view the message in your browser.' A red arrow points to it with the word 'Spam'. Below the inbox, there is a snippet of a webpage for '3 BHK Villas' near an IT park, featuring a modern villa image and navigation tabs for Overview, Location, Specifications, Floor Plans, and For Sale.

book service, offering unlimited uses, and books by your favorite

hd that we are offering a one-month voracious readers.

th a subscription to bd.

NEL GAIMAN  
AMERICAN GODS

THE ALCHEMIST

XS Real Villa Boutique

Time to buy now, No Pre-EMI until Possession!!!

- Property type : 3 BHK Villas
- Property Price : Rs. 1.30 Cr (All Inclusive)
- Plot Area : 2364 - 2371 Sq. Ft

Enquire

# Binary Classification

- **Input:** e-mail messages
- **Output:** Spam/Not spam

# Binary Classification

- **Input:** e-mail messages  $\Rightarrow$  some feature space
- **Output:** Spam/Not spam  $\Rightarrow \{+1, -1\}$
- Generally many input/output pairs are given for learning the machine learning model.

# Binary Classification

## Feature Extraction

W Dark matter - Wikipedia, the free encyclopedia [en.wikipedia.org/w/index.php?title=Dark\\_matter&oldid=1132400000](https://en.wikipedia.org/w/index.php?title=Dark_matter&oldid=1132400000)

**Galaxy clusters and gravitational lensing [edit]**

Galaxy clusters are especially important for dark matter studies since their masses can be estimated in three independent ways:

- From the scatter in radial velocities of the galaxies within them (as in Zwicky's early observations, with much larger modern samples).
- From X-rays emitted by very hot gas within the clusters. The temperature and density of the gas can be estimated from the energy and flux of the X-rays, hence the gas pressure; assuming pressure and gravity balance, this enables the mass profile of the cluster to be derived. Many of the experiments of the Chandra X-ray Observatory use this technique to independently determine the mass of clusters. These observations generally indicate a ratio of baryonic to total mass approximately 12-15 percent, in reasonable agreement with the Planck spacecraft cosmic average of 15.5 - 16 percent.<sup>[37]</sup>
- From their gravitational lensing effects on background objects, usually more distant galaxies. This is observed as "strong lensing" (multiple images) near the cluster core, and weak lensing (shape distortions) in the outer parts. Several large Hubble projects have used this method to measure cluster masses.

Generally these three methods are in reasonable agreement, that clusters contain much more matter than the visible galaxies and gas.

A gravitational lens is formed when the light from a more distant source (such as a quasar) is "bent" around a massive object (such as a cluster of galaxies) between the source object and the observer. The process is known as gravitational lensing.

The galaxy cluster Abell 2029 is composed of thousands of galaxies enveloped in a cloud of hot gas, and an amount of dark matter equivalent to more than  $10^{14}$  Suns. At the center of this cluster is an enormous, elliptically shaped galaxy that is thought to have been formed from the mergers of many smaller galaxies.<sup>[38]</sup> The measured orbital velocities of galaxies within galactic clusters have been found to be consistent with dark matter observations.

Another important tool for future dark matter observations is gravitational lensing. Lensing relies on the effects of general relativity to predict masses without relying on dynamics, and so is a completely independent means of measuring the dark matter. Strong lensing, the observed distortion of background galaxies into arcs when the light passes through a gravitational lens, has been observed around a few distant clusters including Abell 1689 (pictured right).<sup>[39]</sup> By measuring the distortion geometry, the mass of the cluster causing the phenomena can be obtained. In the dozens of cases where this has been done, the mass-to-light ratios obtained correspond to the dynamic measurements of clusters.<sup>[40]</sup>

Weak gravitational lensing looks at minute distortions of galaxies observed in vast galaxy surveys due to foreground objects through statistical analyses. By examining the apparent shear of the adjacent background galaxies, astrophysicists can characterize the mean distribution of dark matter by statistical means and have found mass-to-light ratios that correspond to dark matter predictions by other large-scale structure measurements.<sup>[41]</sup> The correspondence of the two gravitational lens techniques to other dark matter measurements has convinced almost all astrophysicists that dark matter actually exists as a major component of the universe's composition.

The most direct observational evidence to date for dark matter is in a system known as the Bullet Cluster. In most regions of the universe, dark matter and visible material are found together,<sup>[42]</sup> as expected because of their mutual gravitational attraction. In the Bullet Cluster, a collision between two galaxy clusters appears to have caused a separation of dark matter and baryonic matter. X-ray observations show that much of the baryonic matter (in the form of  $10^7$ - $10^8$  Kelvin<sup>[43]</sup> gas, or plasma) in the system is concentrated in the center of the system. Electromagnetic interactions between passing gas particles caused them to slow down and settle near the point of impact. However, weak gravitational lensing observations of the same system show that much of the mass resides outside of the central region of baryonic gas. Because dark matter does not interact by electromagnetic forces, it would not have been slowed in the same way as the X-ray visible gas, so the dark matter components of the two clusters passed through each other without slowing down substantially. This accounts for the separation. Unlike the galactic rotation curves, this evidence for dark matter is independent of the details of Newtonian gravity, so it is claimed to be direct evidence of the existence of dark matter.<sup>[43]</sup>

Another galaxy cluster, known as the Train Wreck Cluster/Abell 520, appears to have an unusually massive and dark core containing few of the cluster's galaxies, which presents problems for standard dark matter models.<sup>[44]</sup>

This may be explained by the dark core actually being a long, low-density dark matter filament (containing few galaxies) along the line of sight, projected onto the cluster core.<sup>[45]</sup>

The observed behavior of dark matter in clusters constrains whether and how much dark matter scatters off other dark matter particles, quantified as its

**Paraphasis**

**Capitalized First Letter**

**Numbers**

**Quote**

**Superscripts**

**Bigram**

**Trigram**

**Hyphenated**

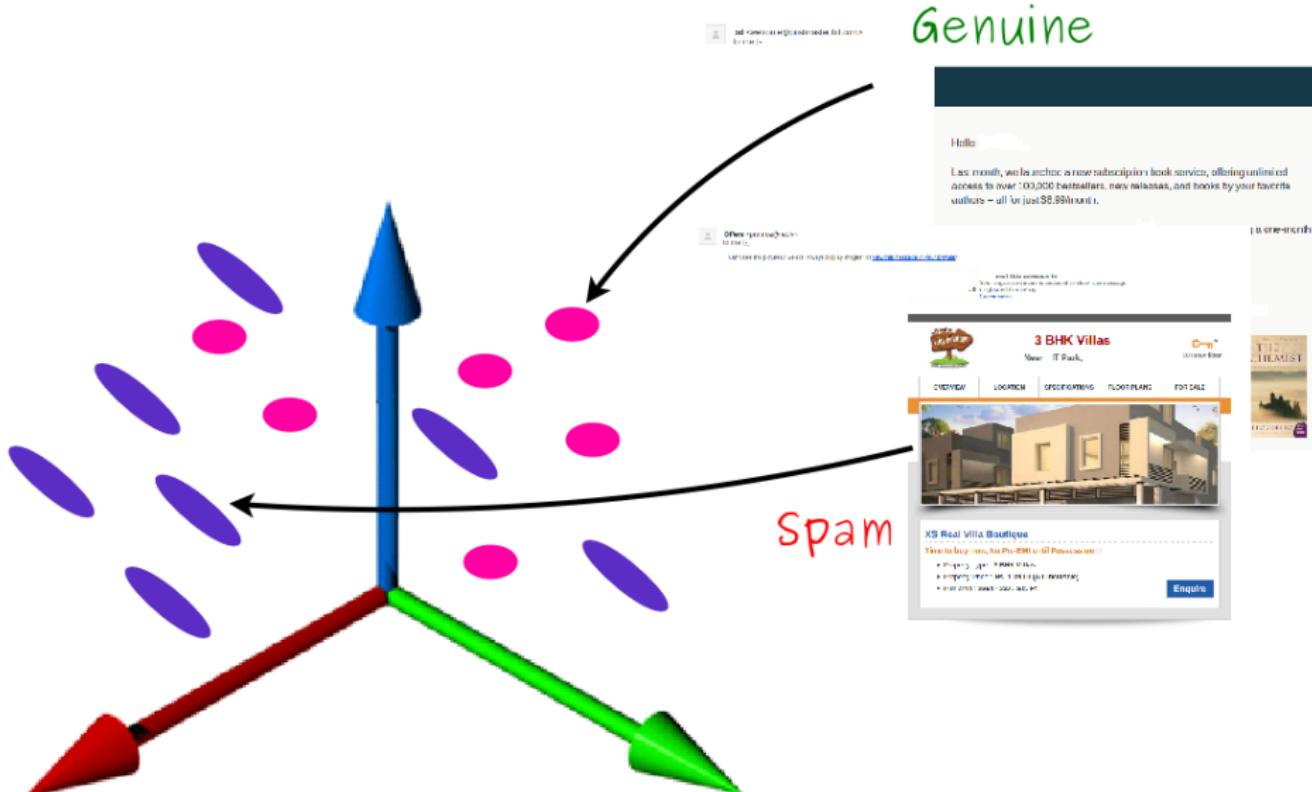
**Name**






# Binary Classification

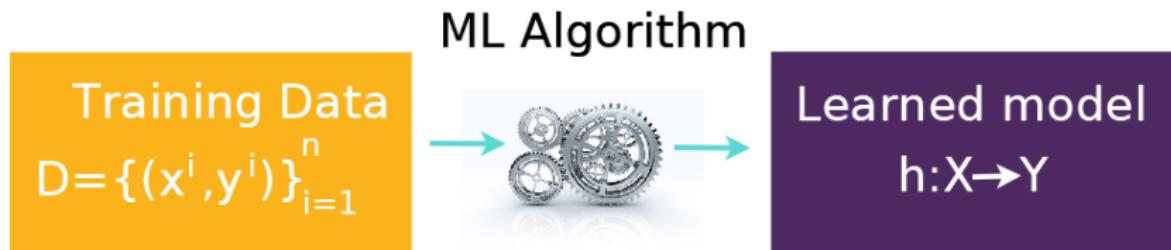
## Feature Extraction



# Binary Classification

- **Input:** e-mail messages  $\implies$  some feature space  $\subseteq \mathbb{R}^d$ .
  - ▶  $x \in \mathcal{X} \subseteq \mathbb{R}^d$
- **Output:** Spam/Not spam  $\implies \{+1, -1\}$ 
  - ▶  $y \in \mathcal{Y} = \{+1, -1\}$
- Generally  $n$  input/output pairs  $\{(x^i, y^i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$  are given for learning the machine learning model.
- $D = \{(x^i, y^i)\}_{i=1}^n$  called the training data.

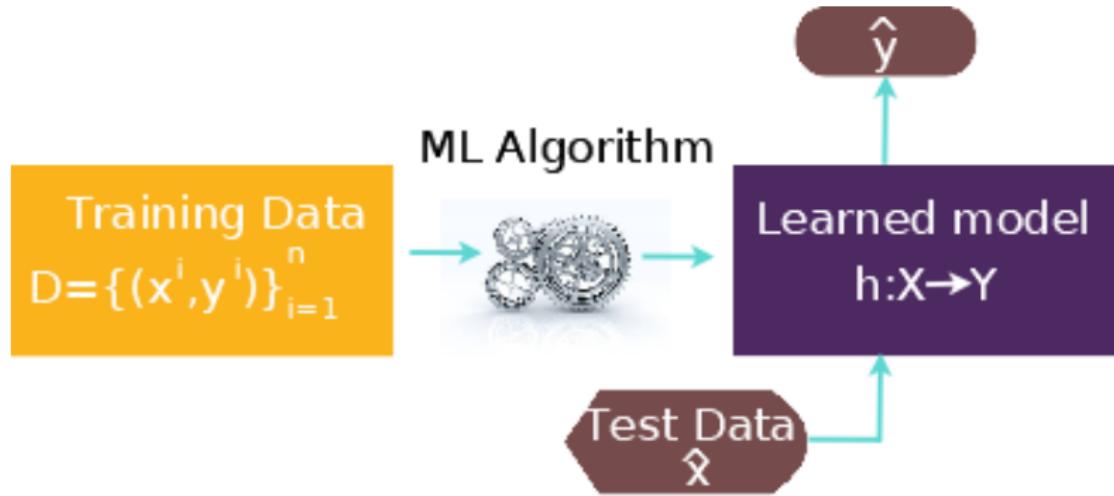
# General Nature of a Supervised Machine Learning Task



## Training

- **Input:** Training data  $D = \{(x^i, y^i)\}_{i=1}^n$
- **Aim:** Learn a model  $h : \mathcal{X} \rightarrow \mathcal{Y}$

# General Nature of a Supervised Machine Learning Task



## Training

- **Input:** Training data  $D = \{(x^i, y^i)\}_{i=1}^n$
- **Aim:** Learn a model  $h : \mathcal{X} \rightarrow \mathcal{Y}$

## Testing

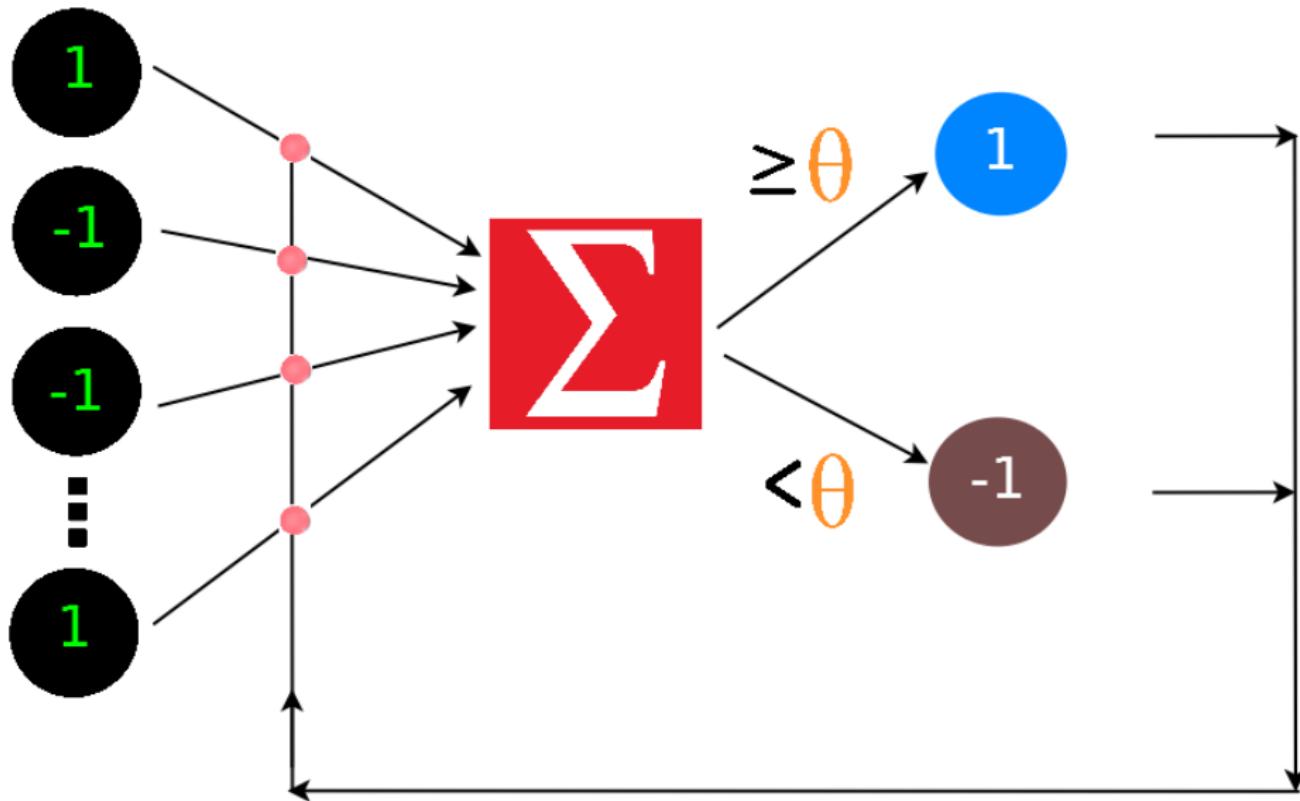
- Given  $\hat{x}$ , predict  $\hat{y} = h(\hat{x})$

# A sample data set for email classification

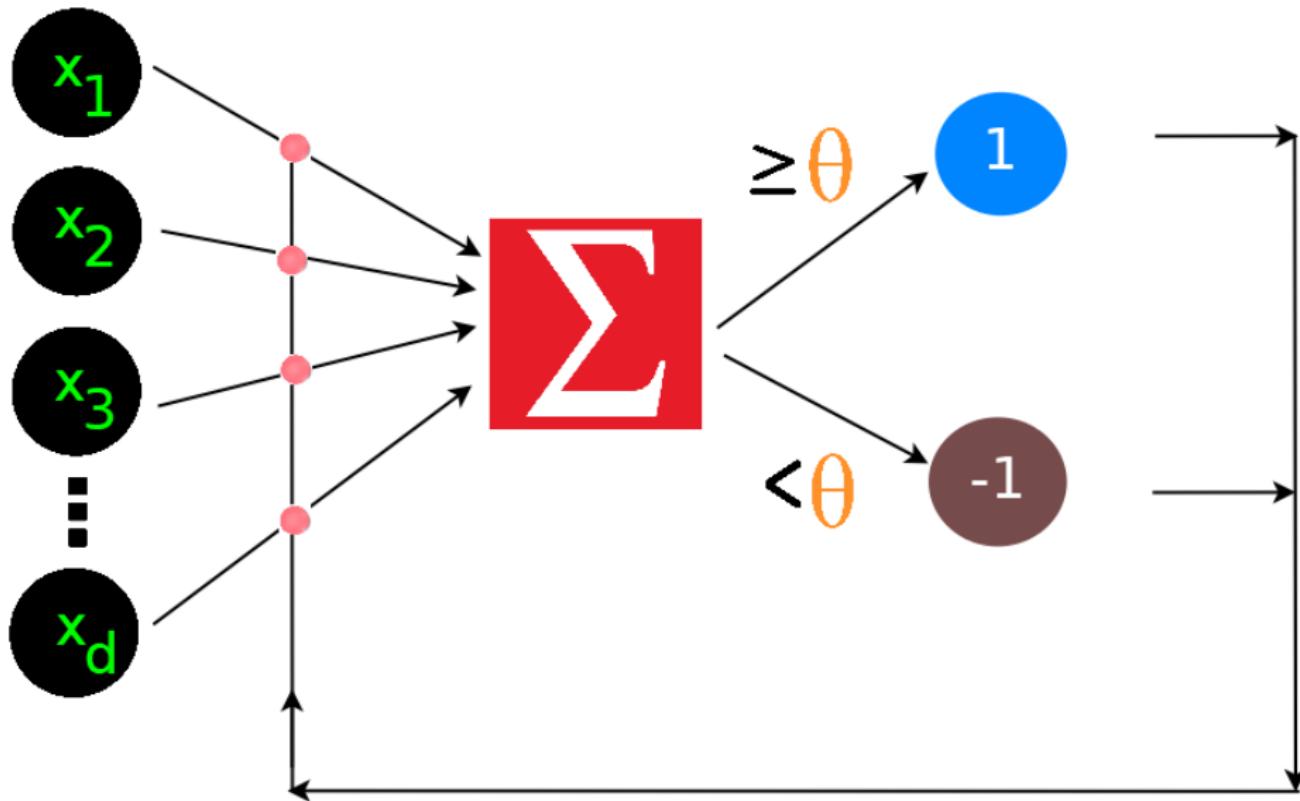
spam	to_multiple	from	cc	sent_email	image	attach	dollar	winner	inherit	viagra	password	num_char
0	0	1	0	0	0	0	0	no	0	0	0	11.37
0	0	1	0	0	0	0	0	no	0	0	0	10.504
0	0	1	0	0	0	0	0	no	0	0	0	13.256
0	0	1	0	0	0	0	0	no	0	0	2	1.231
1	0	1	0	0	0	0	0	no	0	0	0	5.108
1	0	1	4	0	0	2	0	no	0	0	0	0.631
1	0	1	0	0	0	0	1	yes	0	0	0	0.963
1	0	1	0	0	0	0	2	no	0	0	0	2.182

# Perceptron and Learning

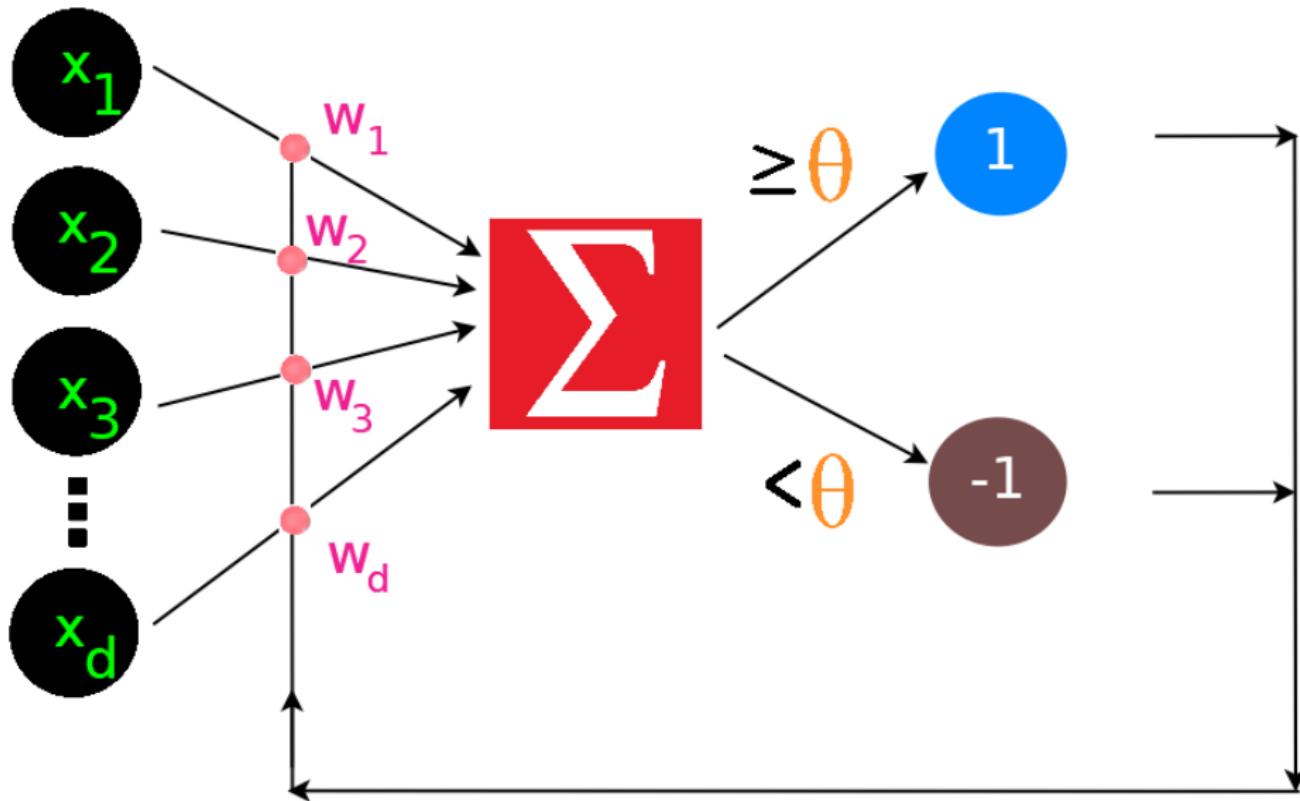
# Perceptron



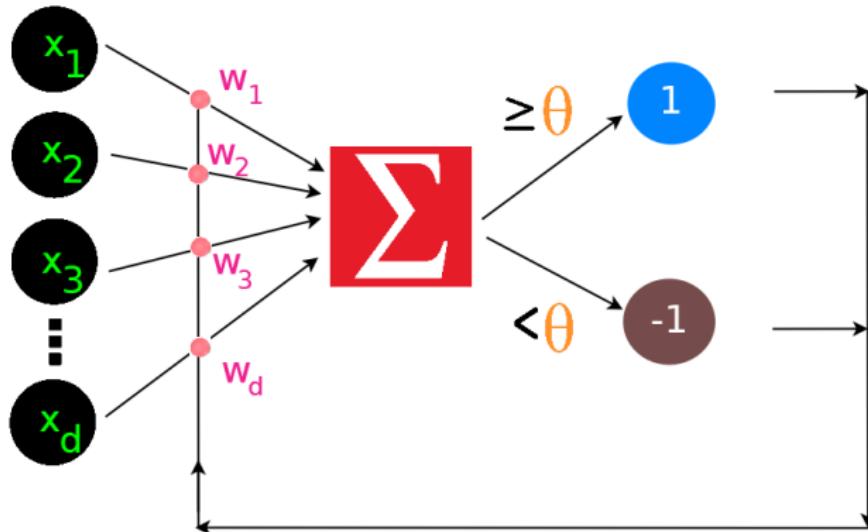
# Perceptron



# Perceptron

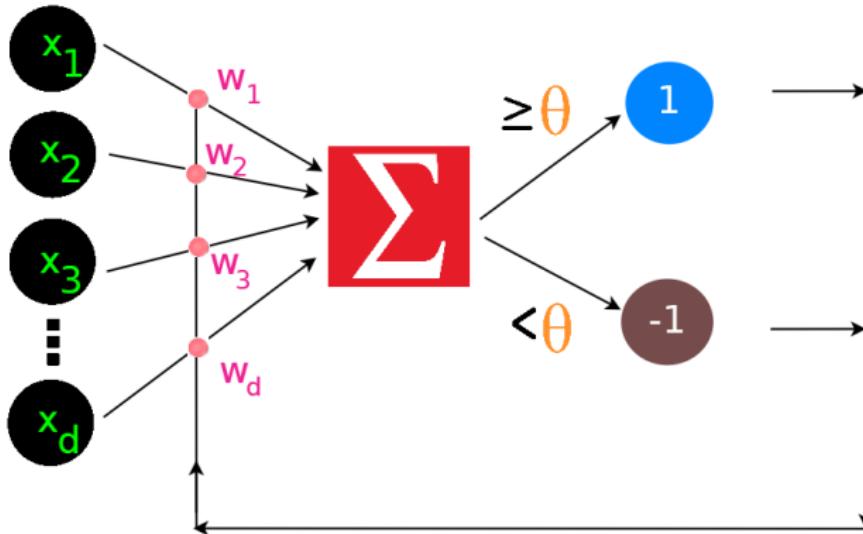


# Perceptron



- Input is of the form  $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ .
- We associate weights  $w = (w_1, w_2, \dots, w_d) \in \mathbb{R}^d$  to the connections.
- **Prediction Rule:**
  - ▶  $\sum_{i=1}^d w_i x_i \geq \theta \implies \text{predict } 1$ .
  - ▶  $\sum_{i=1}^d w_i x_i < \theta \implies \text{predict } -1$ .

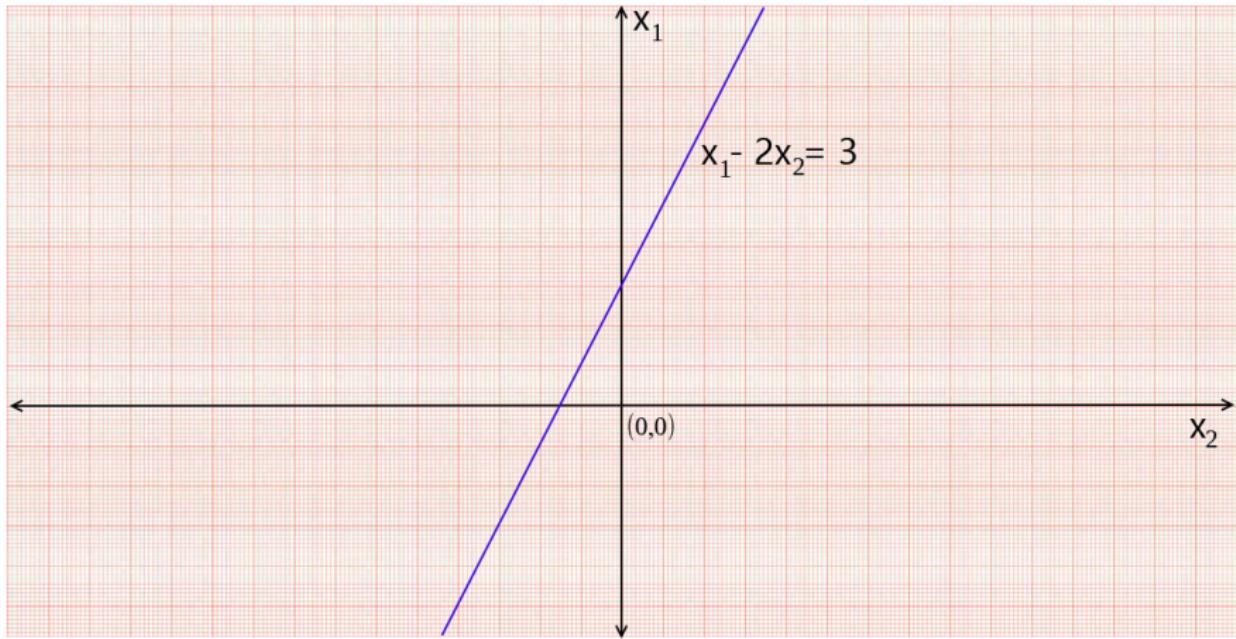
# Perceptron



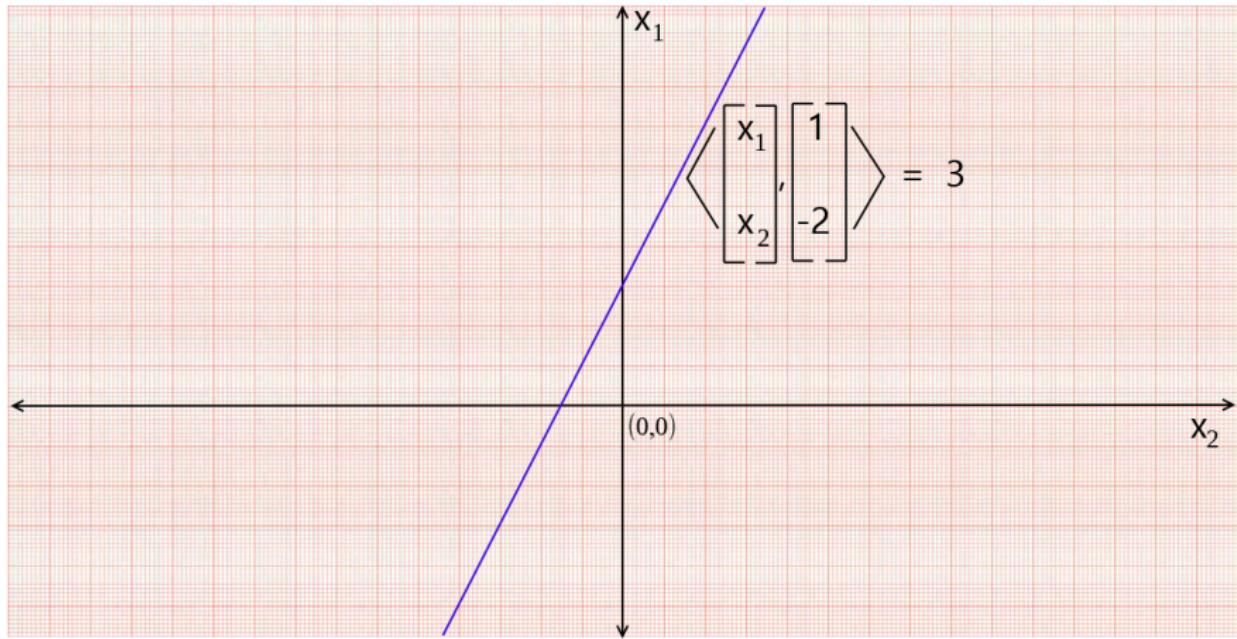
- Input is of the form  $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ .
- We associate weights  $w = (w_1, w_2, \dots, w_d) \in \mathbb{R}^d$  to the connections.
- **Prediction Rule:**
  - ▶  $\langle w, x \rangle \geq \theta \implies \text{predict } 1$ .
  - ▶  $\langle w, x \rangle < \theta \implies \text{predict } -1$ .

**Note:**  $\langle w, x \rangle = \sum_{i=1}^d w_i x_i$  denotes inner product between  $w$  and  $x$ .

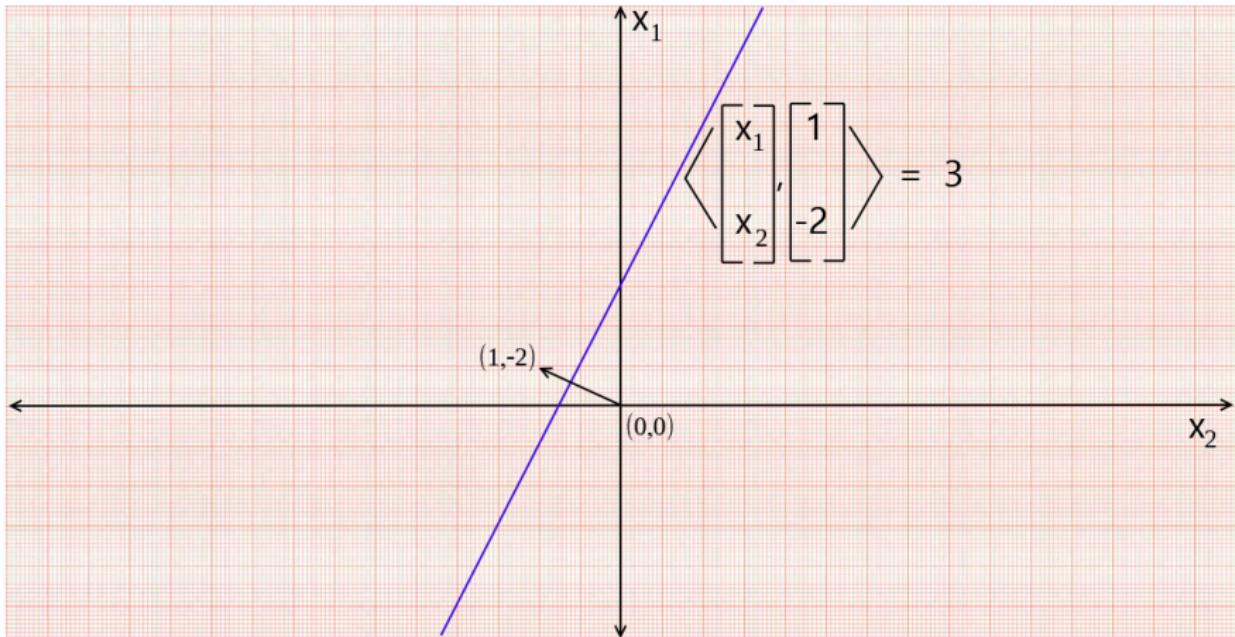
# Perceptron - Geometry



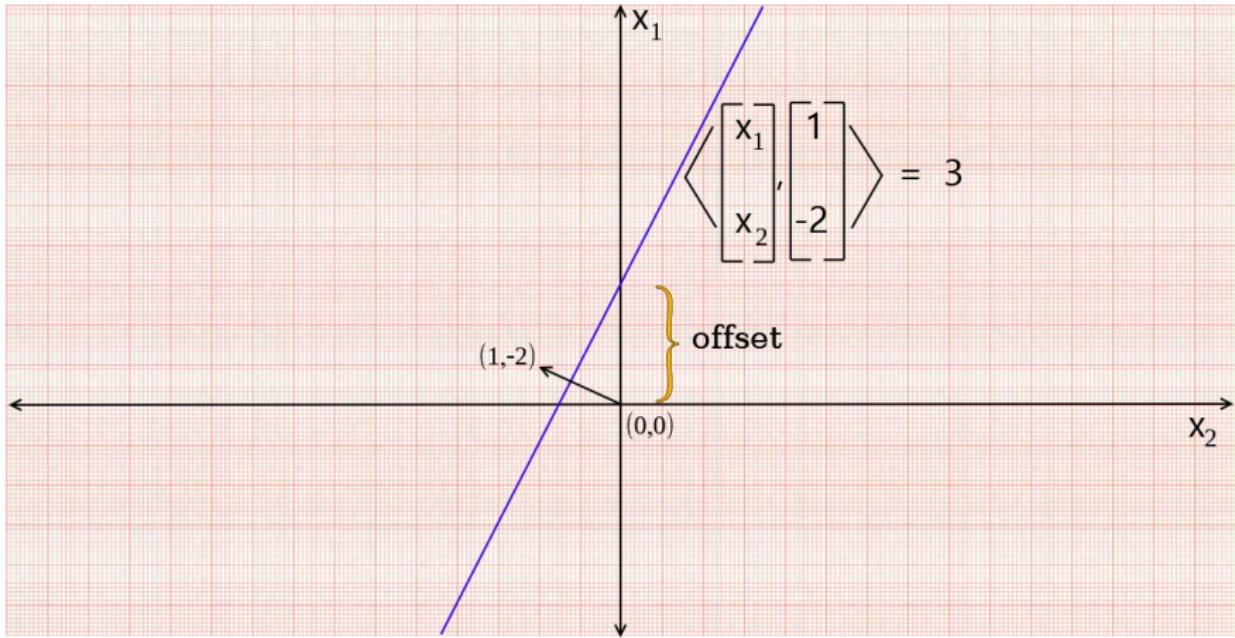
# Perceptron - Geometry



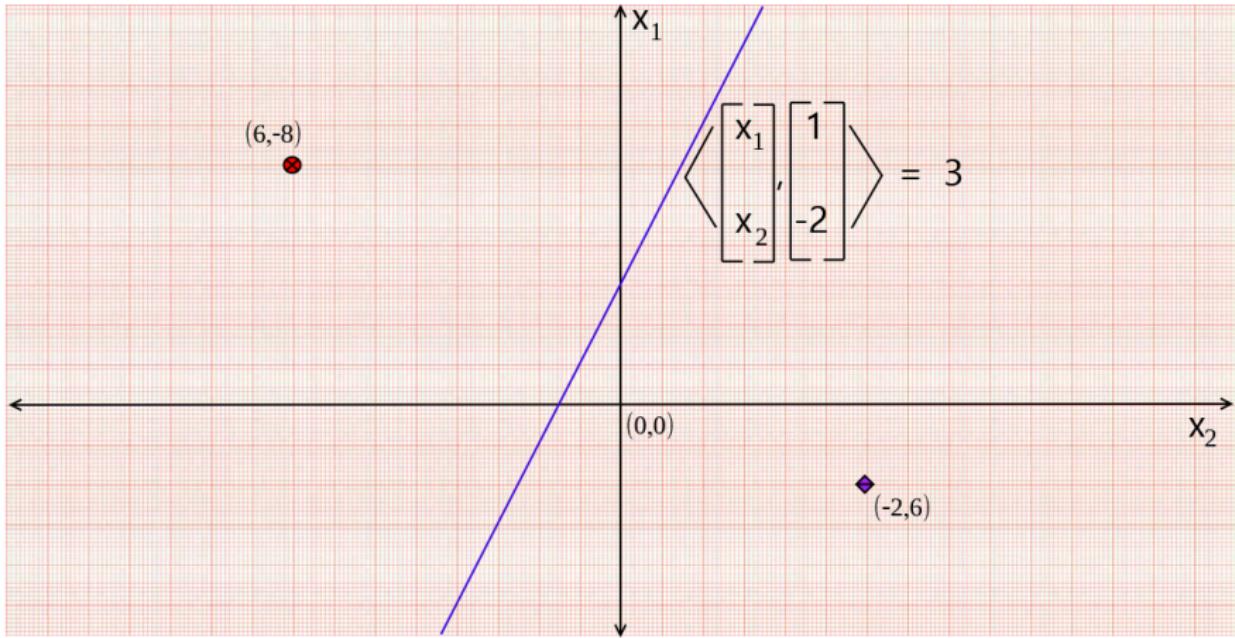
# Perceptron - Geometry



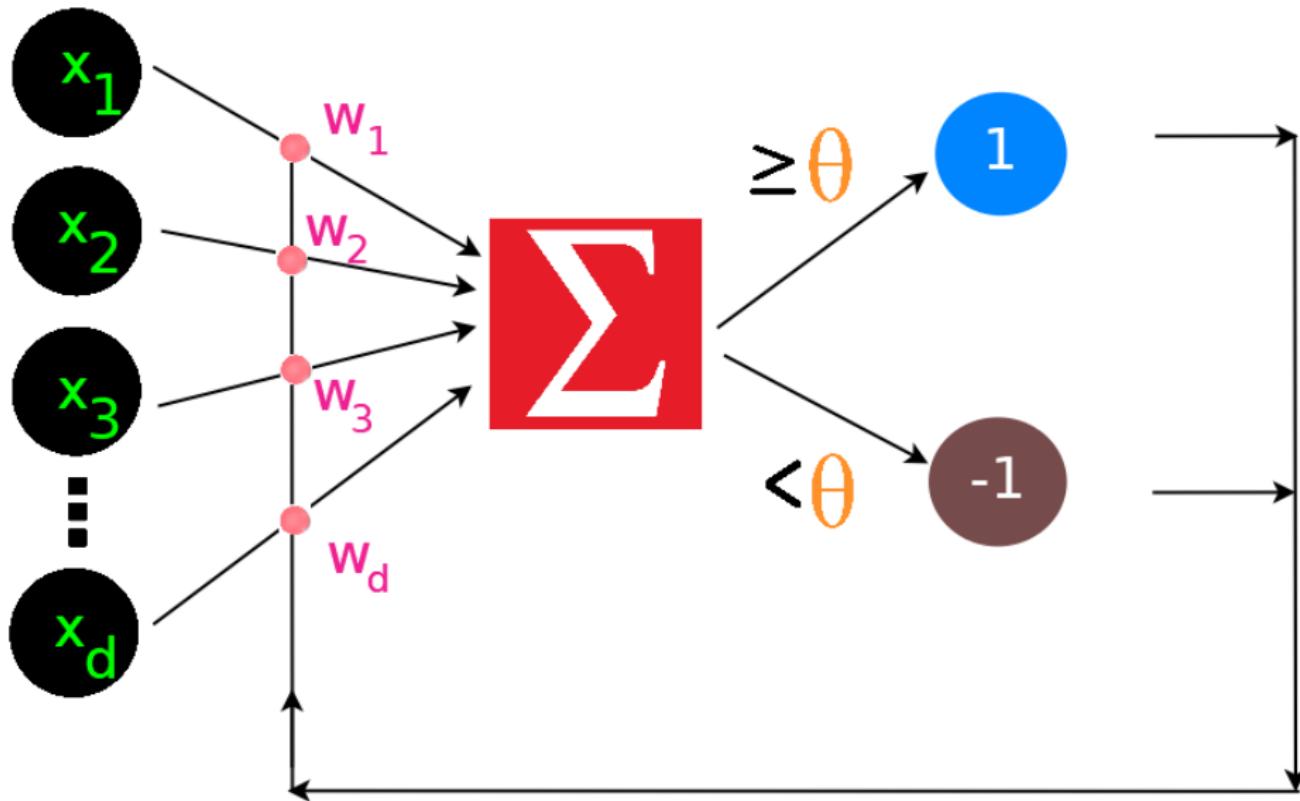
# Perceptron - Geometry



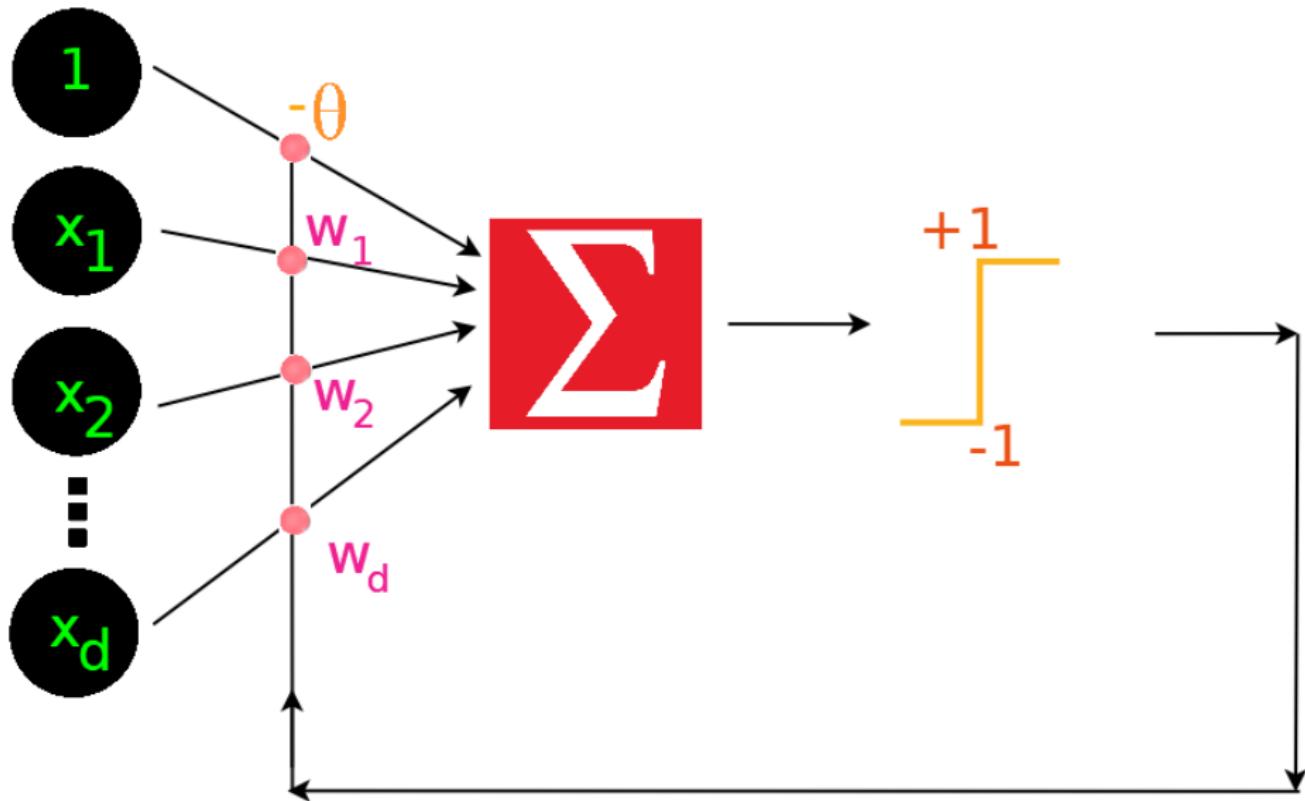
# Perceptron - Geometry



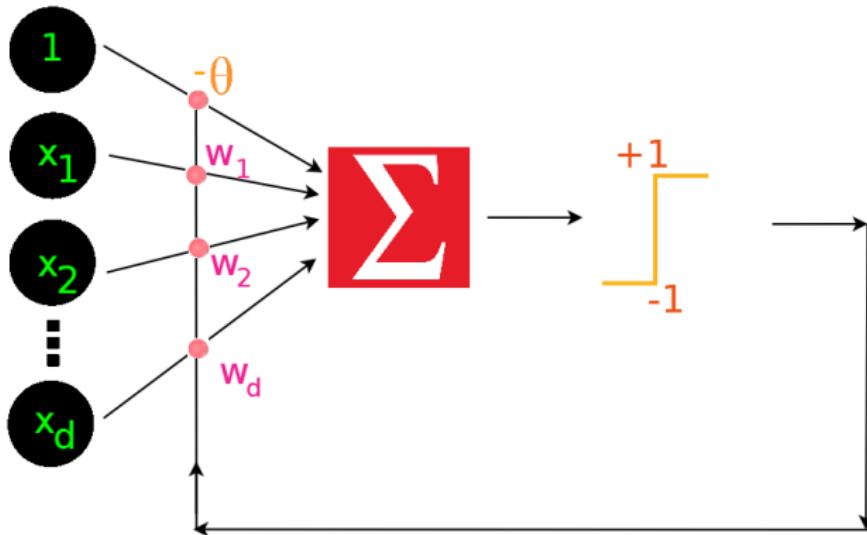
# Perceptron



# Perceptron

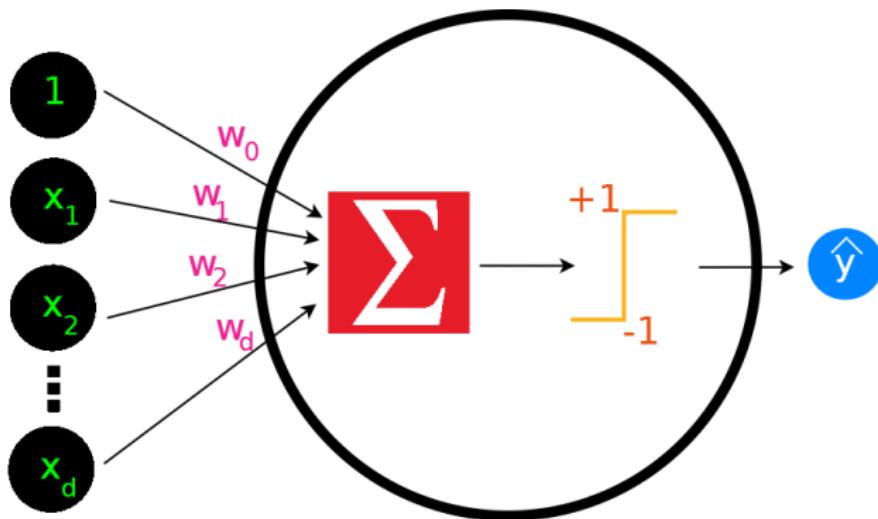


# Perceptron



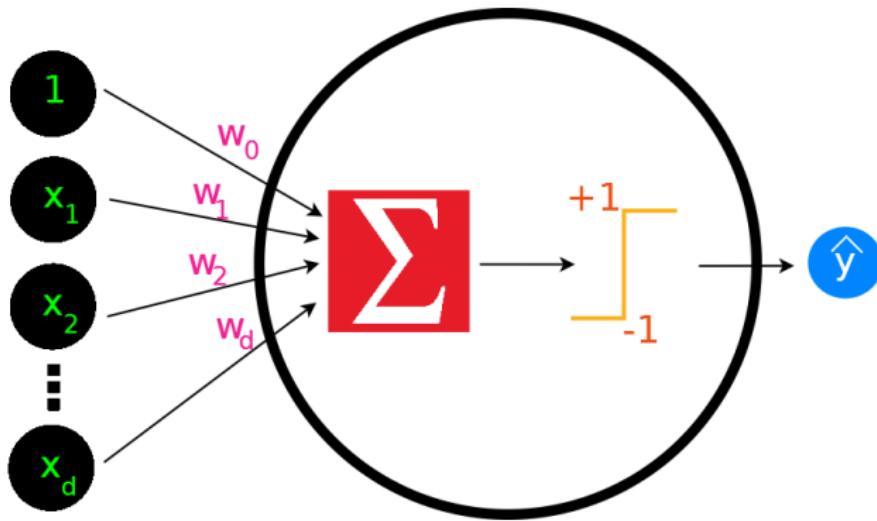
- Input is of the form  $\tilde{x} = (1, x) = (1, x_1, x_2, \dots, x_d) \in \mathbb{R}^{d+1}$ .
- We associate weights  $\tilde{w} = (-\theta, w) = (-\theta, w_1, w_2, \dots, w_d) \in \mathbb{R}^{d+1}$  to the connections.
- **Prediction Rule:**
  - ▶  $\langle w, x \rangle - \theta \geq 0 \implies \text{predict } 1$ .
  - ▶  $\langle w, x \rangle - \theta < 0 \implies \text{predict } -1$ .

# Perceptron - Data Perspective



- **Actual Input:**  $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$
- **Transformed Input:**  $\tilde{x} = (1, x_1, x_2, \dots, x_d) \in \mathbb{R}^{d+1}$
- **Actual Weights:**  $w = (w_1, w_2, \dots, w_d) \in \mathbb{R}^d, \theta \in \mathbb{R}$
- **Transformed Weights:**  $\tilde{w} = (w_0, w_1, w_2, \dots, w_d) \in \mathbb{R}^{d+1}$

# Perceptron - Data Perspective



Equivalently we might use:

- **Transformed Input:**  $\tilde{x} = (x_1, x_2, \dots, x_d, 1) \in \mathbb{R}^{d+1}$
- **Transformed Weights:**  $\tilde{w} = (w_1, w_2, \dots, w_d, w_0) \in \mathbb{R}^{d+1}$

# Perceptron - Data Perspective

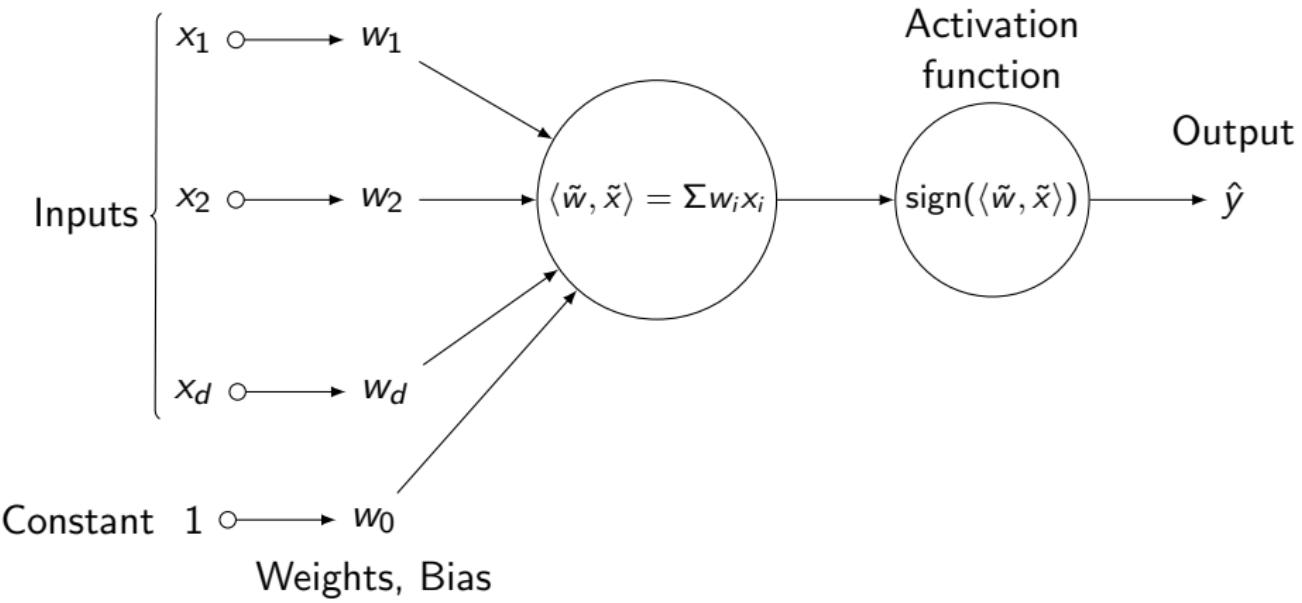
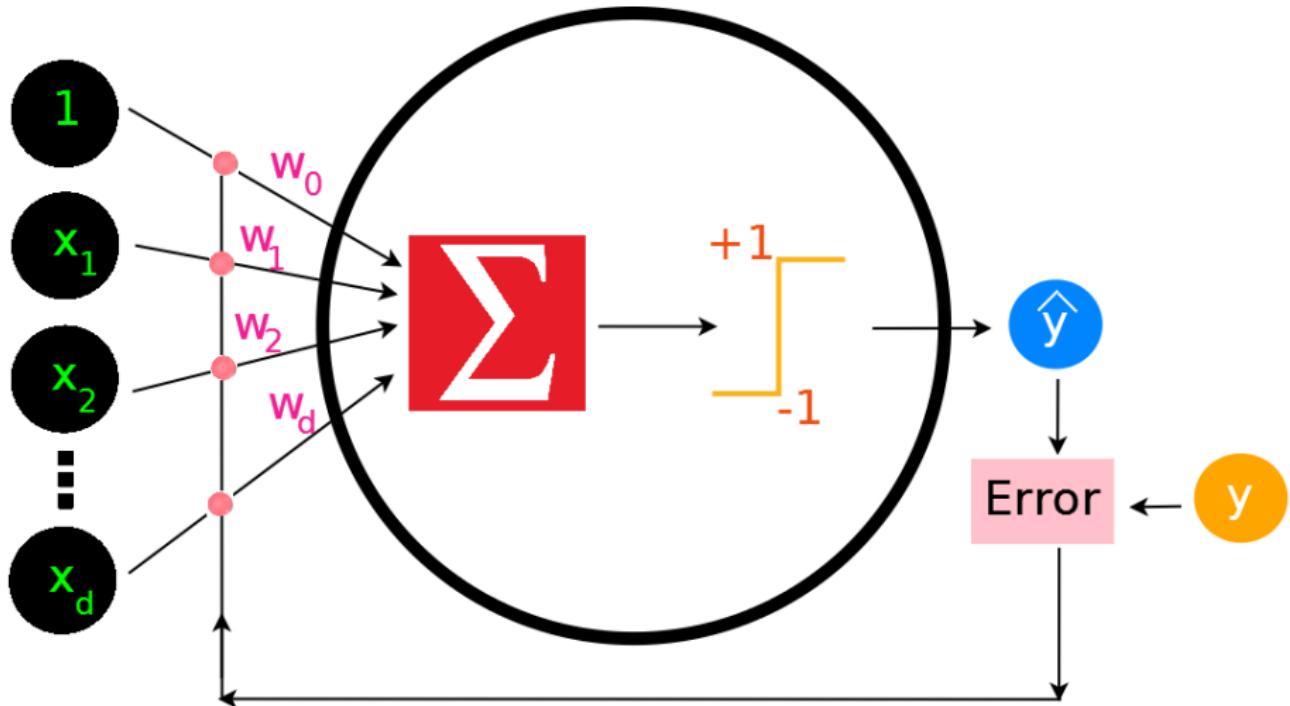


Figure: Perceptron unit

# Perceptron - Data Perspective



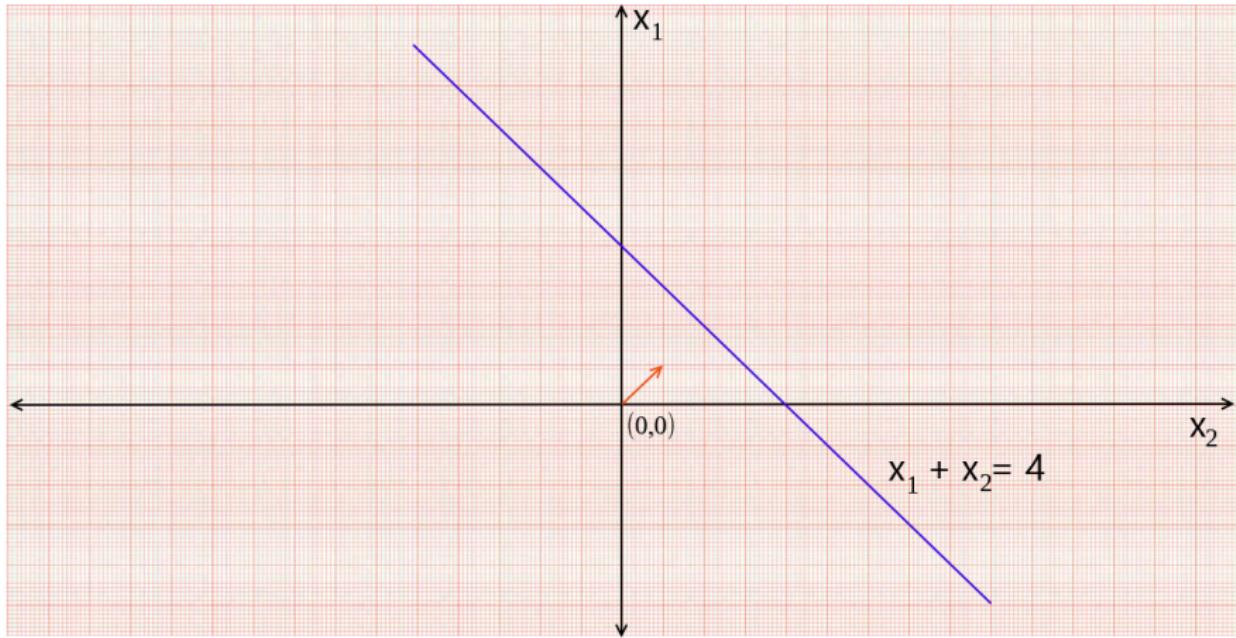
- **Input:** data point  $x = (x_1, x_2, \dots, x_d)$ , label  $y \in \{+1, -1\}$ .

# Perceptron - Training

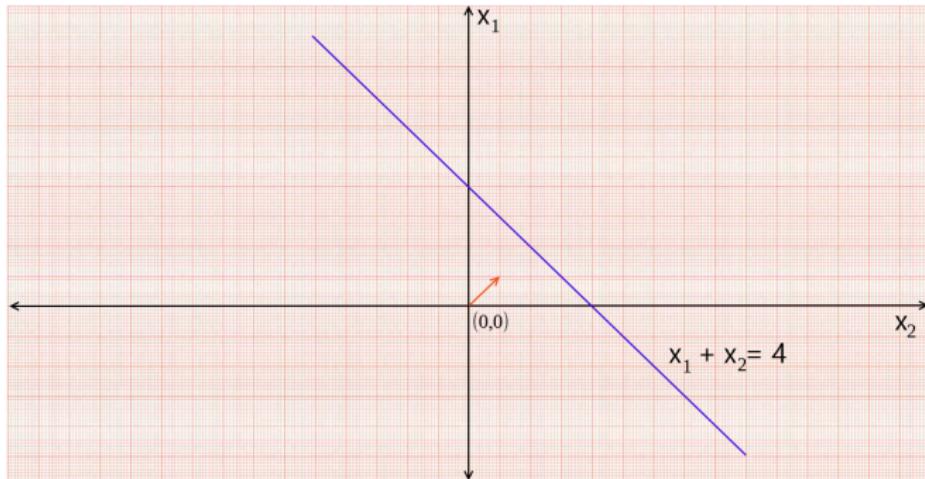
## Perceptron Training Procedure

```
1:  $\tilde{w}^1 = 0$ 
2: for  $t \leftarrow 1, 2, 3, \dots$  do
3:   receive  $(x^t, y^t)$ ,  $x^t \in \mathbb{R}^d$ ,  $y^t \in \{+1, -1\}$ .
4:   Transform  $x^t$  into  $\tilde{x}^t = (x^t, 1) \in \mathbb{R}^{d+1}$ .
5:    $\hat{y} = \text{Perceptron}(\tilde{x}^t; \tilde{w}^t)$ 
6:   if  $\hat{y} \neq y^t$  then
7:      $\tilde{w}^{t+1} = \tilde{w}^t + y^t \tilde{x}^t$ 
8:   else
9:      $\tilde{w}^{t+1} = \tilde{w}^t$ 
```

# Perceptron Update Rule - Geometric Idea

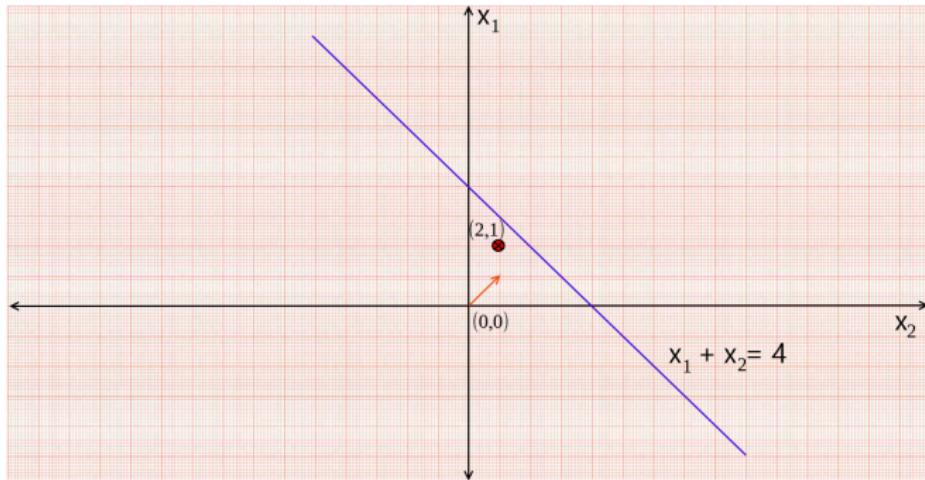


# Perceptron Update Rule - Geometric Idea



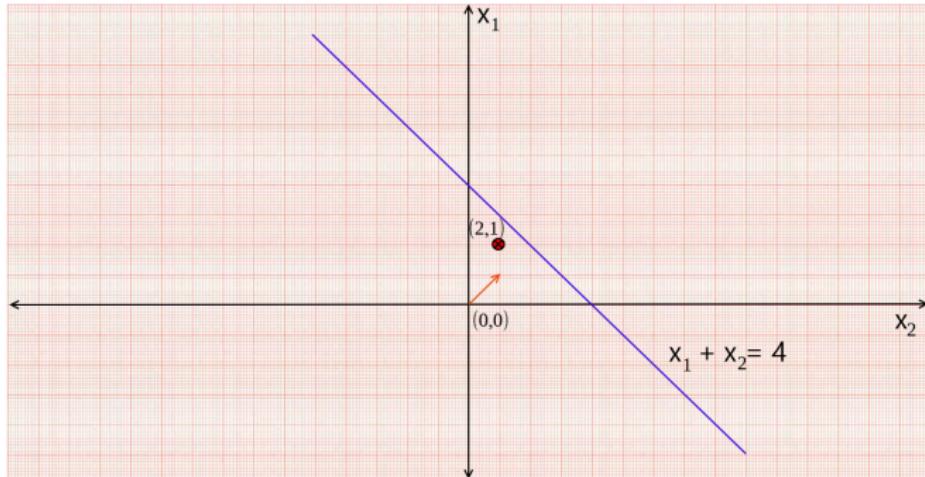
- The current weights are  $\tilde{w}^t = (1, 1, -4)$ .

# Perceptron Update Rule - Geometric Idea



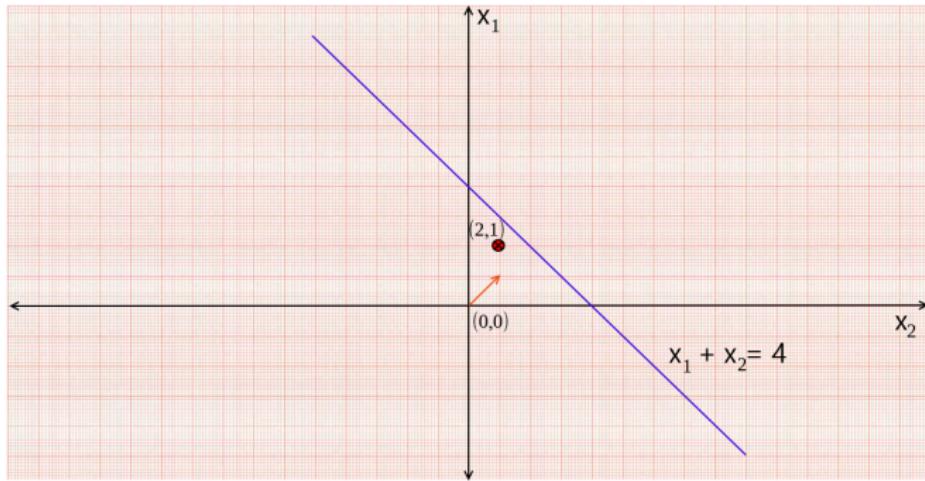
- Suppose a new point  $x^t = (2, 1)$ ,  $y^t = 1$  arrives.

# Perceptron Update Rule - Geometric Idea



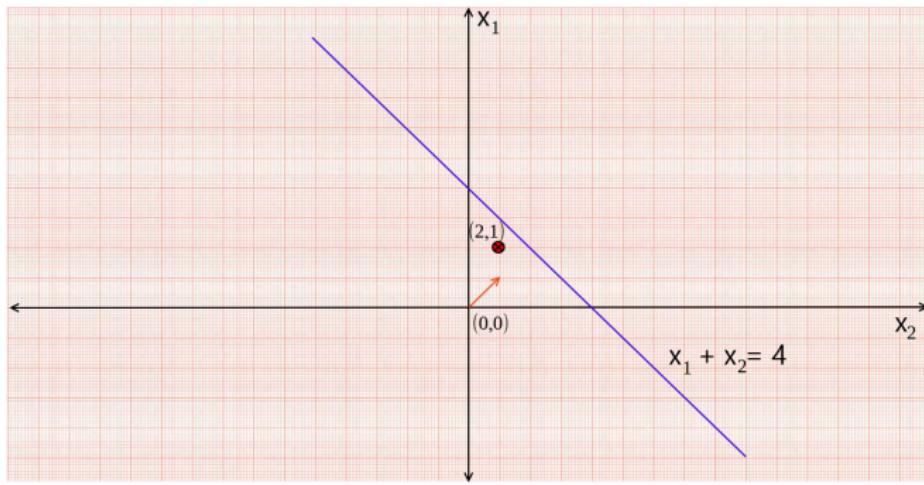
- Suppose a new point  $x^t = (2, 1)$ ,  $y^t = 1$  arrives.
- Transform  $x^t$  into  $\tilde{x}^t = (2, 1, 1)$ .

# Perceptron Update Rule - Geometric Idea



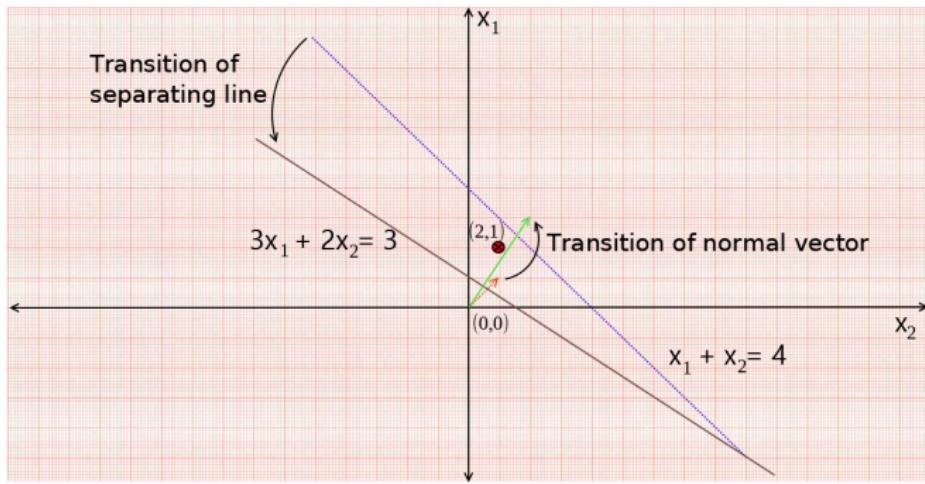
- Suppose a new point  $x^t = (2, 1)$ ,  $y^t = 1$  arrives.
- However with the current weights, perceptron outputs  $\hat{y}^t = -1$ .  
*(why?)*

# Perceptron Update Rule - Geometric Idea



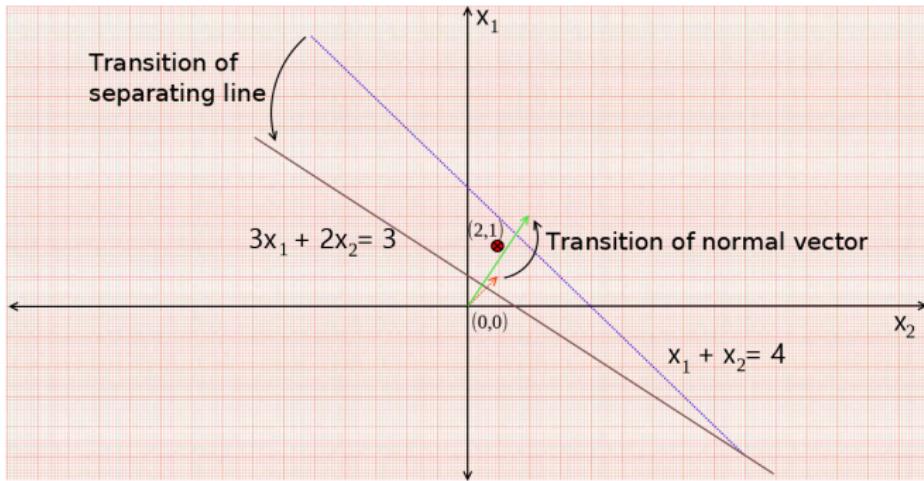
- Suppose a new point  $x^t = (2, 1)$ ,  $y^t = 1$  arrives.
- However with the current weights, perceptron outputs  $\hat{y}^t = -1$ .
- Hence an error occurs and  $\tilde{w}^t$  gets updated to  $\tilde{w}^{t+1} = \tilde{w}^t + y^t \tilde{x}^t$ .

# Perceptron Update Rule - Geometric Idea



- Suppose a new point  $x^t = (2, 1)$ ,  $y^t = 1$  arrives.
- However with the current weights, perceptron outputs  $\hat{y}^t = -1$ .
- After update, the new weights become  $\tilde{w}^{t+1} = (3, 2, -3)$ .

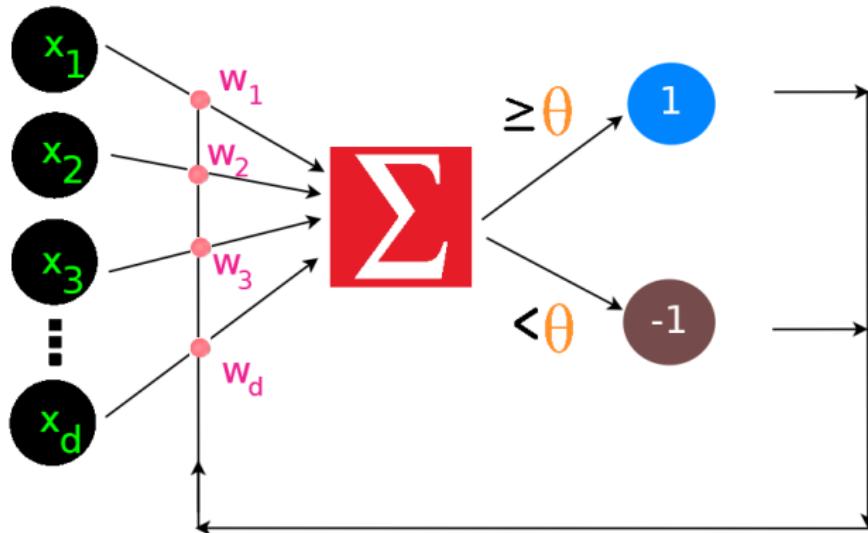
# Perceptron Update Rule - Geometric Idea



## Homework:

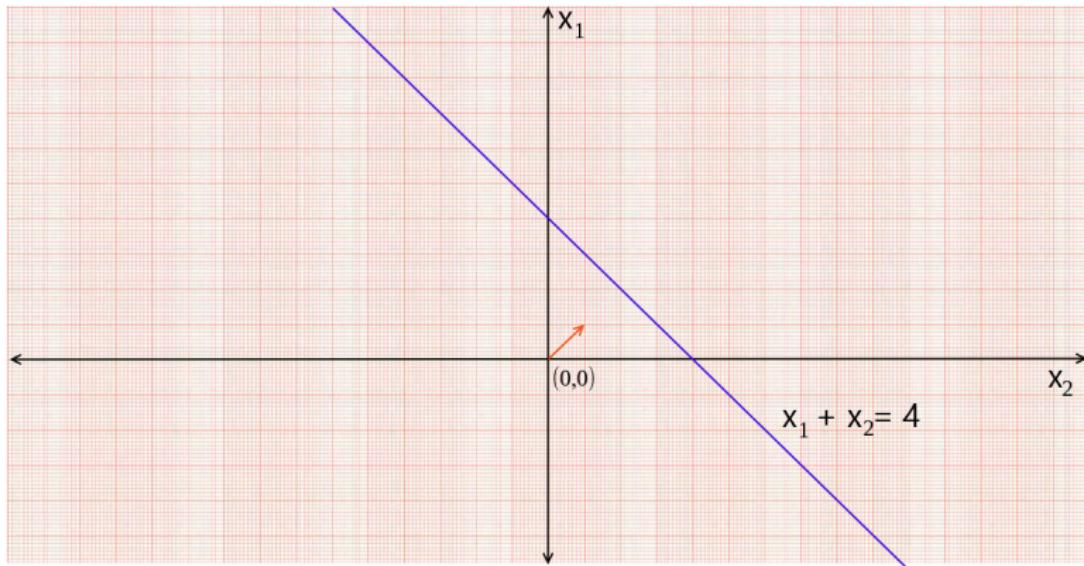
- Suppose now a new point  $x^t = (-1, -1)$  with label  $-1$  comes up.  
How will the weights change?
- Suppose a different new point  $x^t = (-2, 3)$  with label  $+1$  comes up.  
How will the weights change?

# Perceptron - Geometric Idea



- Input is of the form  $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ .
- We associate weights  $w = (w_1, w_2, \dots, w_d) \in \mathbb{R}^d$  to the connections.
- **Prediction Rule:**
  - ▶  $\langle w, x \rangle \geq \theta \implies \text{predict } 1$ .
  - ▶  $\langle w, x \rangle < \theta \implies \text{predict } -1$ .

# Perceptron - Geometric Idea

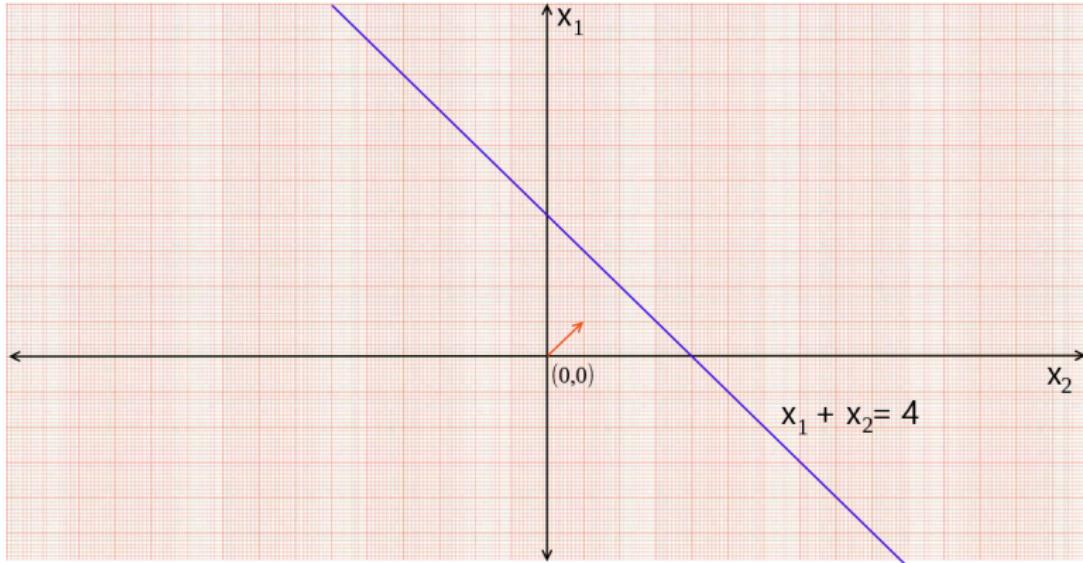


## Prediction Rule

- $\langle w, x \rangle \geq \theta \implies \text{predict } 1.$
- $\langle w, x \rangle < \theta \implies \text{predict } -1.$

**Geometric Idea:** To find a separating hyperplane.

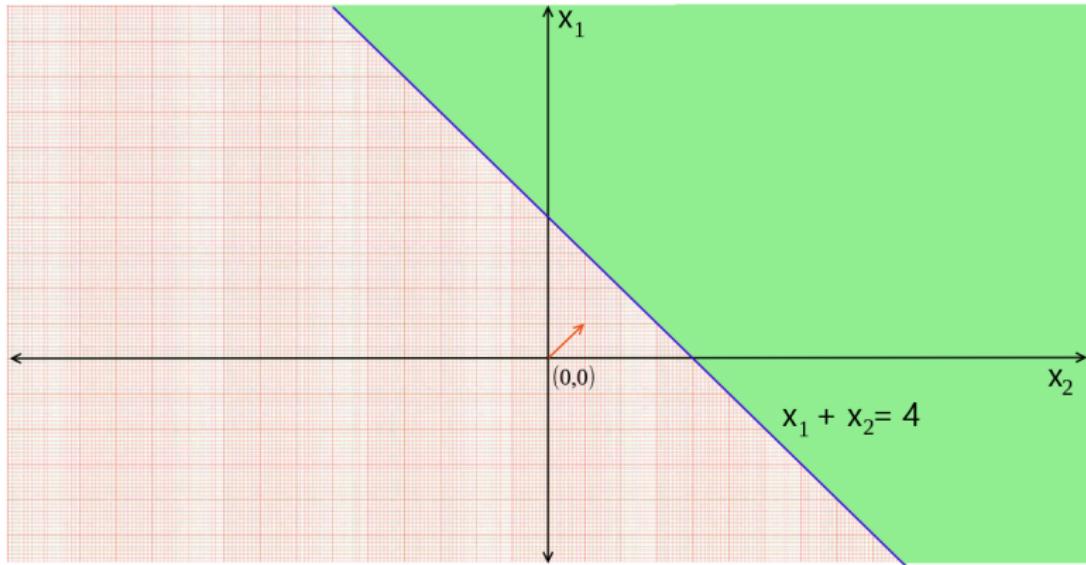
# Perceptron - Geometric Idea



Hyperplane in  $\mathbb{R}^d$

$$\mathcal{H} = \{x \in \mathbb{R}^d : \exists w \neq 0 \text{ and } b \in \mathbb{R} \text{ s.t. } \langle w, x \rangle = b\}.$$

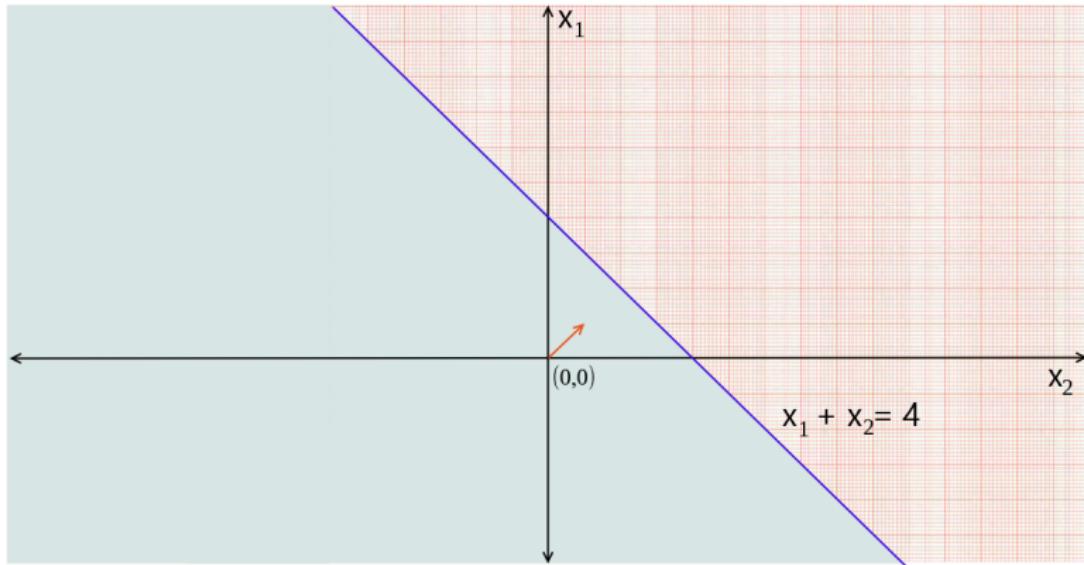
# Perceptron - Geometric Idea



Open Halfspaces associated with a hyperplane  $\langle w, x \rangle = b$

$$Q_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle > b\} \text{ and } Q_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle < b\}.$$

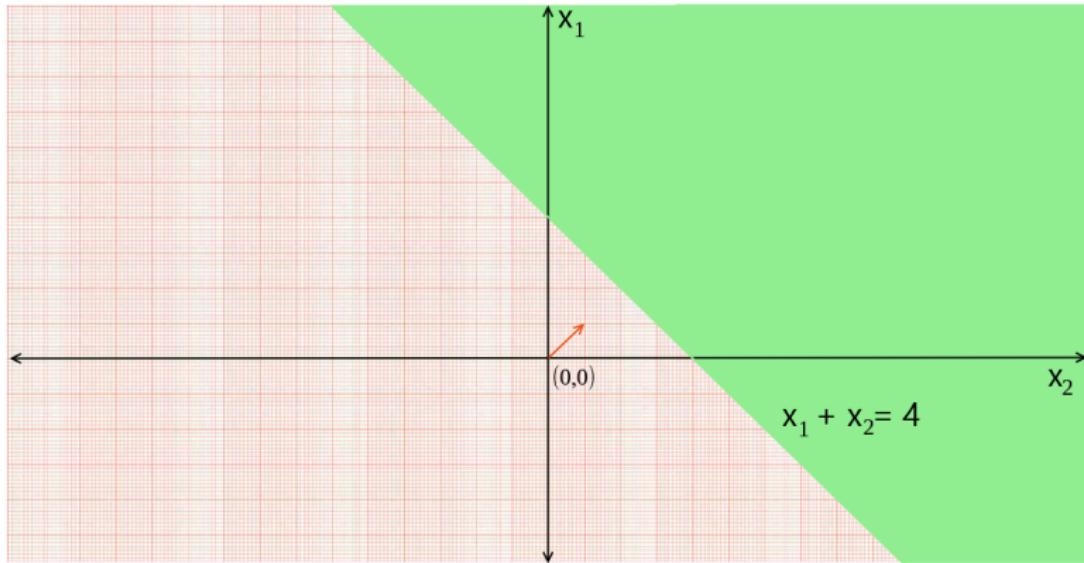
# Perceptron - Geometric Idea



**Open Halfspaces associated with a hyperplane  $\langle w, x \rangle = b$**

$$Q_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle > b\} \text{ and } Q_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle < b\}.$$

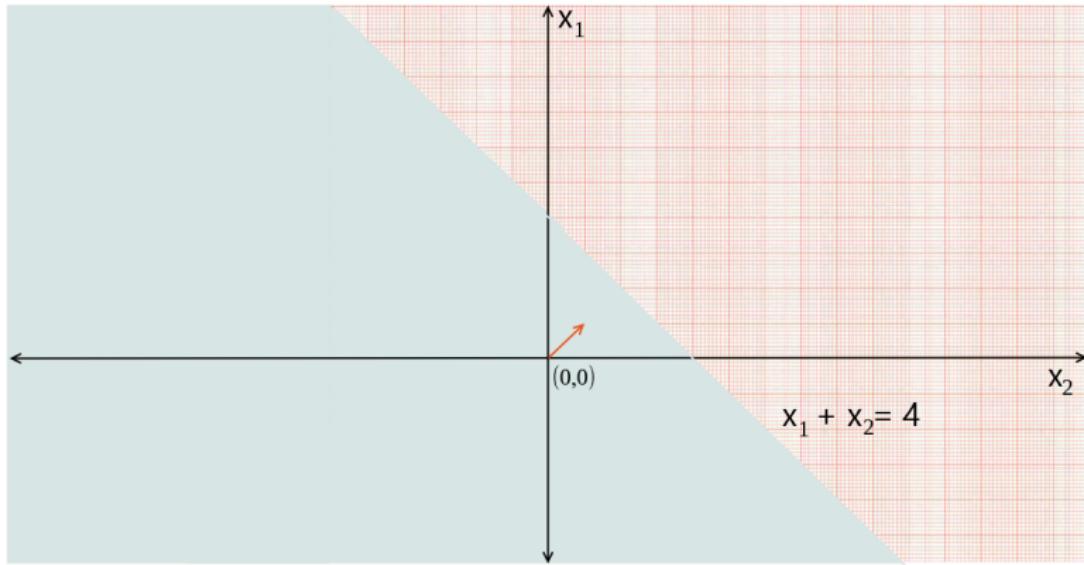
# Perceptron - Geometric Idea



Closed Halfspaces associated with a hyperplane  $\langle w, x \rangle = b$

$$\mathcal{S}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle \geq b\} \text{ and } \mathcal{S}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle \leq b\}.$$

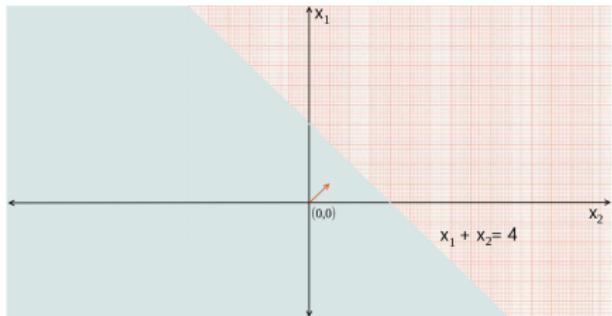
# Perceptron - Geometric Idea



Closed Halfspaces associated with a hyperplane  $\langle w, x \rangle = b$

$$\mathcal{S}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle \geq b\} \text{ and } \mathcal{S}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle \leq b\}.$$

# Perceptron - Geometric Idea

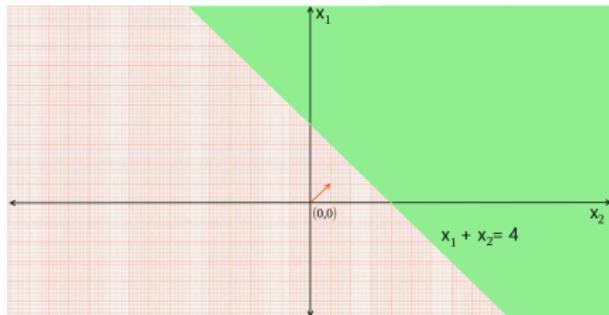
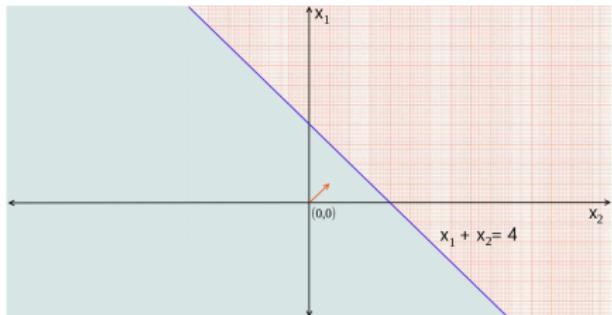


## Prediction Rule

- $\langle w, x \rangle \geq \theta \implies \text{predict } 1.$
- $\langle w, x \rangle < \theta \implies \text{predict } -1.$

**Geometric Idea:** To find a separating hyperplane  $(w, \theta)$  such that samples with class labels 1 and  $-1$  lie on alternate sides of the hyperplane.

# Perceptron - Geometric Idea



## Prediction Rule

- $\langle w, x \rangle \geq \theta \implies \text{predict } 1.$
- $\langle w, x \rangle < \theta \implies \text{predict } -1.$

**Geometric Idea:** To find a separating hyperplane  $(w, \theta)$  such that samples with class labels 1 and  $-1$  lie on alternate sides of the hyperplane.

# Perceptron - Convergence

## Convergence of Perceptron Training Procedure

Under a suitable **separation assumption** of the positive and negative labeled data, the training procedure for Perceptron converges in finite time.