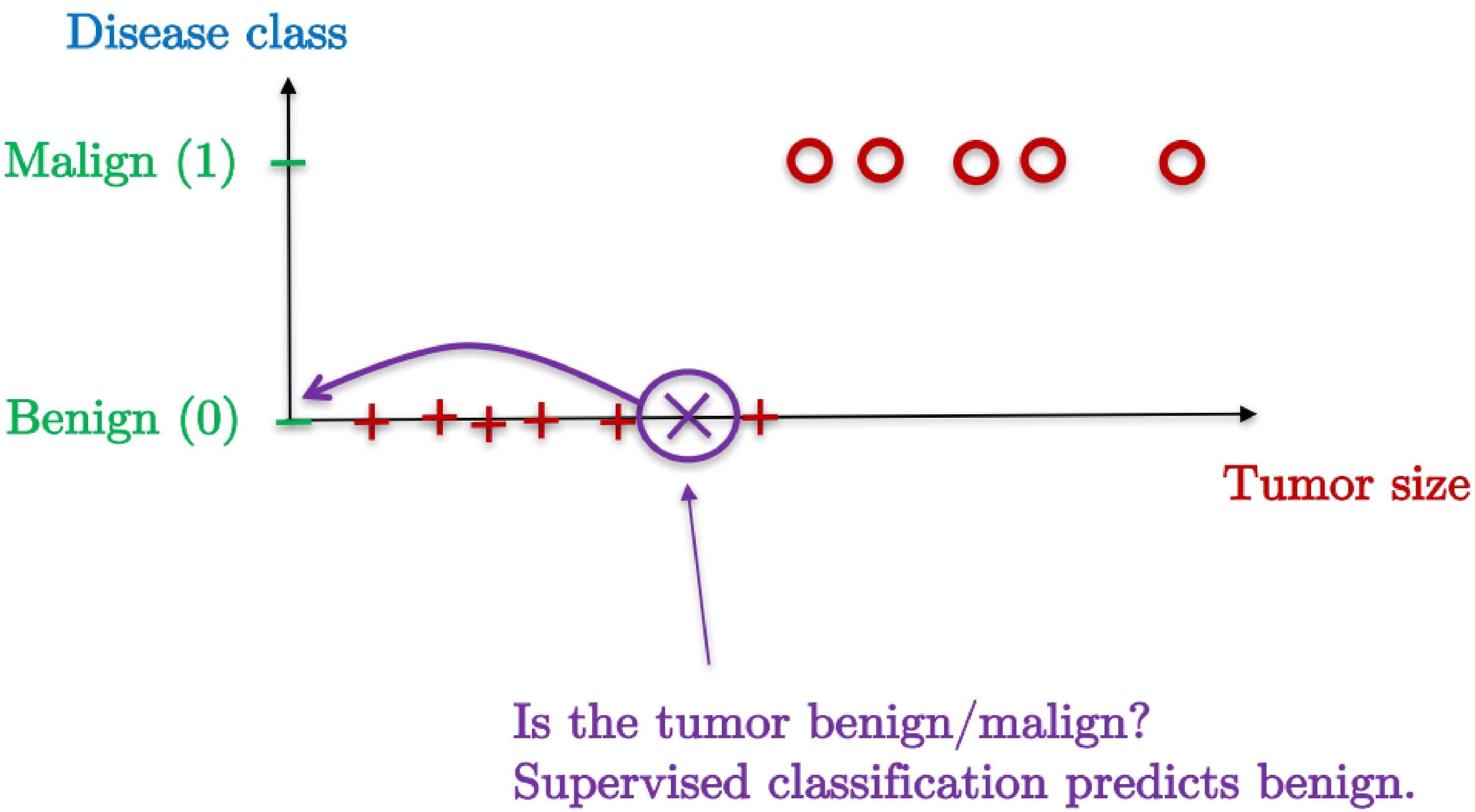
인공지능과 수학적 배경

Classification

Disease class prediction

- **Supervised classification problem: Predict the disease class (discrete value) of patient given existing medical data features (tumor size).**



More examples

- Examples of **binary classification** tasks:
  - Email: Spam (1) or not spam (0)
  - Online financial transaction: Fraudulent (1) or legitimate (0)

$y = \{0, 1\}$ Binary variable ≠ Continuous variable (regression)

Benign Malign

$y \in \mathbb{R}$

- From binary to **multi-class classification**:
  - Email: Spam (0), work (1), friends (2), family (3)
  - Medical diseases: Benign (0), malign I (1), malign II (2), malign III (3)

Multi-value variable: $y = \{0, 1, 2, ..., K\}$

Classification
Formalization

- **Supervised classification learning:**

Reminder:
For regression,
y is a scalar

For classification,
y is a discrete variable

Training dataset
(x, y)

↓

Input: x (data) → Supervised classification algorithm *f* → Output: y (data label)

here: tumor size

here: disease class

f : class prediction function

f maps data feature x to data class y

$f: x \rightarrow y$
or
$y = f(x)$

Classification

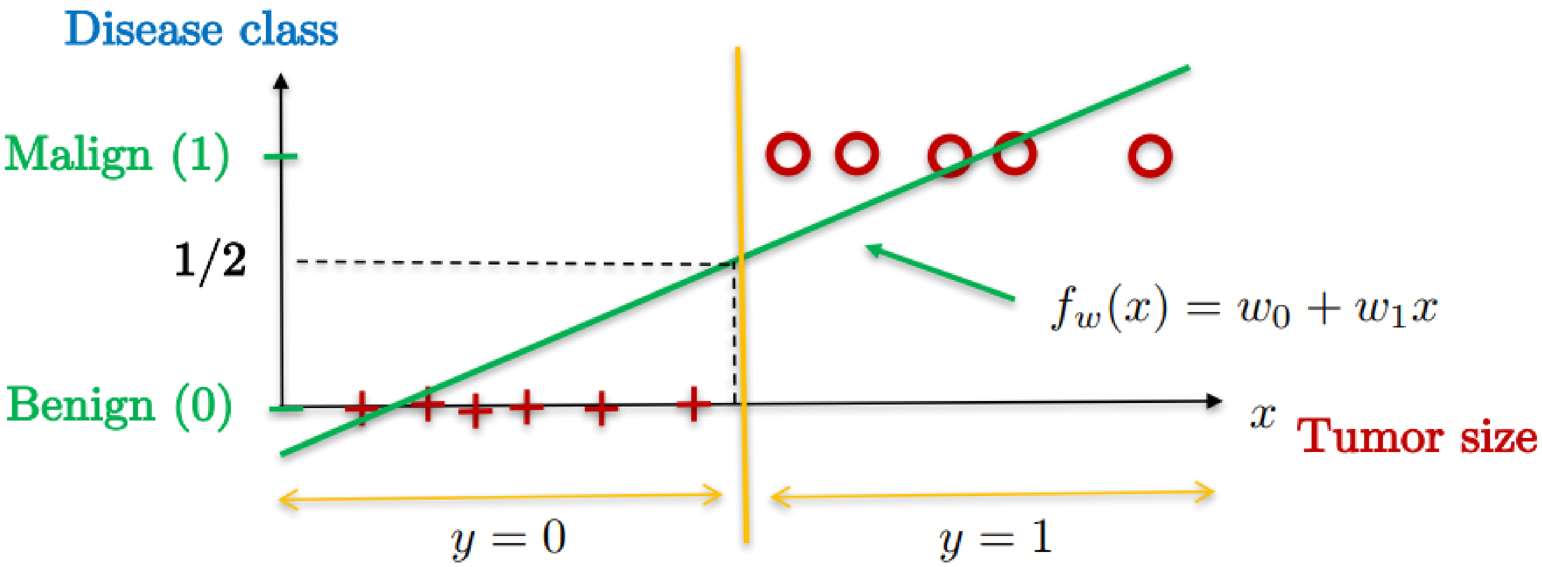
Model representation

- **How to represent a (discrete) class prediction function?**
  - **Linear model? (like for regression)**

$$f_w(x) = w_0 + w_1 x$$

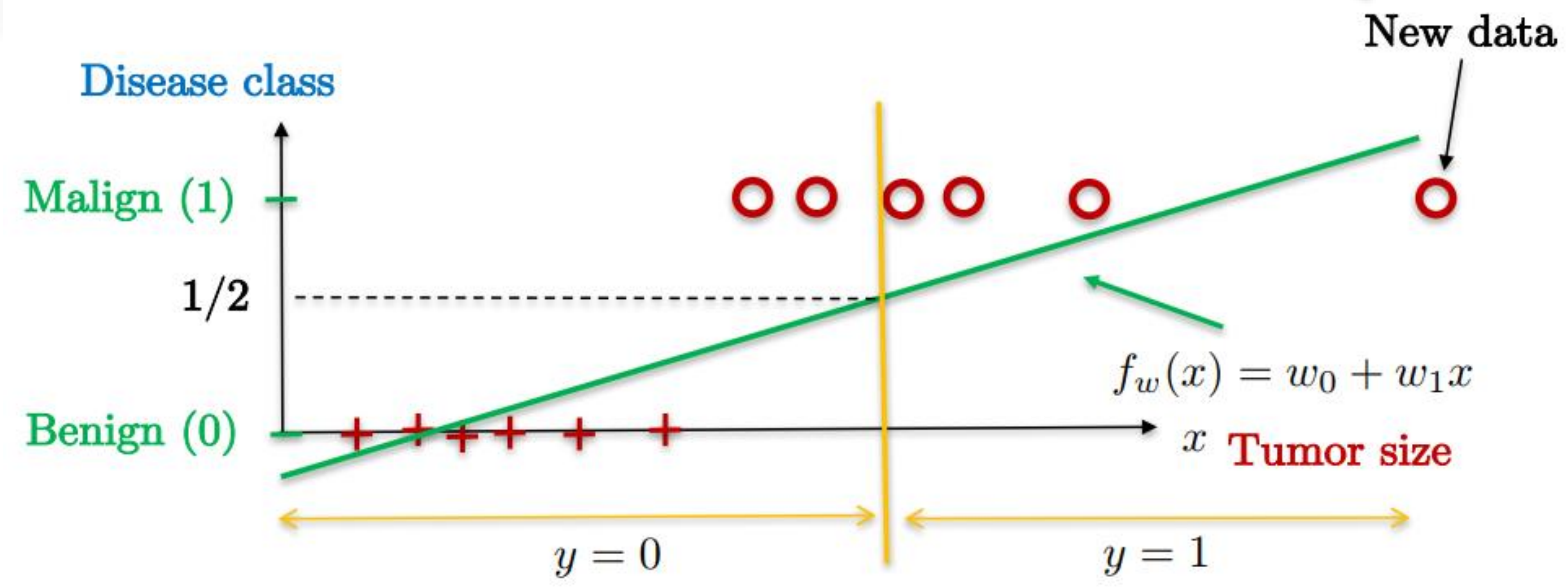
  - **Class prediction might be:**

$$\text{if } f_w(x) \geq 0.5 \text{ then predict } y = 1$$
$$\text{if } f_w(x) < 0.5 \text{ then predict } y = 0$$



Classification
Limitation of linear model

- Linear classification models are not robust to large variations of data features:



- The new data has changed significantly the classification result.

  ⇒ Linear model is not a good solution to the classification problem.

Classification
Model representation

- **Prediction function for classification of *d*-dim data:**

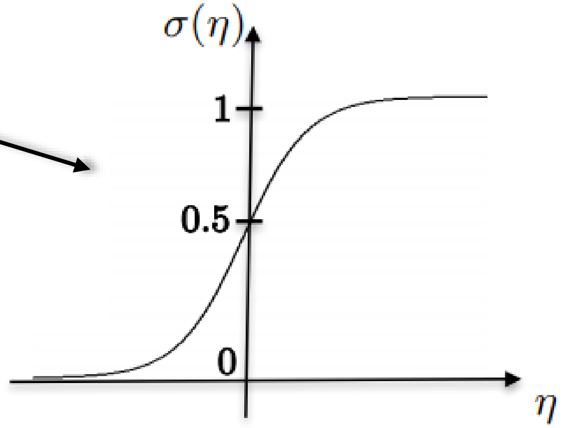
$$\begin{cases} f_w(x) = \sigma(w^T x) \\ \sigma(\eta) = \dfrac{1}{1+e^{-\eta}} \end{cases}$$

Logistic/ sigmoid function

$$f_w(x) = \frac{1}{1+e^{-w^T x}}$$

Logistic regression/ classification function

with $w^T x = w_0 + w_1 x_{(1)} + \ldots + w_d x_{(d)}$



Sigmoid is like a smooth gate

$$x = \begin{bmatrix} 1 \\ x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(d)} \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

Classification

Probabilistic interpretation

- **The prediction function with logistic regression is a probability function:**

$$f_w(x) = \Pr_w(y = 1|x)$$

Probability to have y=1 given data x

Probability is parametrized by w

Example: If x = 5mm (tumor size) and f_w(x) = 0.3 then the patient has 30% chance of tumor being malign.

- **New notation for prediction function:**

$$f_w(x) \quad \Rightarrow \quad p_w(x) = \Pr_w(y = 1|x) = \frac{1}{1 + e^{-w^T x}}$$

Probability function

Classification

Class prediction

- **Soft (continuous) class predictive function:**

$$p_w(x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

- **Observe this predicative function is not a "hard" prediction (discrete value {0, 1} to have class 1 or class 2), but a "soft" prediction (probability value between [0, 1] to have class 1 or class 2).**
- **Hard (discrete) class predicative function:**

$$\text{if } p_w(x) \geq 0.5 \text{ then } y = 1$$
$$\text{if } p_w(x) < 0.5 \text{ then } y = 0$$

Classification

Decision boundary

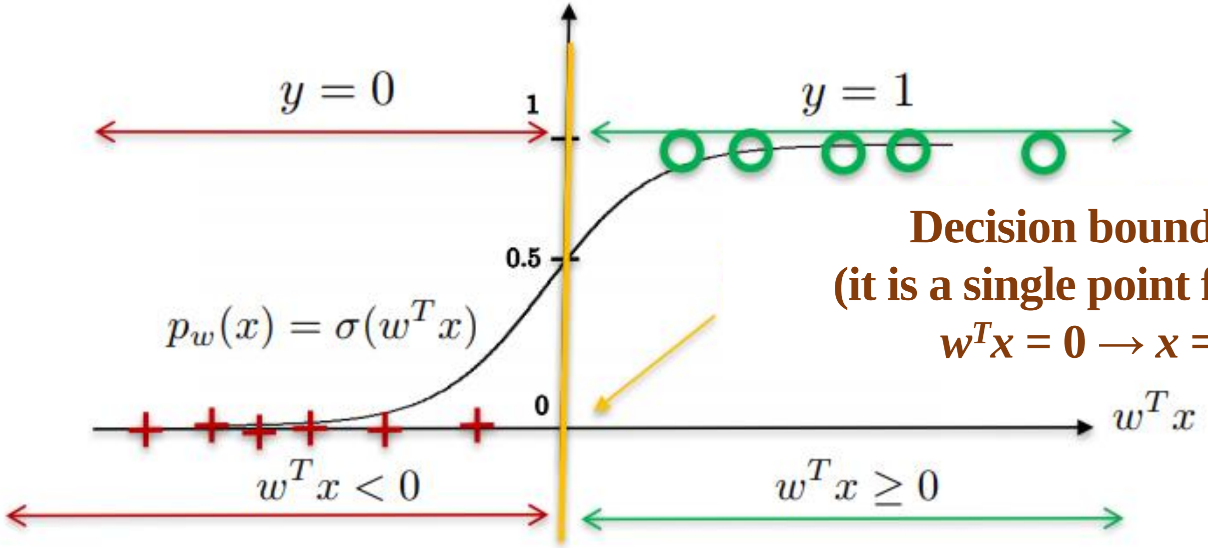
- **Interpretation of**

$$\begin{cases} \text{if } p_w(x) \geq 0.5 \text{ then } y = 1 \\ \text{if } p_w(x) < 0.5 \text{ then } y = 0 \end{cases}$$

As $\sigma(\eta = w^T x) \geq 0.5$ when $\eta = w^T x \geq 0$

Therefore $p_w(x) = \sigma(w^T x) \geq 0.5$ if $w^T x \geq 0$ (and $y = 1$)

And $p_w(x) = \sigma(w^T x) < 0.5$ if $w^T x < 0$ (and $y = 0$)



Decision boundary
(it is a single point for d=1)
$w^T x = 0 \rightarrow x = 0$

Classification

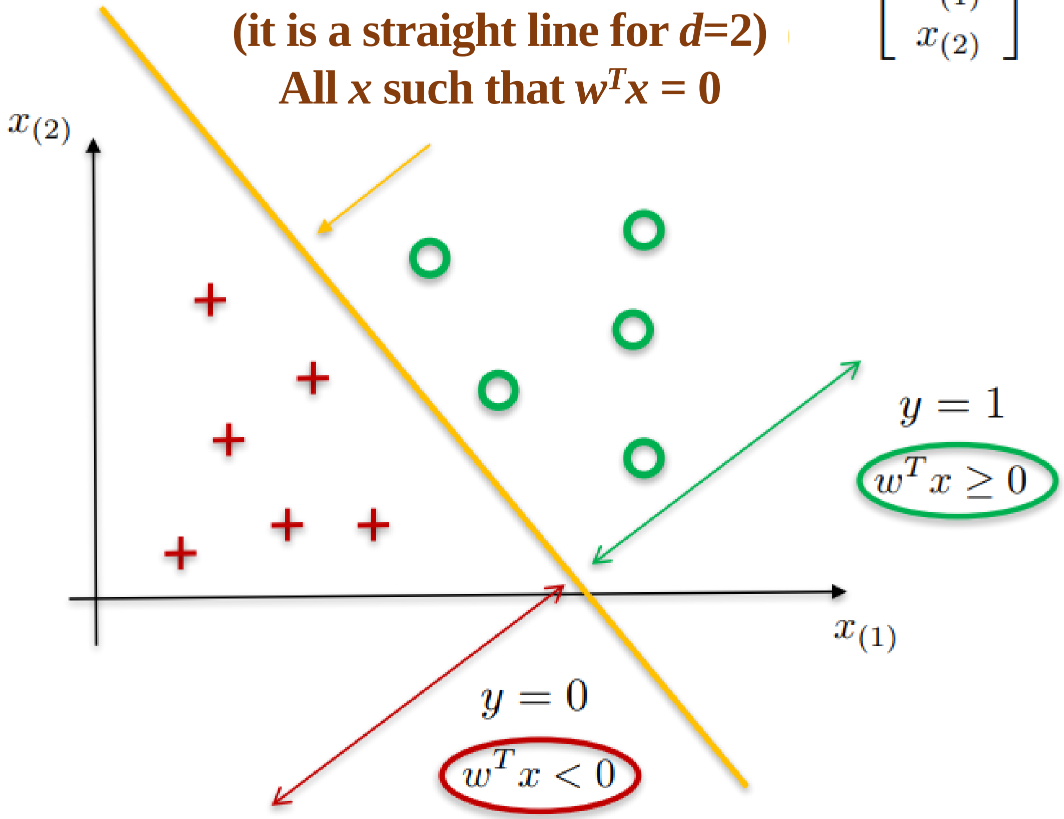
Decision boundary for d = 2 features

- **Decision boundary in higher dimensional spaces:**

$$p_w(x) = \sigma(w_0 + w_1 x_{(1)} + w_2 x_{(2)}) = \sigma(w^T x)$$

Decision boundary
(it is a straight line for d=2)
All x such that $w^T x = 0$

$$x = \begin{bmatrix} 1 \\ x_{(1)} \\ x_{(2)} \end{bmatrix} \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

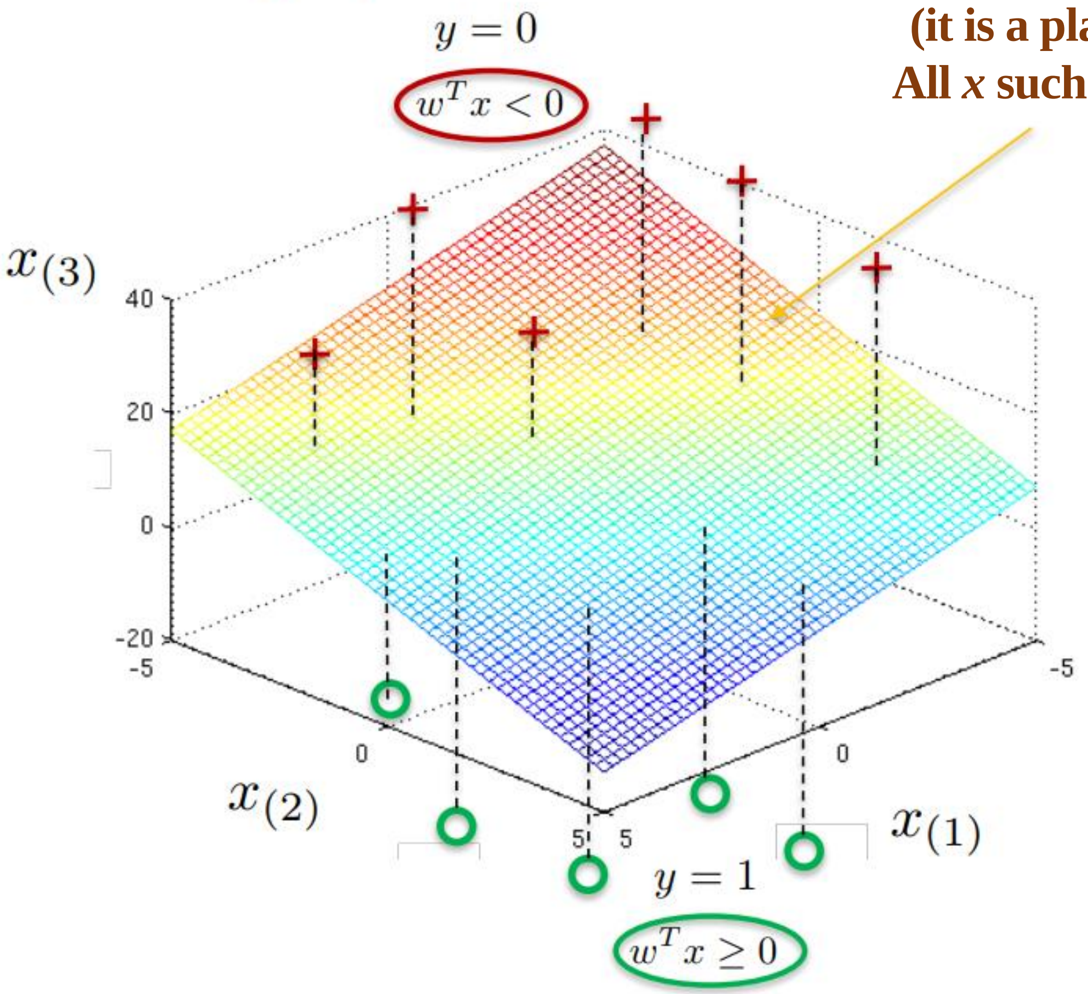


Classification

Decision boundary for *d* features

- **Plan** in 3D and **hyper-plan** in *d*-D

Decision boundary
(it is a plane for *d*=3)
All *x* such that $w^T x = 0$



Classification

Non-linear decision boundary

- **Beyond flat boundaries (straight lines, plans):**

$$p_w(x) = \sigma(\underbrace{w_0 + w_1 x_{(1)} + w_2 x_{(2)} + w_3 x_{(1)}^2 + w_4 x_{(2)}^2}_{\text{Quadratic function}}) = \sigma(w^T x)$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ x_{(1)} \\ x_{(2)} \\ x_{(1)}^2 \\ x_{(2)}^2 \end{bmatrix}$$

- **Class decision function:**

$$\text{if } p_w(x) \geq 0.5 \text{ or } w^T x \geq 0 \text{ then } y = 1$$
$$\text{if } p_w(x) < 0.5 \text{ or } w^T x < 0 \text{ then } y = 0$$

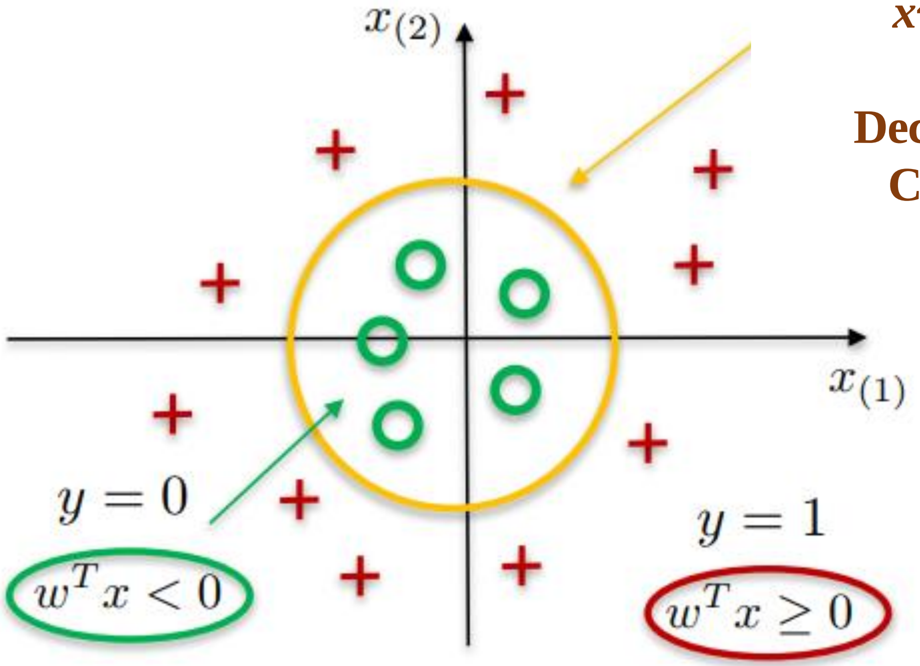
Non-linear decision boundary

- **Example:**

$$w_0=-R^2,\ w_1=w_2=0,\ w_3=w_4=1$$

$$\Downarrow$$

$$w^T x = -R^2 + x_{(1)}^2 + x_{(2)}^2$$



All x such that $w^T x = 0$

$x^2_{(1)} + x^2_{(2)} = R^2$

Decision boundary

Circle equation

Classification

Loss function

- **Predictive function:**

$$p_w(x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

- **How to choose the parameters w of the predictive function p_w?**

  We need:

  - **A loss/cost function to assess the prediction.**
  - **A training set of examples (x_i, y_i) (supervised learning)**

- **Candidate:** **Loss function used for regression?**

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \left(p_w(x_i) - y_i \right)^2 \quad \text{Mean square error (MSE)}$$

Good choice for classification?

Classification

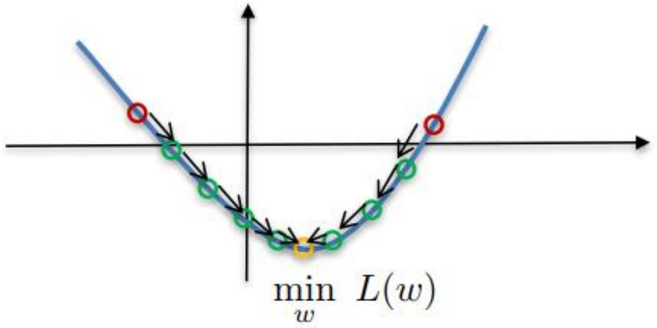
MSE loss for regression and classification

- **Linear regression predictive function:**

$$f_w(x) = w^T x$$

- **MSE loss:**

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(w^T x_i - y_i\right)^2$$



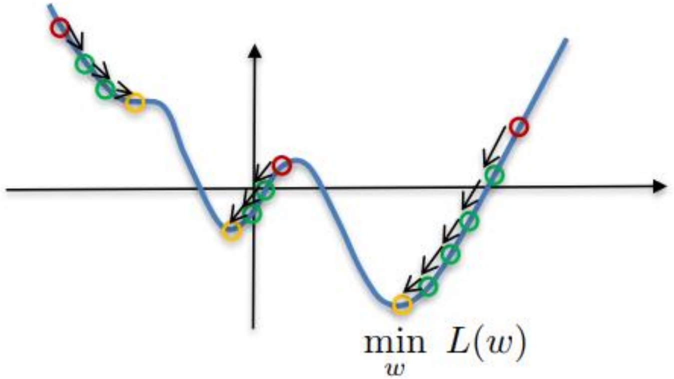
- **L function is convex ☺**
  - **GD guarantees to find (global) minimum**

- **Classification predictive function:**

$$p_w(x) = \frac{1}{1+e^{-w^T x}}$$

- **MSE loss:**

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{1+e^{-w^T x_i}} - y_i\right)^2$$



- **L function is non-convex ☹**
  - **GD no guaranteed to converge to global minimum**