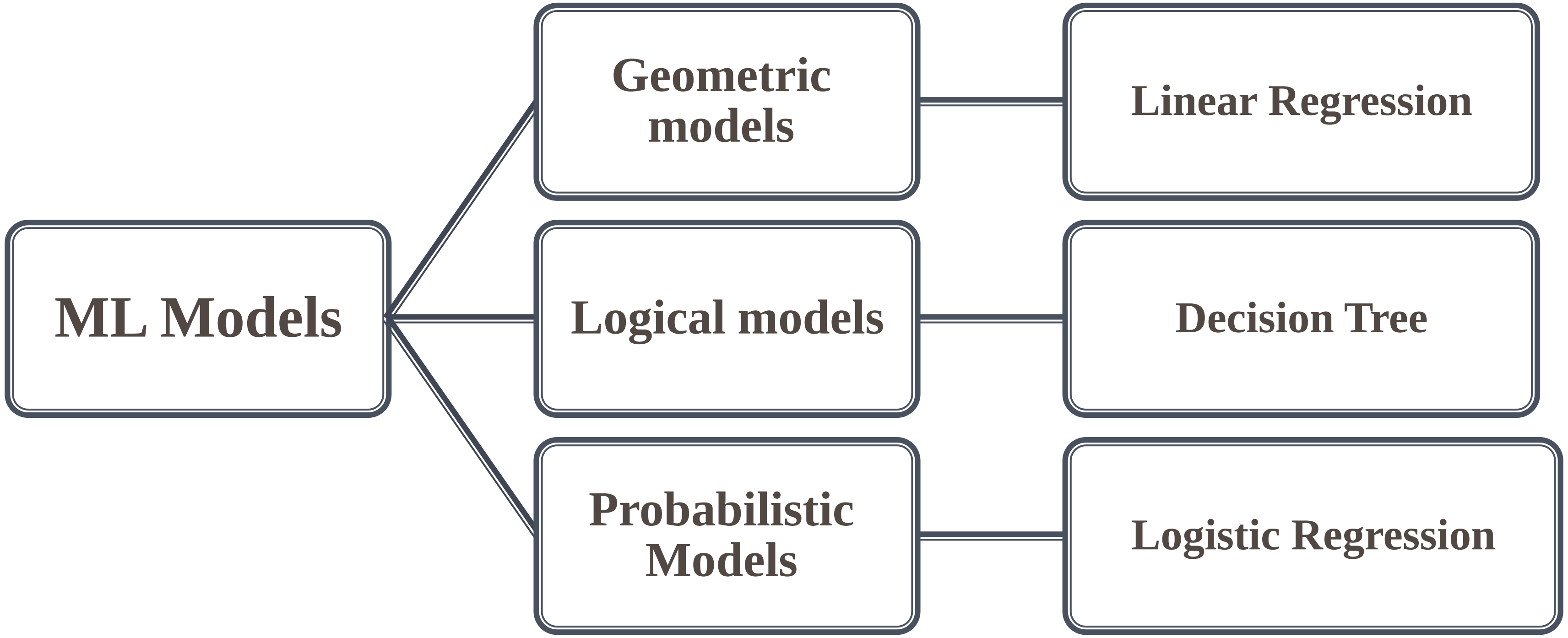
Machine learning

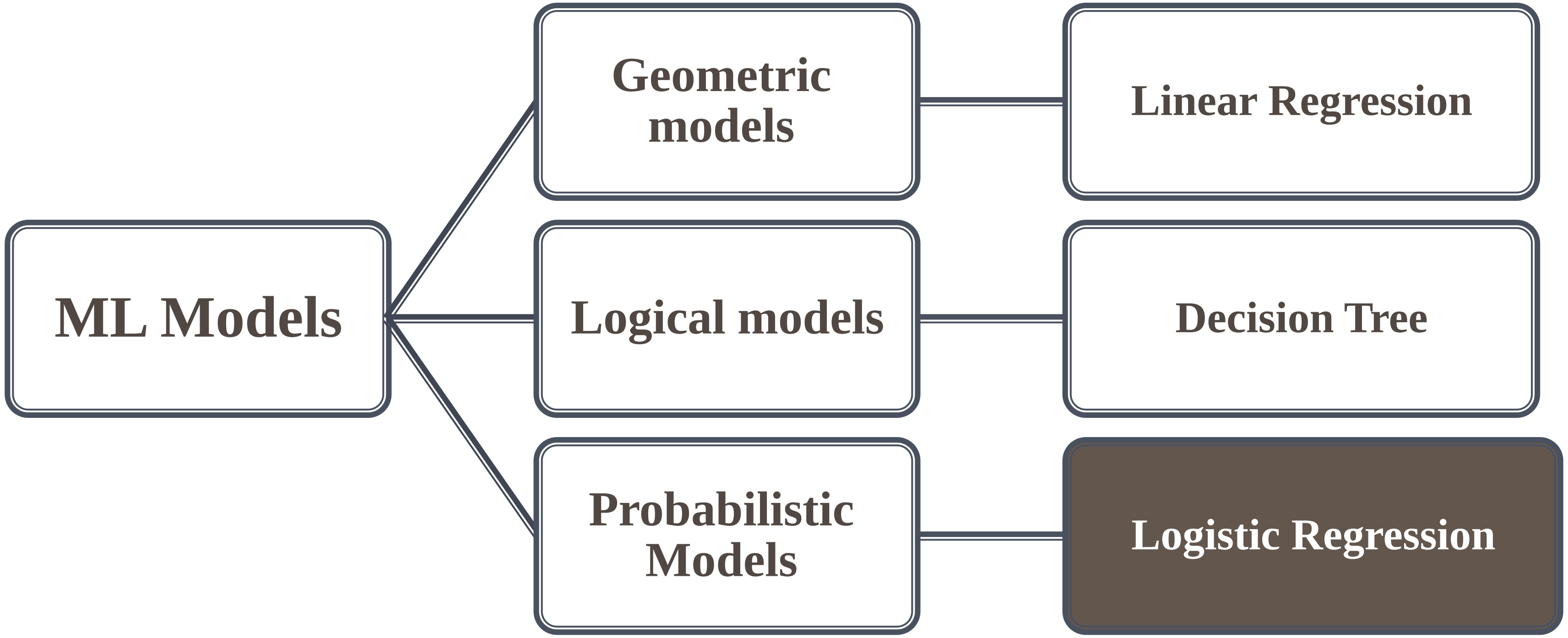
Prepared by : Dr. Hanaa Bayomi
Updated By: Prof Abeer ElKorany

Lecture 4 : Logistic Regression

Flach talks about three types of Machine Learning models [Fla12]



Flach talks about three types of Machine Learning models [Fla12]



CLASSIFICATION

The classification problem is just like the regression problem, except that the values **y** we now want to predict take on only a **small number of discrete values**.

Some Example of Classification problem

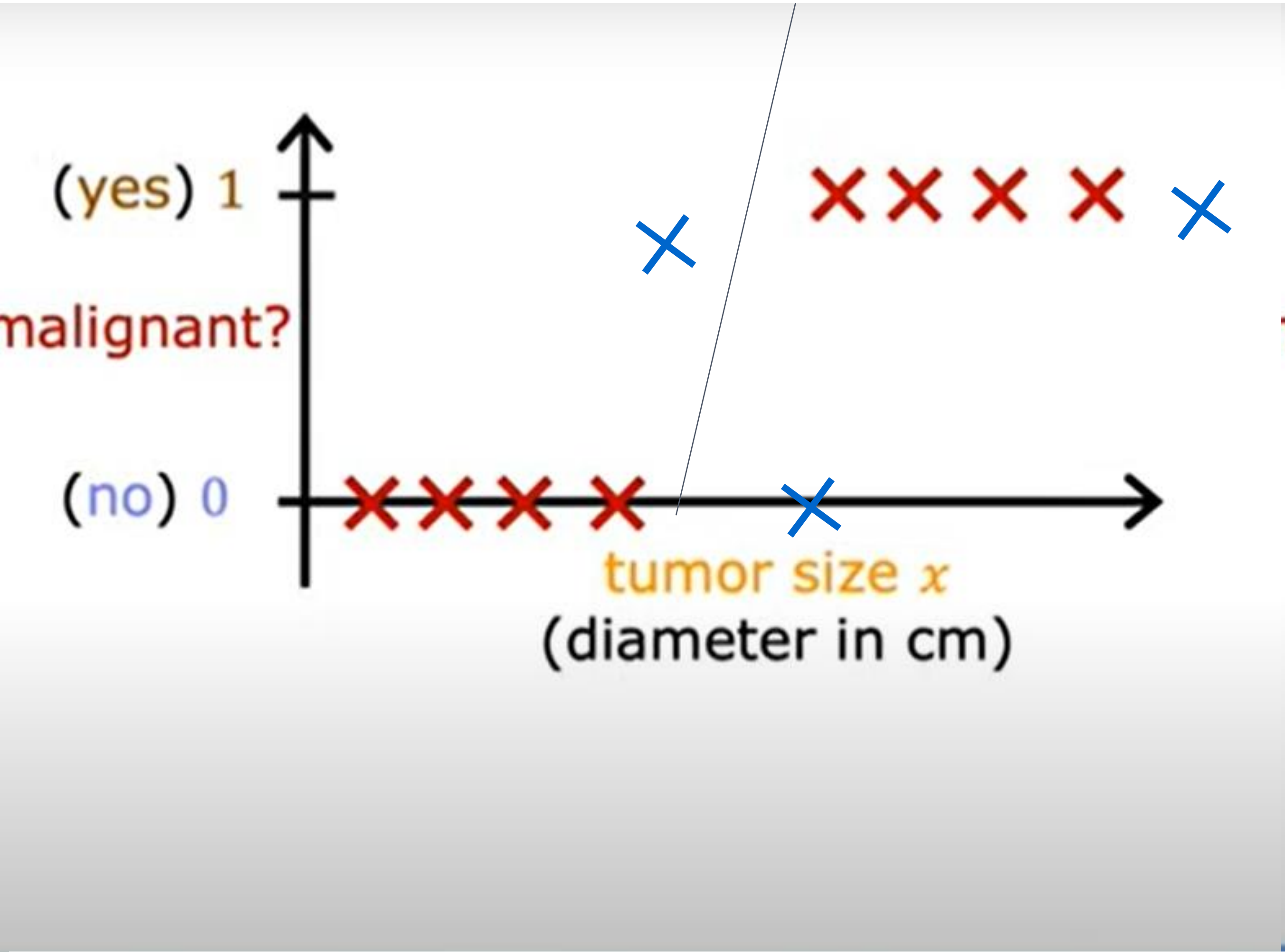
- Email : Spam / Not spam
- Tumor: Malignant/ Benign
- Transaction : Fraud /NO

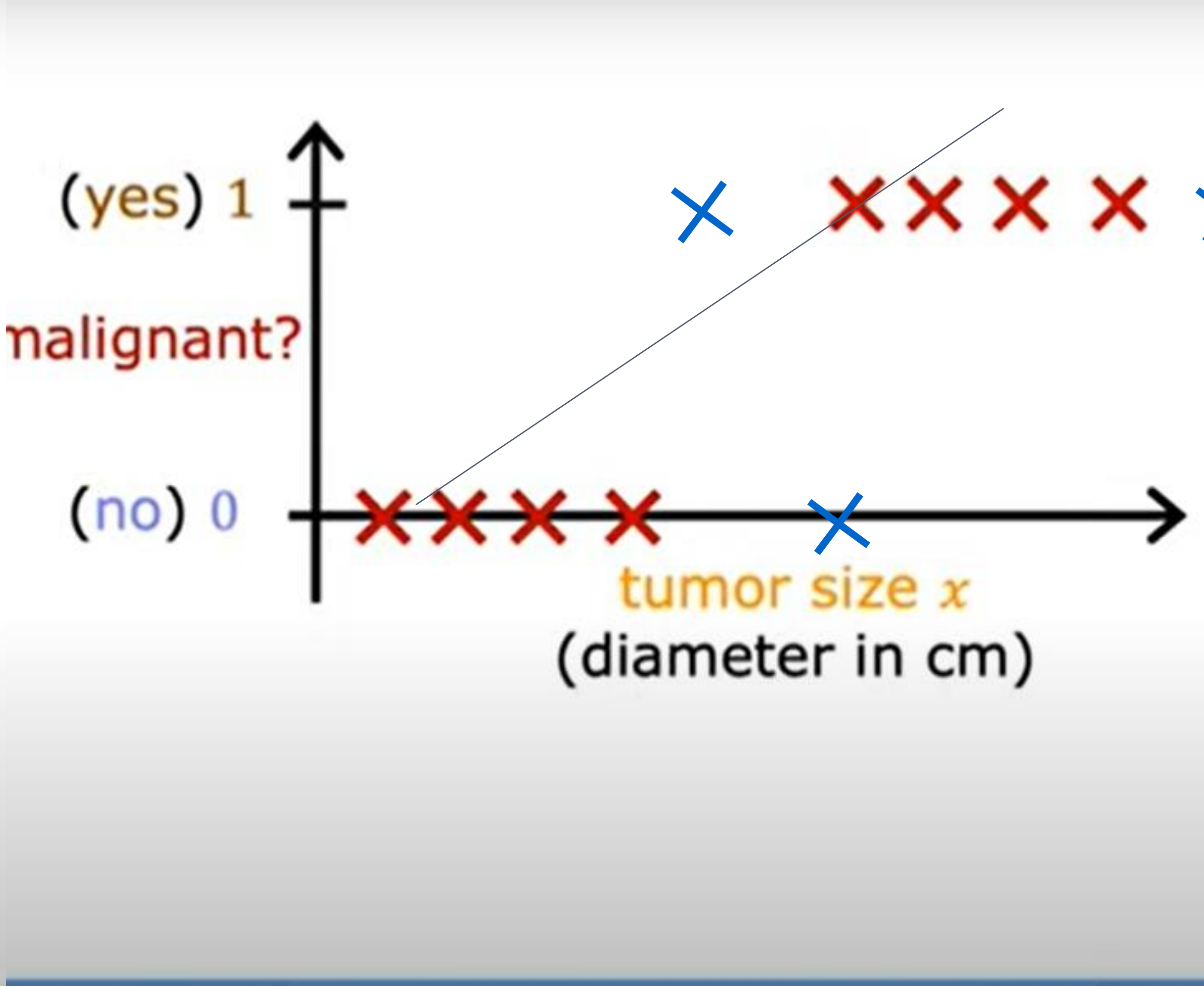
$y \in \{0, 1\}$

0: "Negative Class" (e.g., benign tumor)

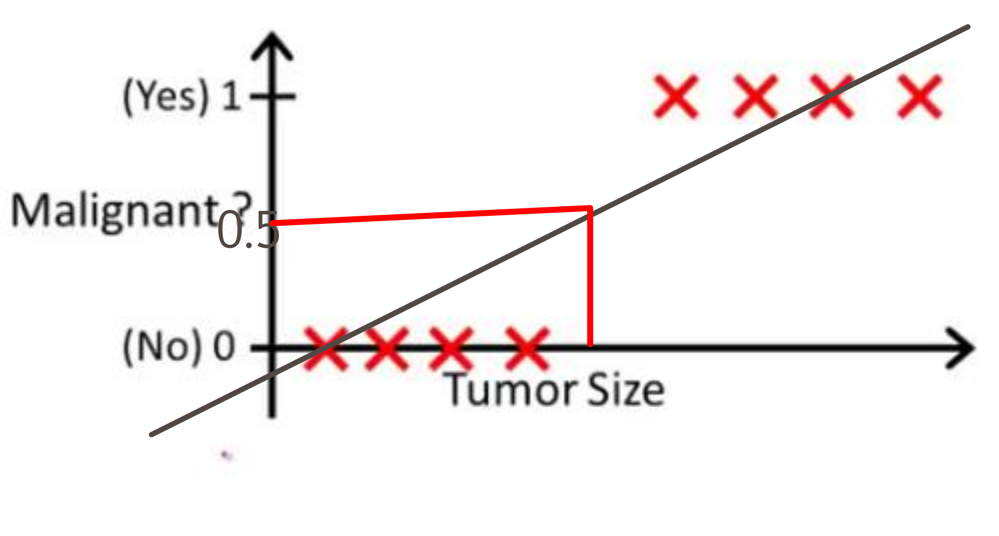
1: "Positive Class" (e.g., malignant tumor)

CLASSIFICATION





CLASSIFICATION

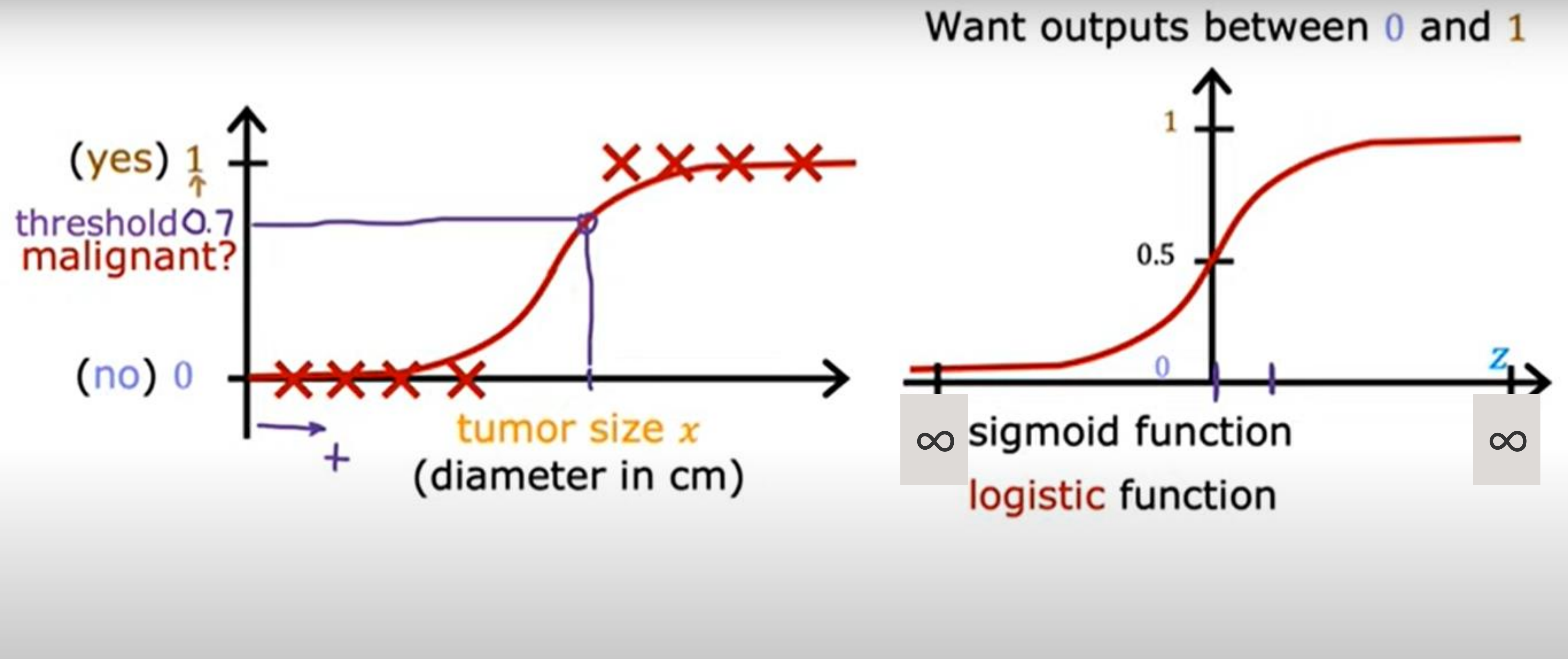


Threshold classifier output $h_\theta(x)$ at 0.5:

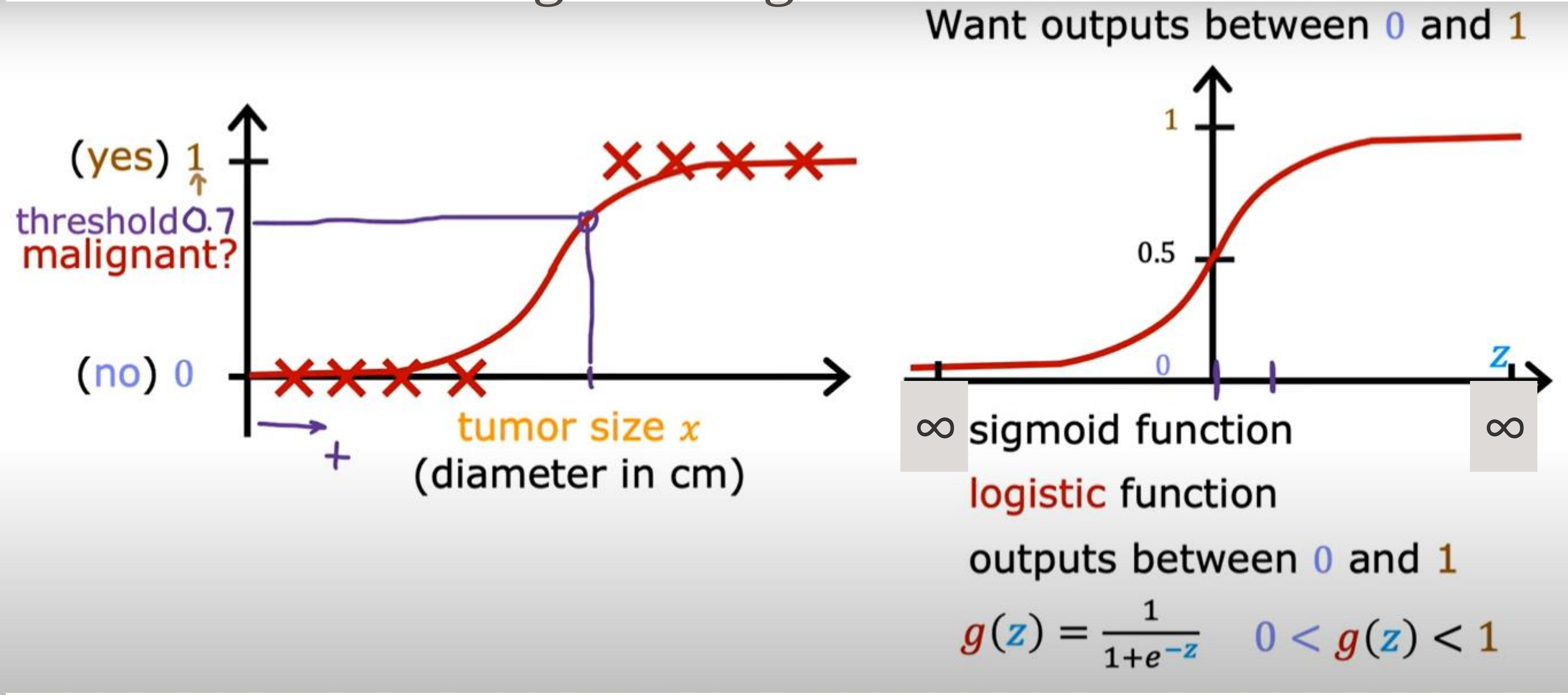
If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

CLASSIFICATION

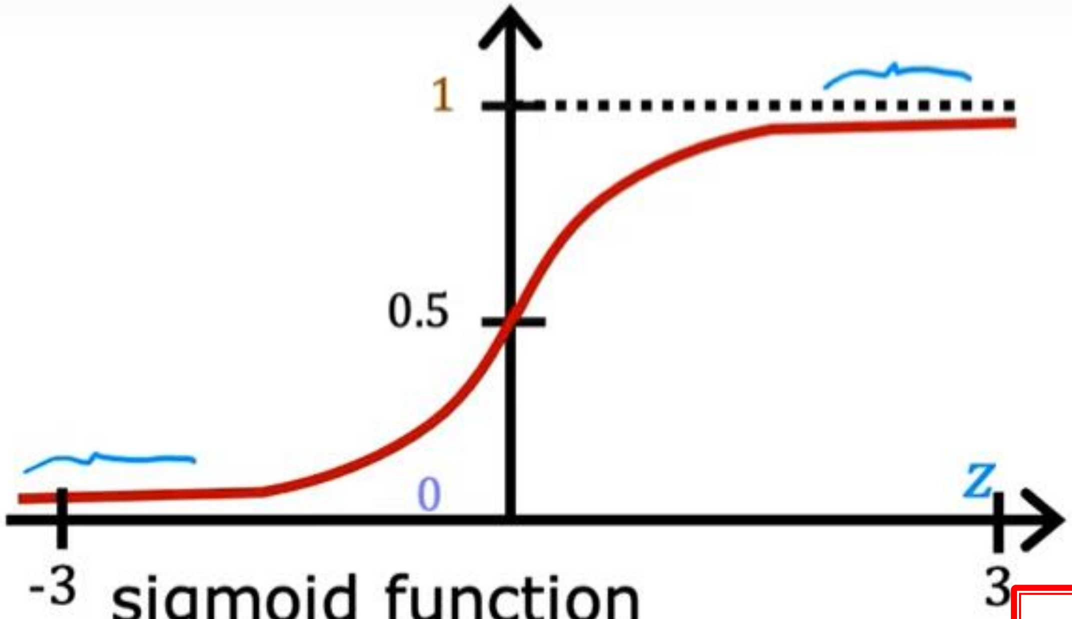


Logistic Regression



Logistic Regression

Want outputs between 0 and 1



sigmoid function

logistic function

outputs between 0 and 1

$$g(z) = \frac{1}{1+e^{-z}} \qquad 0 < g(z) < 1$$

$$f_{\vec{w},b}(\vec{x})$$

$$\boxed{z = \vec{w} \cdot \vec{x} + b}$$

$$\downarrow$$

$$z$$

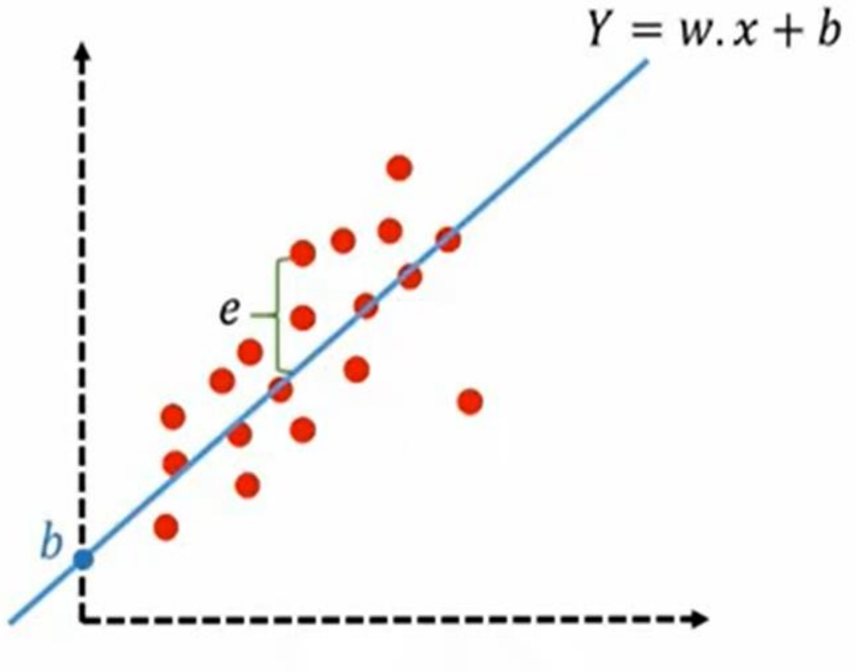
$$\downarrow$$

$$\boxed{g(z) = \frac{1}{1+e^{-z}}}$$

$$f_{\vec{w},b}(\vec{x}) = g(\underbrace{\vec{w} \cdot \vec{x} + b}_{z}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

"logistic regression"

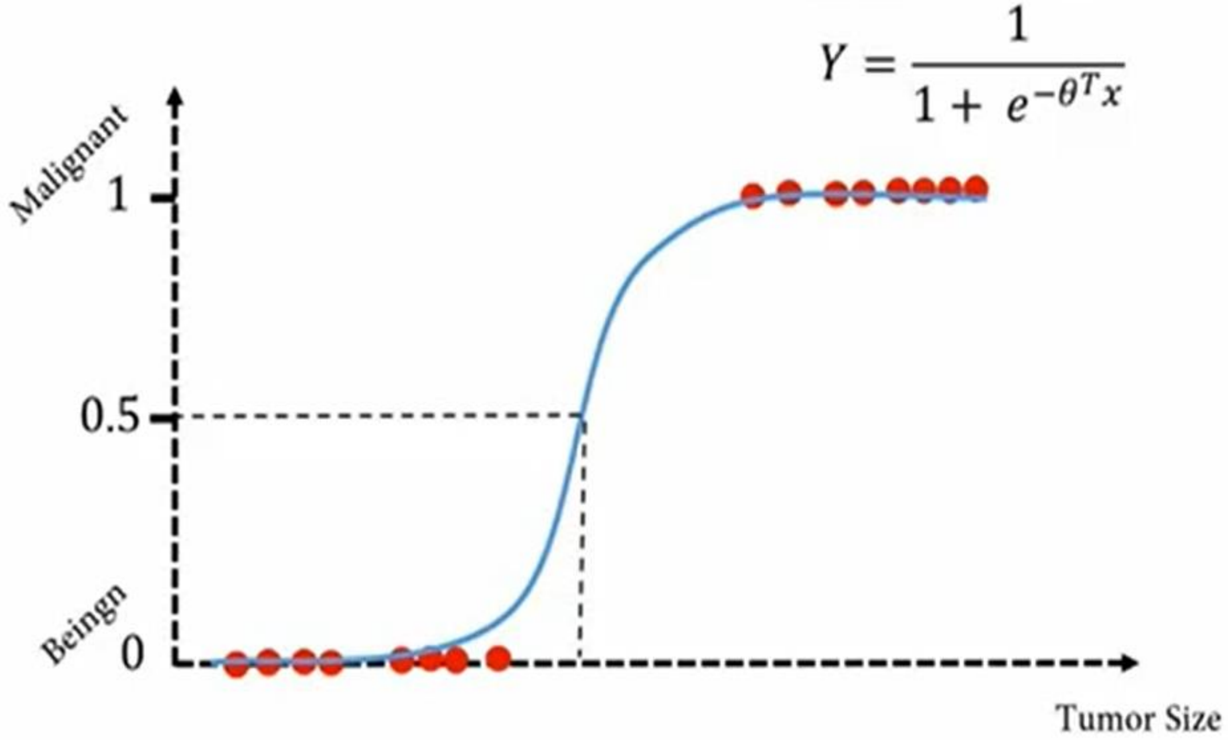
Logistic Regression



Linear regression

Regression Probelm: Continous

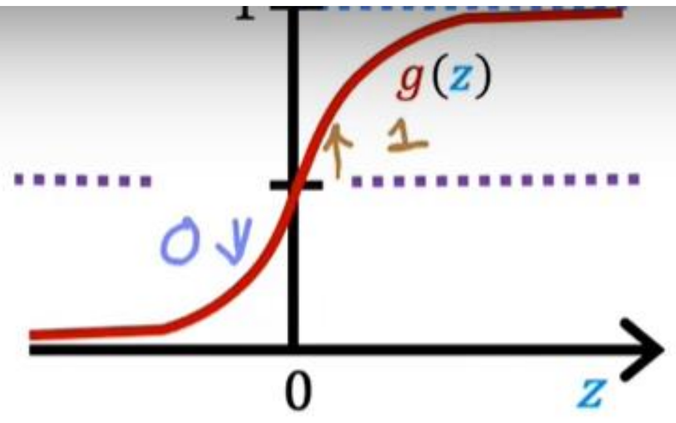
- Stock prices



Logistic regression

Classification Probelm: Discrete

- Malignant or benign tumor



$$f_{\vec{w},b}(\vec{x})$$

$$\boxed{z = \vec{w} \cdot \vec{x} + b}$$

$$\downarrow$$

$$z$$

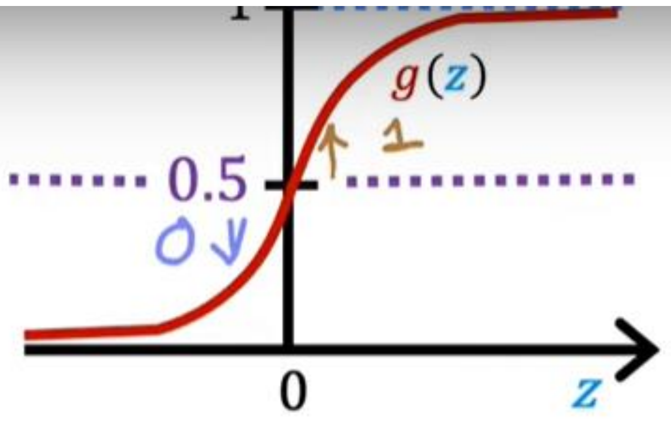
$$\downarrow$$

$$\boxed{g(z) = \frac{1}{1+e^{-z}}}$$

$$f_{\vec{w},b}(\vec{x}) = g(\underbrace{\vec{w} \cdot \vec{x} + b}_{z}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}}$$

$$= P(y = 1|x; \vec{w}, b) \quad 0.7 \quad 0.3$$

0 or 1?



$f_{\vec{w},b}(\vec{x})$

$$\boxed{z = \vec{w}\cdot\vec{x} + b}$$

$$\downarrow$$

$$z$$

$$\downarrow$$

$$\boxed{g(z) = \frac{1}{1+e^{-z}}}$$

$$f_{\vec{w},b}(\vec{x}) = g(\underbrace{\vec{w}\cdot\vec{x} + b}_{z}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}}$$

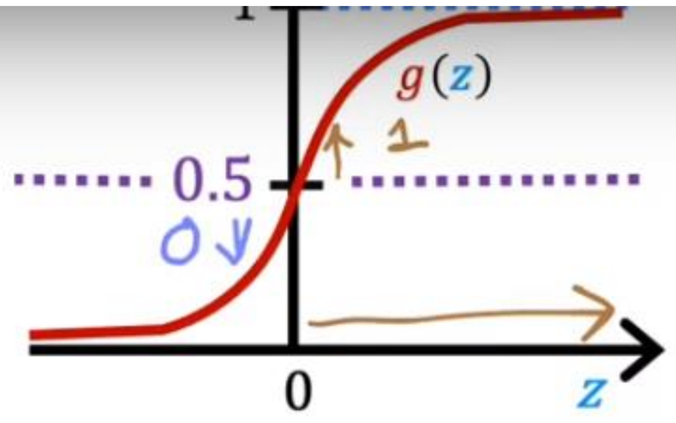
$$= P(y = 1|x;\vec{w},b) \quad 0.7 \quad 0.3$$

0 or 1?

Is $f_{\vec{w},b}(\vec{x}) \geq \overbrace{0.5}^{\text{threshold}}$?

Yes: $\hat{y} = 1$ No: $\hat{y} = 0$

When is $f_{\vec{w},b}(\vec{x}) \geq 0.5$?



$f_{\vec{w},b}(\vec{x})$

$$\boxed{z = \vec{w} \cdot \vec{x} + b}$$

$\downarrow$

z

$\downarrow$

$$\boxed{g(z) = \frac{1}{1+e^{-z}}}$$

$$f_{\vec{w},b}(\vec{x}) = g(\underbrace{\vec{w} \cdot \vec{x} + b}_{z}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}}$$

$$= P(y = 1 | x; \vec{w}, b) \quad 0.7 \quad 0.3$$

0 or 1?

Is $f_{\vec{w},b}(\vec{x}) \geq \overbrace{0.5}^{\text{threshold}}$?

Yes: $\hat{y} = 1$ No: $\hat{y} = 0$

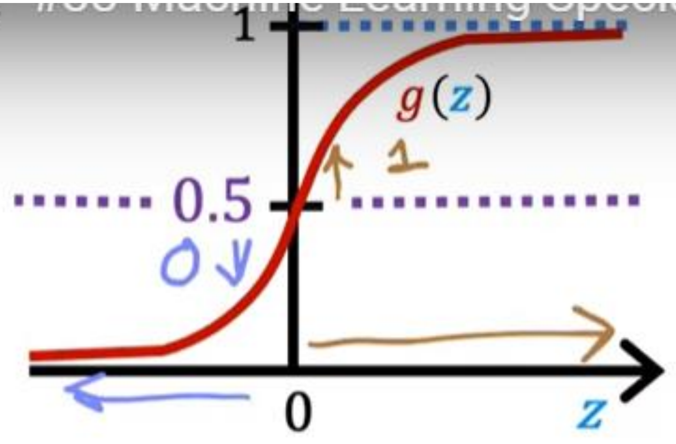
When is $f_{\vec{w},b}(\vec{x}) \geq 0.5$?

$$g(z) \geq 0.5$$

$$z \geq 0$$

$$\vec{w} \cdot \vec{x} + b \geq 0$$

$$\hat{y} = 1$$



$f_{\vec{w},b}(\vec{x})$

$$\boxed{z = \vec{w} \cdot \vec{x} + b}$$

$$\downarrow$$

$$z$$

$$\downarrow$$

$$\boxed{g(z) = \frac{1}{1+e^{-z}}}$$

$$f_{\vec{w},b}(\vec{x}) = g(\underbrace{\vec{w} \cdot \vec{x} + b}_{z}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

$$= P(y = 1 | x; \vec{w}, b) \quad 0.7 \quad 0.3$$

0 or 1? threshold

Is $f_{\vec{w},b}(\vec{x}) \geq \overbrace{0.5}$?

Yes: $\hat{y} = 1$ No: $\hat{y} = 0$

When is $f_{\vec{w},b}(\vec{x}) \geq 0.5$?

$$g(z) \geq 0.5$$

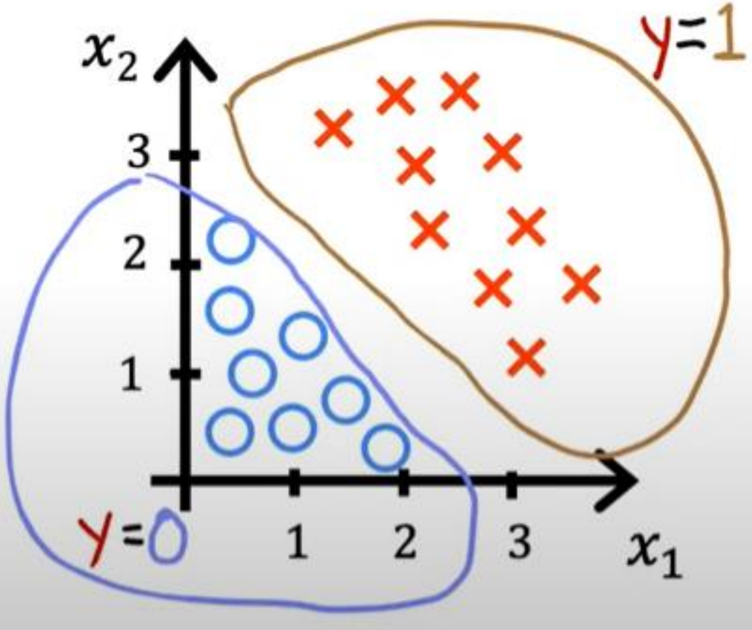
$$z \geq 0$$

$\vec{w} \cdot \vec{x} + b \geq 0$ $\qquad$ $\vec{w} \cdot \vec{x} + b < 0$

$\hat{y} = 1$ $\qquad$ $\hat{y} = 0$

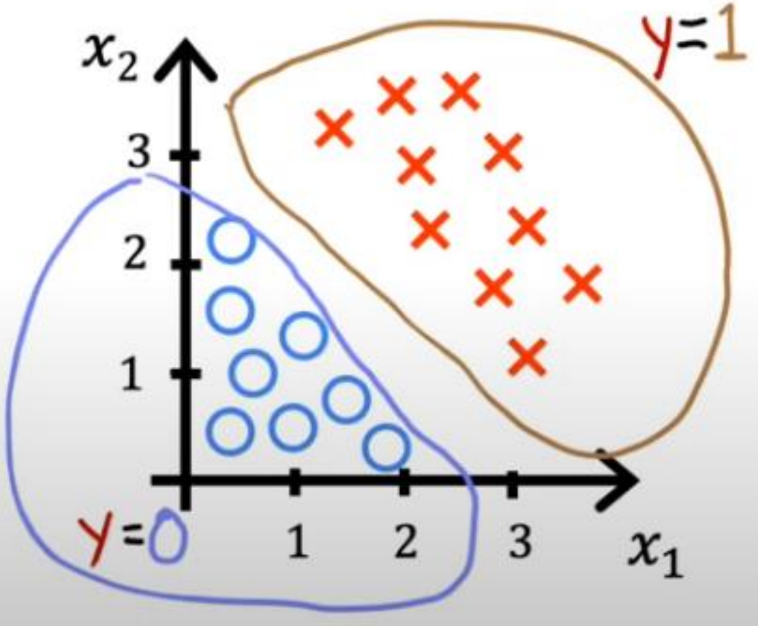
Decision boundary

Logistic regression with two parameters : X1, X2 Range from 0-3



Decision boundary

$$f_{\vec{w},b}(\vec{x}) = g(z) = g(\underset{1}{w_1}x_1 + \underset{1}{w_2}x_2 + \underset{-3}{b})$$



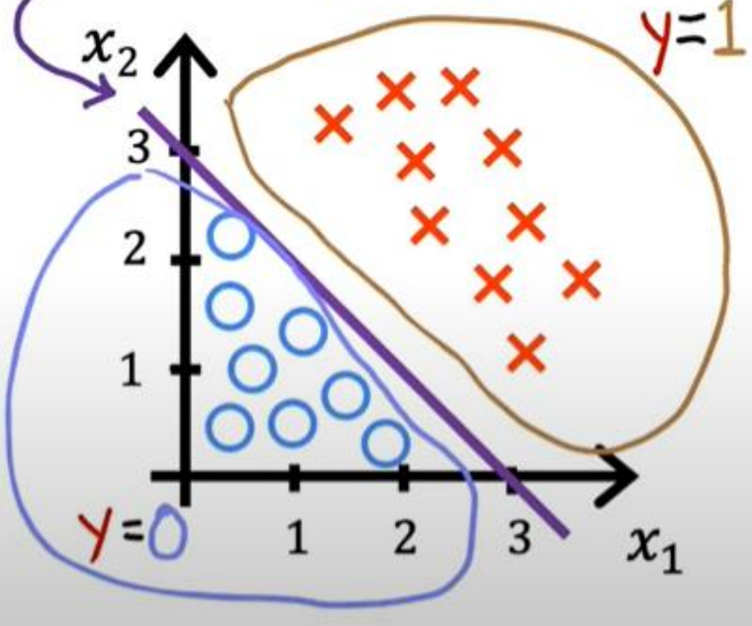
Decision boundary

$$f_{\vec{w},b}(\vec{x}) = g(z) = g(\underset{1}{w_1}x_1 + \underset{1}{w_2}x_2 + \underset{-3}{b})$$

Decision boundary $z = \vec{w} \cdot \vec{x} + b = 0$

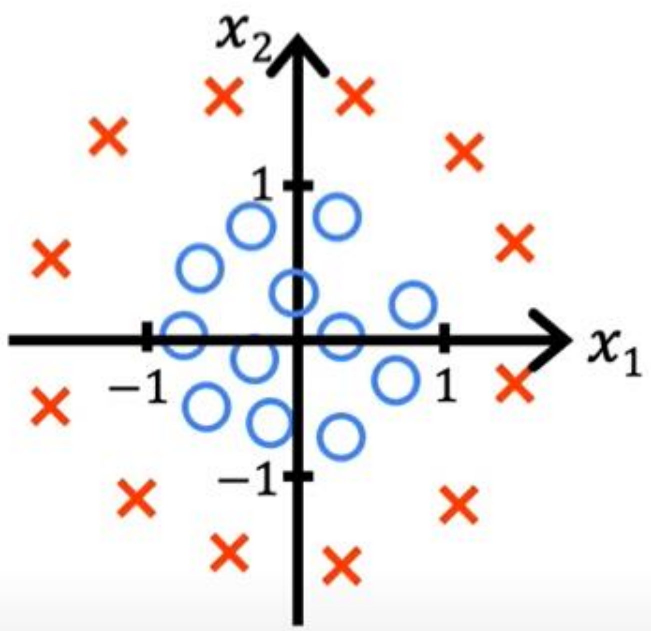
$$z = x_1 + x_2 - 3 = 0$$

$$x_1 + x_2 = 3$$



Logistic regression Using polynomial function

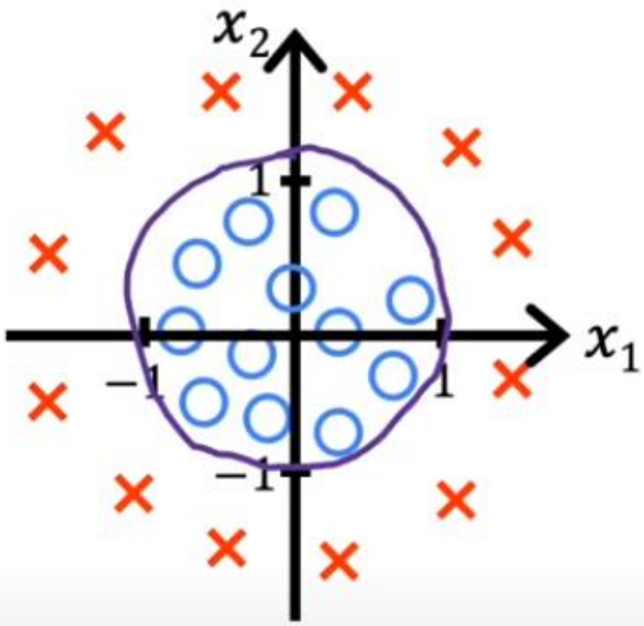
Non-linear decision boundaries



$$f_{\vec{\mathrm{w}},b}(\vec{\mathrm{x}}) = g(z) = g(\overbrace{w_1x_1^2 + w_2x_2^2 + b}^{z})$$

Logistic regression Using polynomial function

Non-linear decision boundaries



$$f_{\vec{w},b}(\vec{x}) = g(z) = g(\overbrace{\underset{1}{w_1}x_1^2 + \underset{1}{w_2}x_2^2 + \underset{-1}{b}}^{z})$$

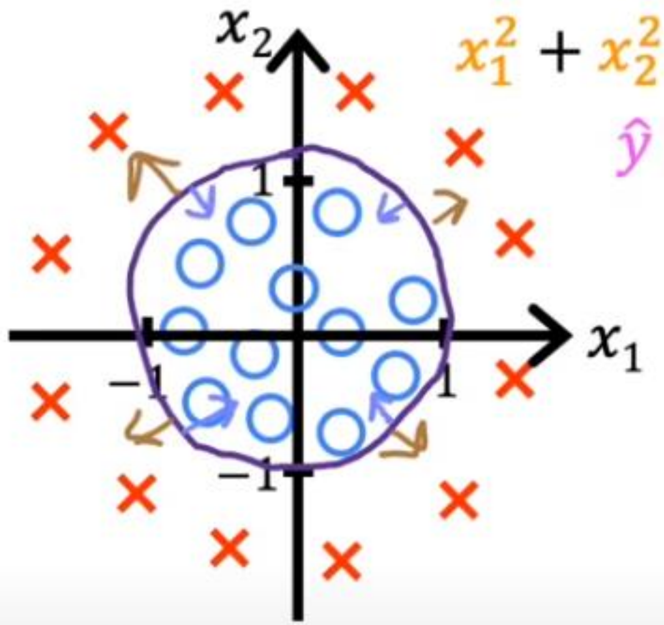
decision boundary

$$z = x_1^2 + x_2^2 - 1 = 0$$

$$x_1^2 + x_2^2 = 1$$

Logistic regression Using polynomial function

Non-linear decision boundaries



$x_1^2 + x_2^2 \ge 1$

$\hat{y} = 1$

$$f_{\vec{w},b}(\vec{x}) = g(z) = g(\overbrace{\underset{1}{w_1} x_1^2 + \underset{1}{w_2} x_2^2 + \underset{-1}{b}}^{z})$$

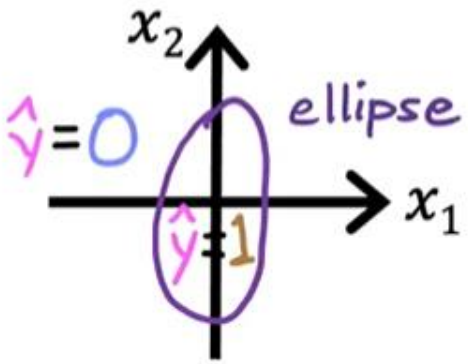
decision boundary $z = x_1^2 + x_2^2 - 1 = 0$

$x_1^2 + x_2^2 = 1$

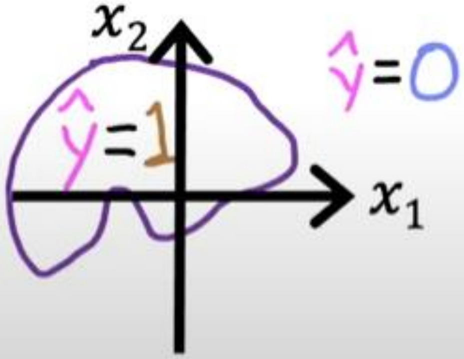
$x_1^2 + x_2^2 < 1$

Logistic regression Using polynomial function

Non-linear decision boundaries



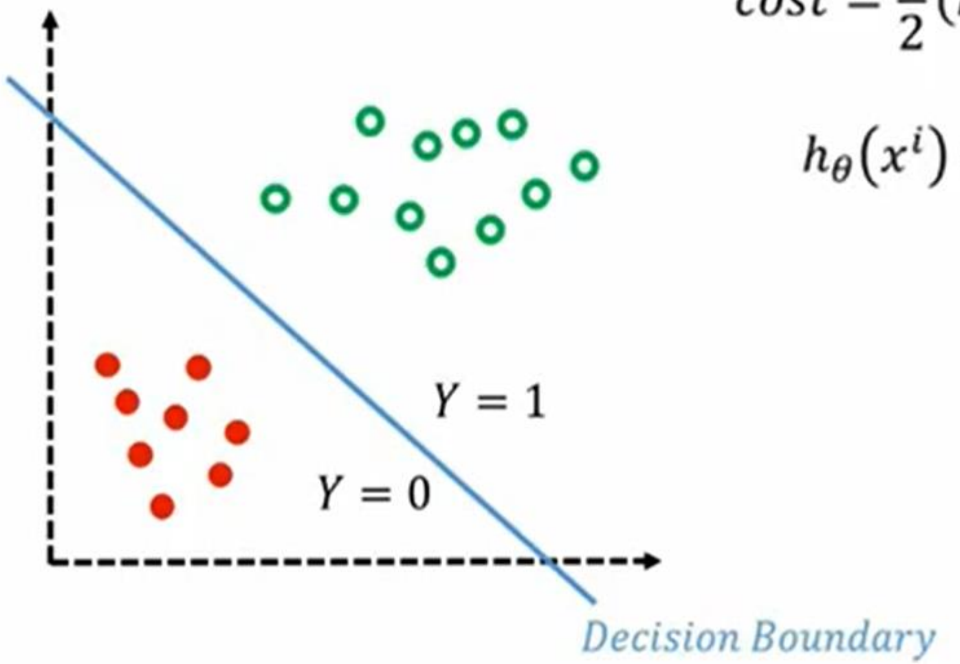
$$f_{\vec{w},b}(\vec{x}) = g(z) = g(w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_1x_2 + w_5x_2^2 + w_6x_1^3 + \cdots + b)$$



Logistic Regression

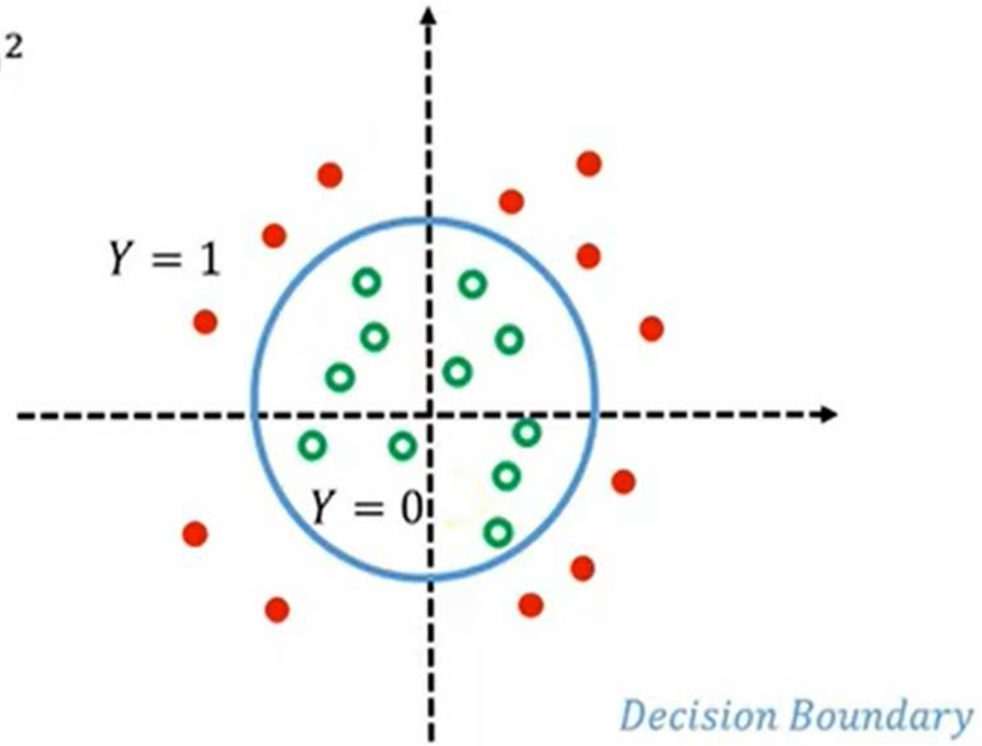
$$cost = \frac{1}{2}(h_\theta(x^i) - y^i)^2$$

$$h_\theta(x^i) = \frac{1}{1+e^{-wx^i+b}}$$



$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h_\theta(x) = -3 + x_1 + x_2$$



$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2$$

$$h_\theta(x) = -1 + x_1^2 + x_2^2$$

Linear Regression VS Logistic Regression

1. Linear Regression: Linear regression is used to model the relationship between a dependent variable and one or more independent variables, assuming a linear relationship. It is primarily used for predicting continuous numeric values.

2. Logistic Regression: Logistic regression is used to model the relationship between a dependent variable and one or more independent variables, with the aim of predicting the probability of an event or a binary outcome. It is commonly used for classification problems where the dependent variable is categorical.

How to choose parameters

Training set

tumor size (cm) x_1	...	patient's age x_n	malignant? y
10		52	1
2		73	0
5		55	0
12		49	1
...		...	...

$i=1$ ⋮ $i=m$

$i = 1, \dots, m$ ← training examples

$j = 1, \dots, n$ ← features

$\boxed{\text{target } y \text{ is } 0 \text{ or } 1}$

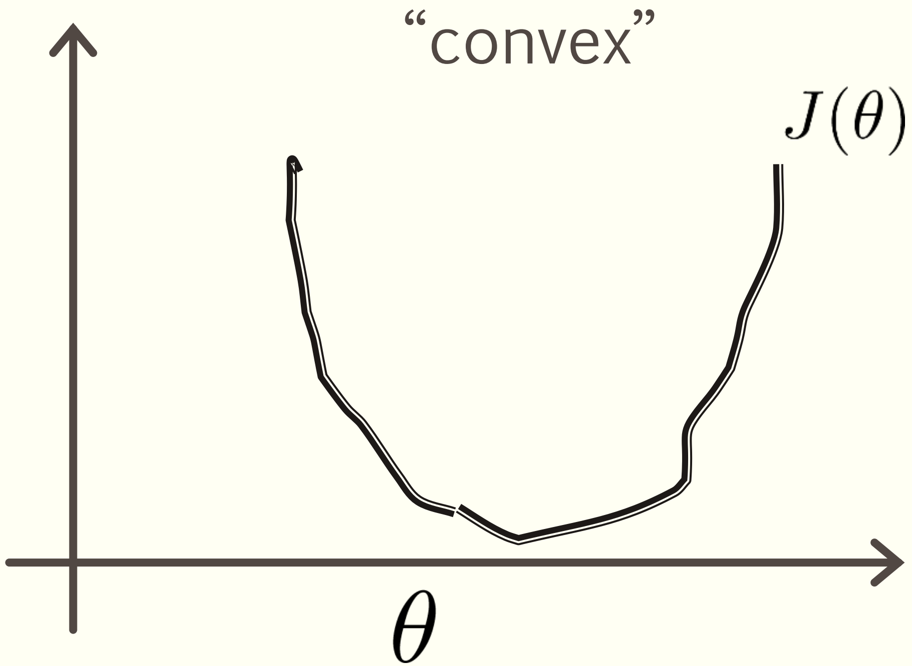
$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}}$$

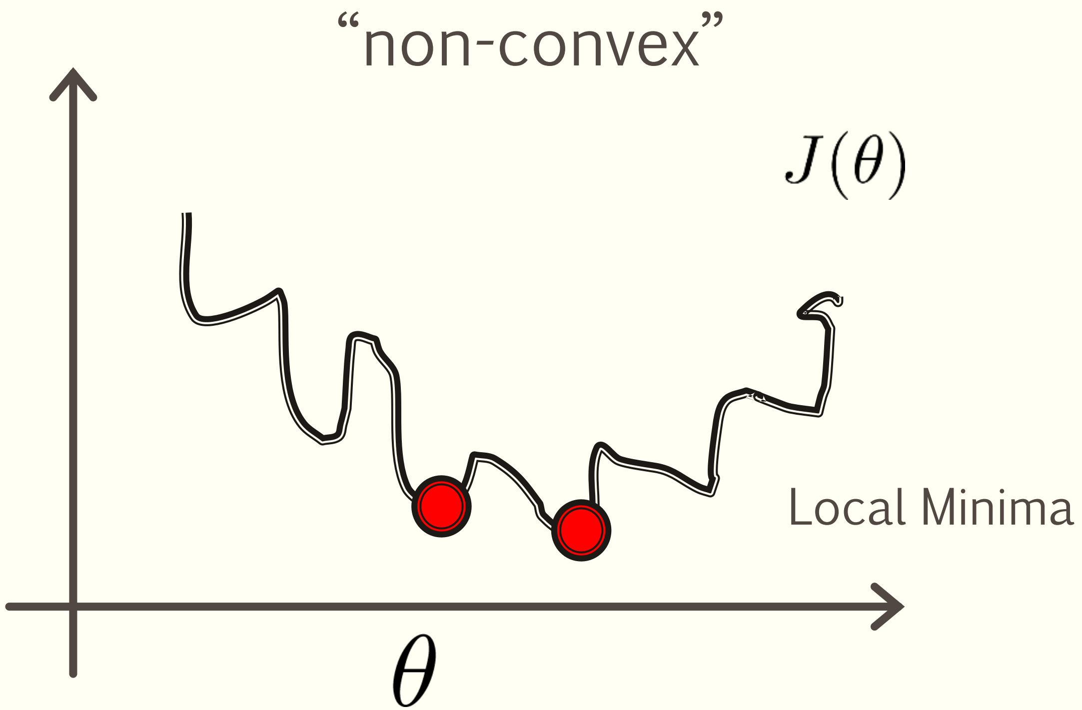
How to choose $\vec{w} = [w_1 \quad w_2 \quad \cdots \quad w_n]$ and b?

Cost function

~~Linear regression:~~ $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

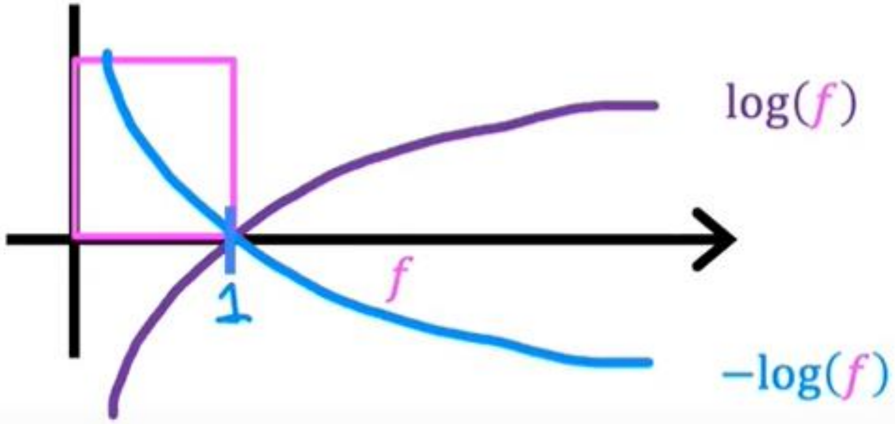
Logistic Regression $\qquad g(z) = \dfrac{1}{1+e^{-z}}$





Logistic cost function

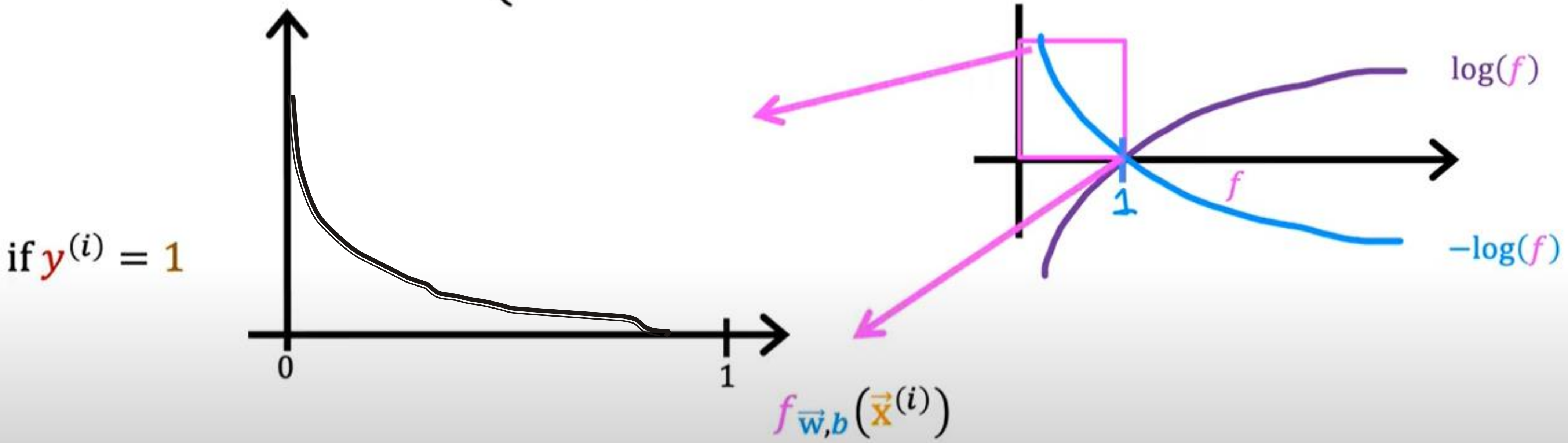
$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$



$$J(\vec{\mathrm{w}}, b) = \frac{1}{m}\sum_{i=1}^{m} L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right)$$

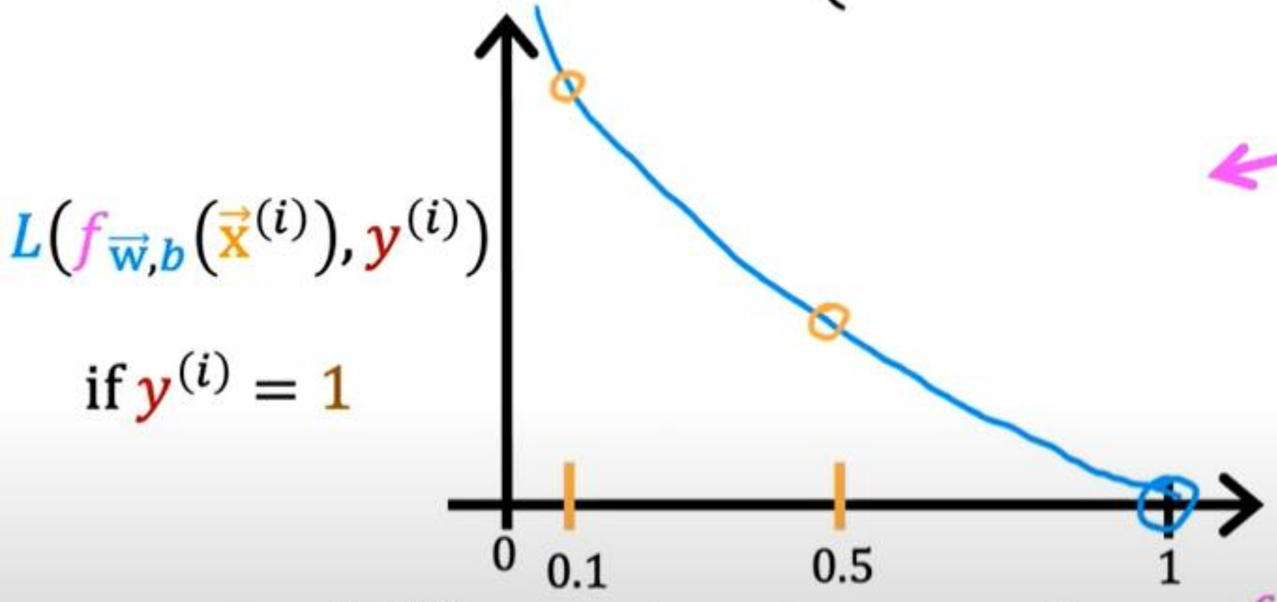
Logistic cost function

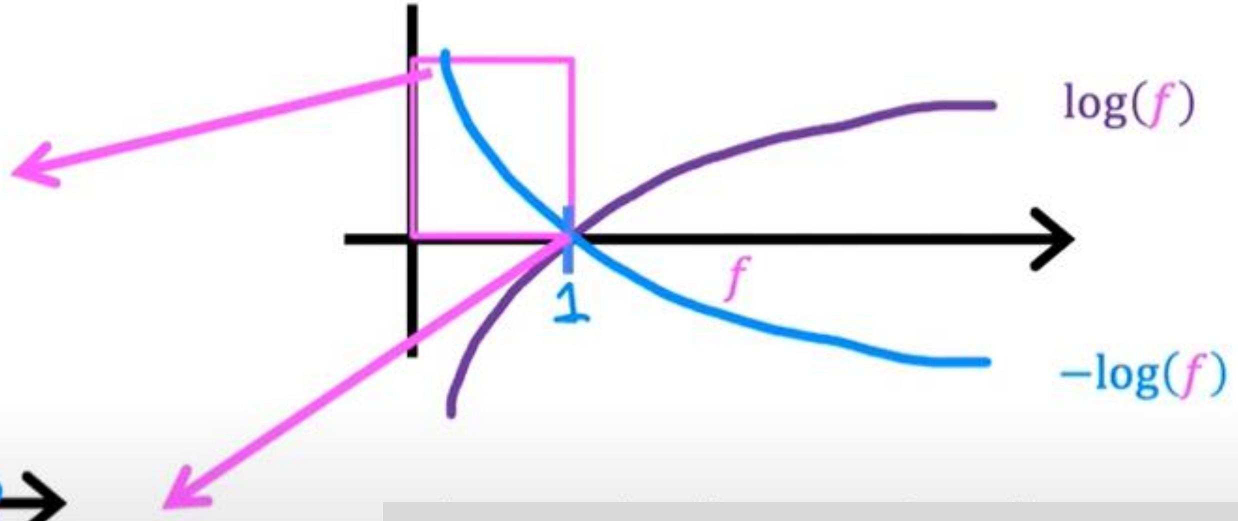
$$L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$



Logistic cost function

$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$



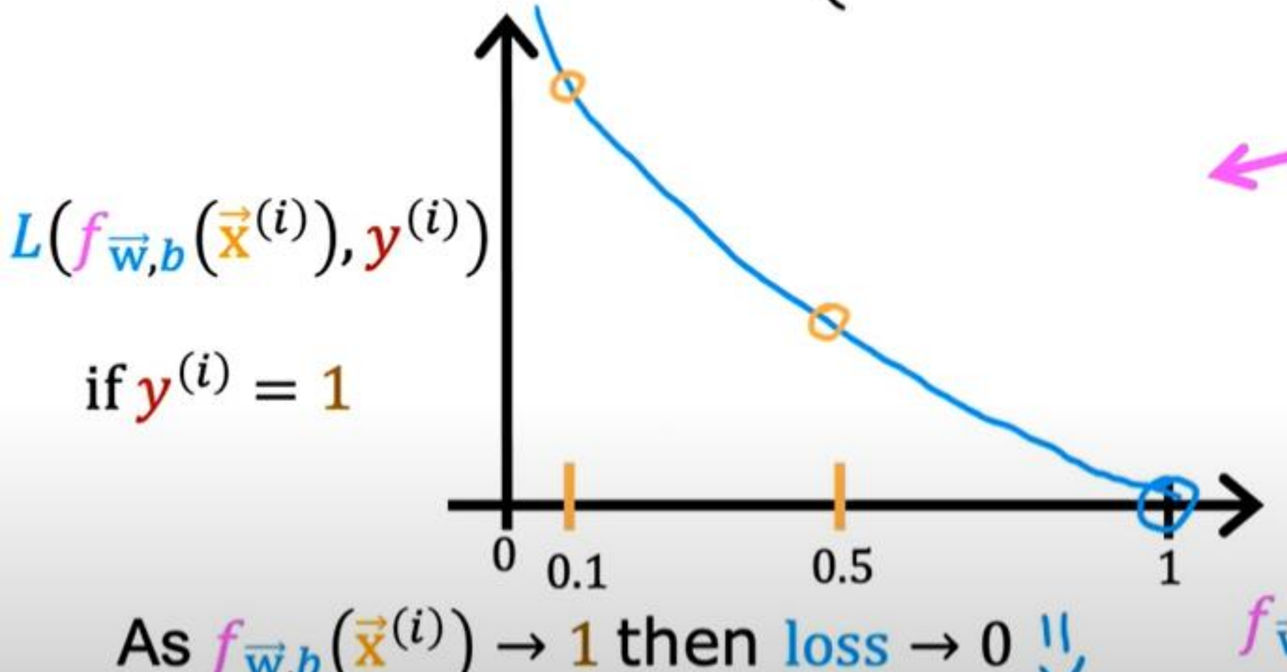


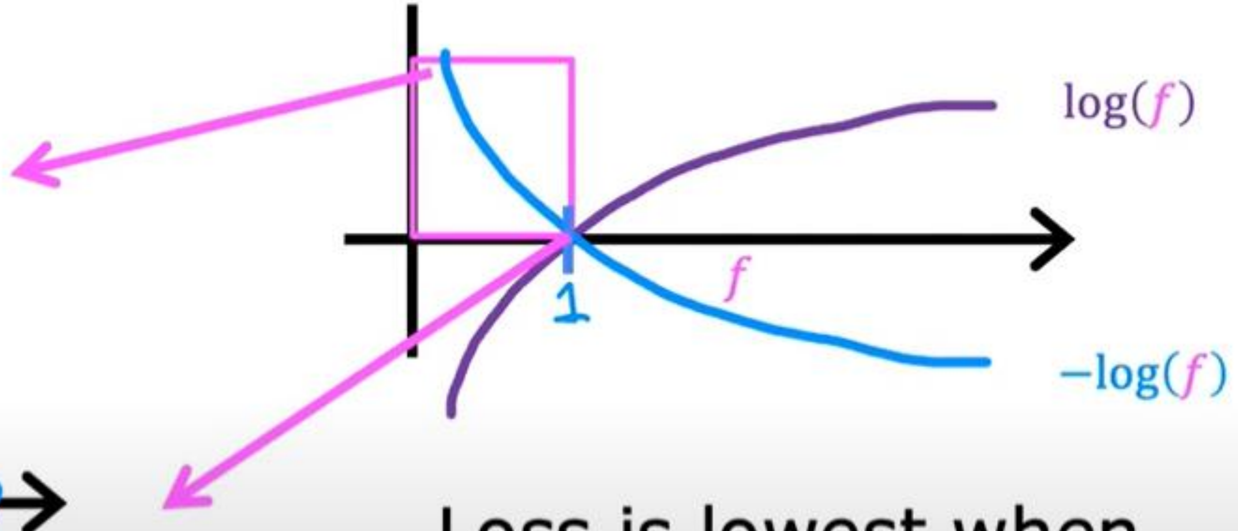
As $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) \rightarrow 1$ then loss $\rightarrow 0$

As $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) \rightarrow 0$ then loss $\rightarrow \infty$

Logistic cost function

$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$





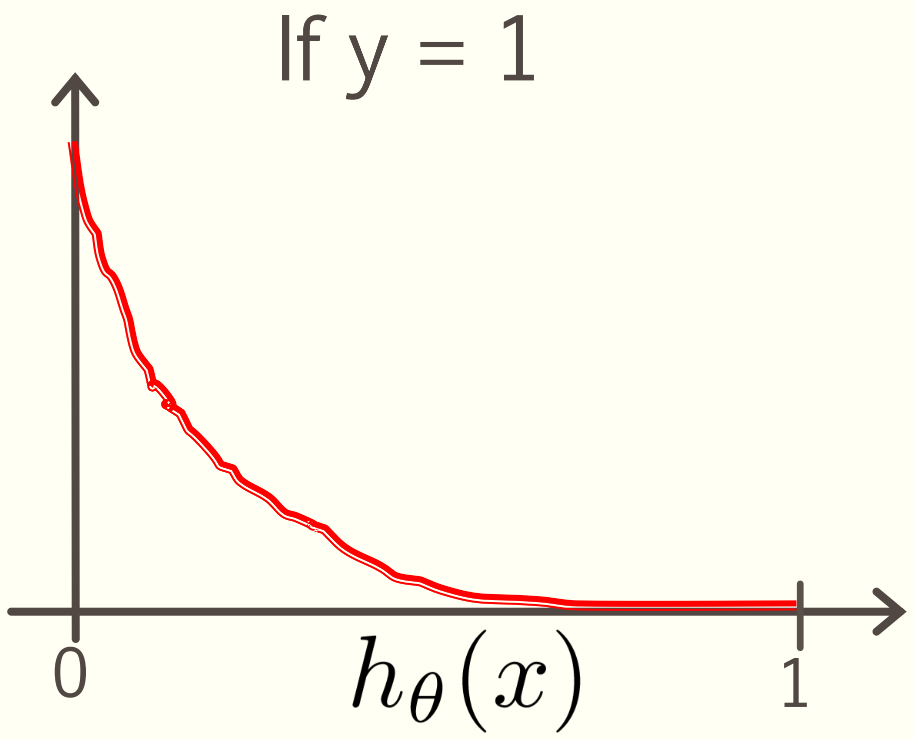
As $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) \to 1$ then loss $\to 0$

As $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) \to 0$ then loss $\to \infty$

Loss is lowest when $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)$ predicts close to true label $y^{(i)}$.

Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



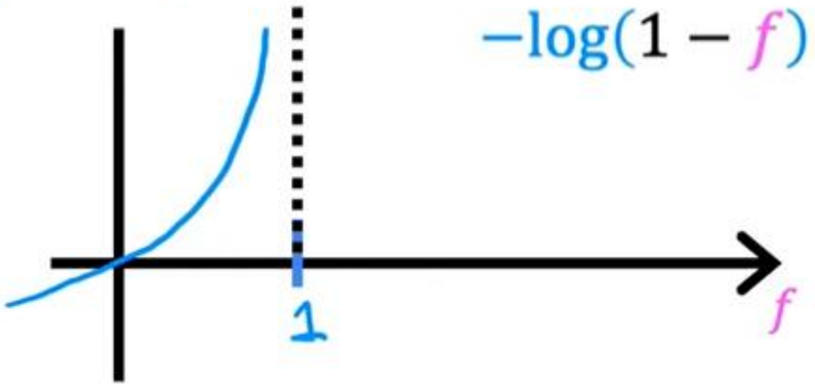
$\mathrm{Cost} = 0$ if $y = 1, h_\theta(x) = 1$
But as $h_\theta(x) \to 0$
$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x; \theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Logistic cost function

$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$

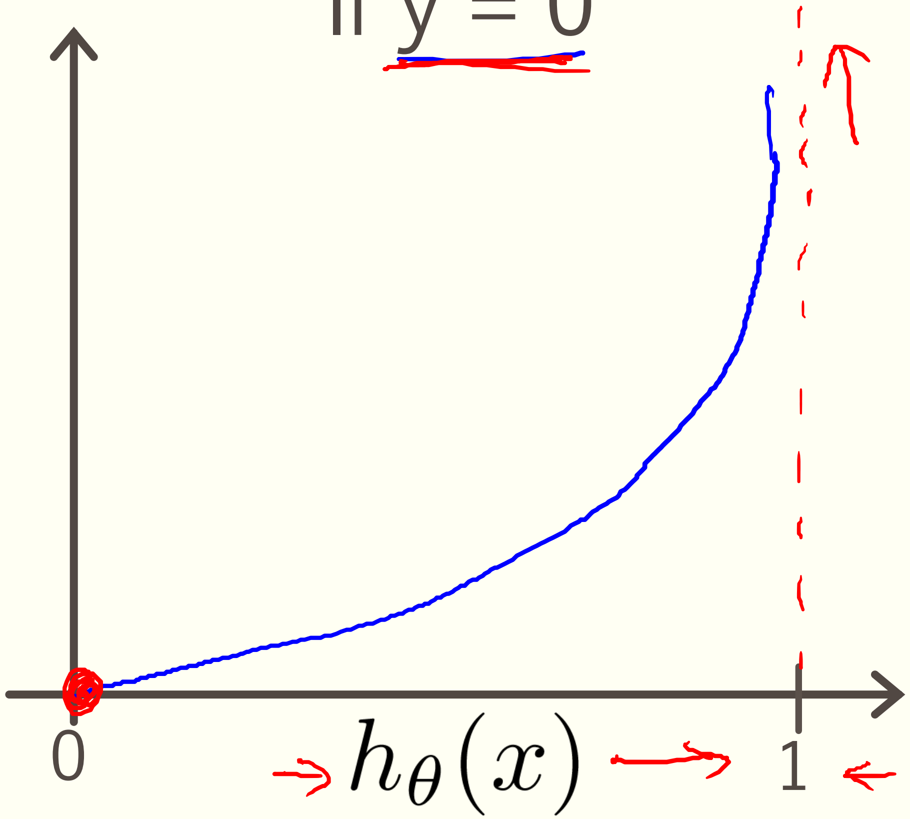
$$J(\vec{\mathrm{w}}, b) = \frac{1}{m}\sum_{i=1}^{m} L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right)$$



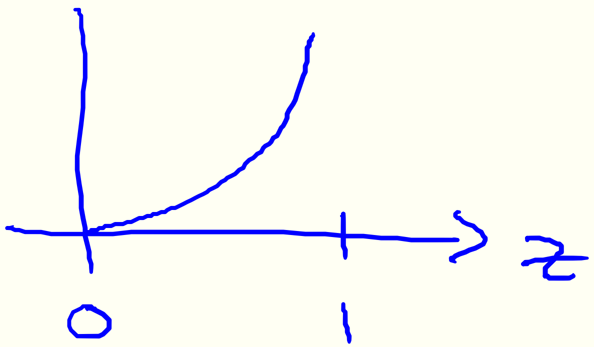
Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ \boxed{-\log(1 - h_\theta(x))} & \text{if } y = 0 \end{cases}$$

If y = 0

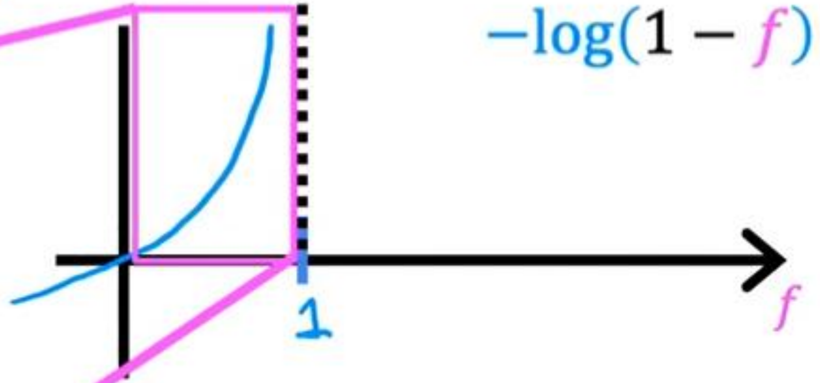


$-\log(1-z)$

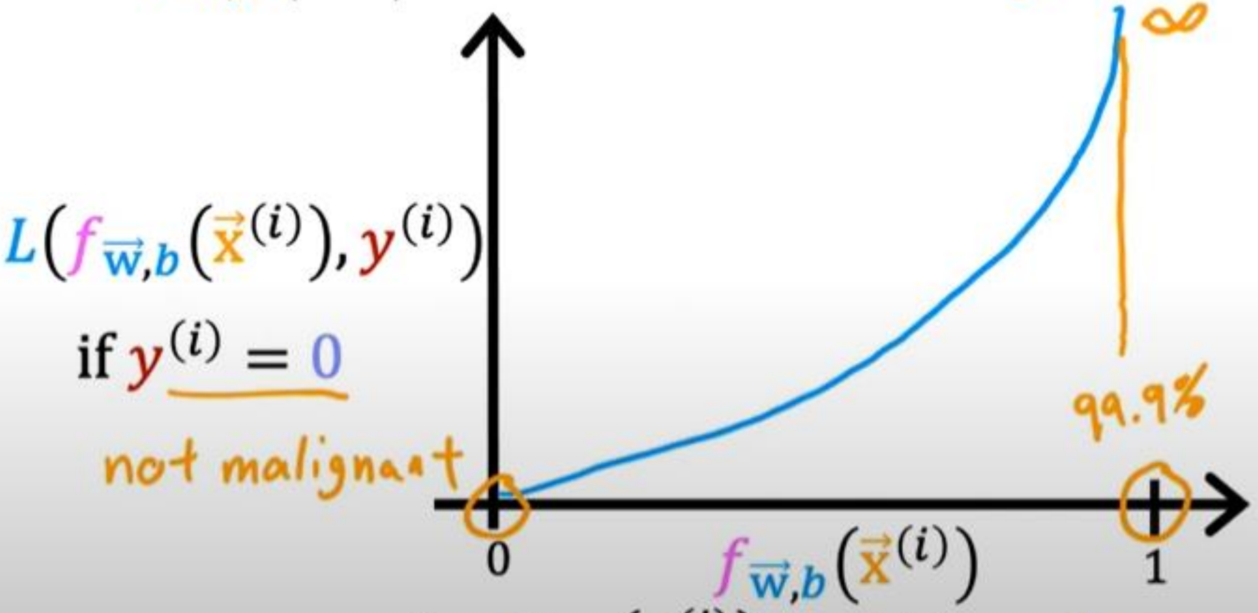


Logistic cost function

$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$



As $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) \to 0$ then loss $\to 0$:)



As $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) \to 1$ then loss $\to \infty$:(

The further prediction $f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)$ is from target $y^{(i)}$, the higher the loss.

Logistic regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$J(\vec{\mathrm{w}}, b) = \frac{1}{m}\sum_{i=1}^{m} L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right)$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note: $y = 0$ or 1 always

It is required to find the parameters w and B that minimize cost

Simplified Loss function

$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$

$$\boxed{L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = -\, y^{(i)}\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) - \left(1 - y^{(i)}\right)\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right)}$$

When Y=1

$$-\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) \quad \text{if } y^{(i)} = 1$$

When Y=0

$$-\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) \quad \text{if } y^{(i)} = 0$$

Cost Function

$$L\left(f_{\vec{\mathbf{w}},b}\left(\vec{\mathbf{x}}^{(i)}\right), y^{(i)}\right) = -y^{(i)}\log\left(f_{\vec{\mathbf{w}},b}\left(\vec{\mathbf{x}}^{(i)}\right)\right) - \left(1 - y^{(i)}\right)\log\left(1 - f_{\vec{\mathbf{w}},b}\left(\vec{\mathbf{x}}^{(i)}\right)\right)$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))\right]$$

This is based on maximum likehood principles from statistics

It is required to find the parameters w and B that minimize cost

Gradient Descent

repeat {

looks like linear regression!

$$w_j = w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right) x_j^{(i)} \right]$$

$$b = b - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right) \right]$$

} simultaneous updates

Linear regression $f_{\vec{w},b}(\vec{x}) = \vec{w} \cdot \vec{x} + b$

Logistic regression $f_{\vec{w},b}(\vec{x}) = \dfrac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$

Gradient Descent

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

(simultaneously update all θ_j)

$\}$

GRADIENT DESCENT

in Linear Regression

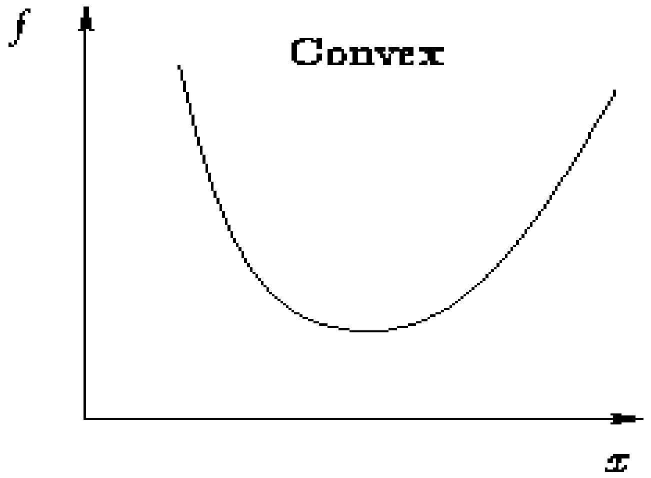
Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

(simultaneously update all θ_j)

$\}$



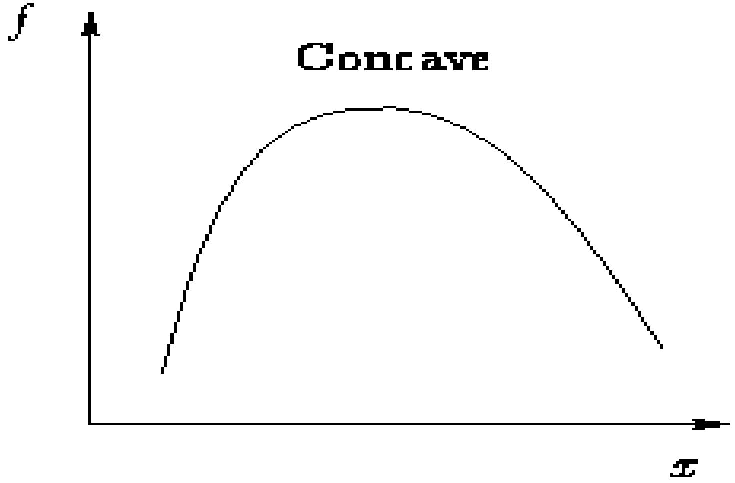
in Logistic Regression

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

$\{$

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{m} \left(y_i - \frac{1}{1 + e^{-\theta^t x_i}} \right) x_{ij}$$

$\}$



We can now use **gradient ascent** to maximize j(θ) The update rule will be: repeat until convergence

DEFINITION

- Binary Logistic Regression

  - We have a set of feature vectors X with corresponding binary outputs

$$X = \{x_1, x_2, \ldots, x_n\}^T$$

$$Y = \{y_1, y_2, \ldots, y_n\}^T, where \quad y_i \in \{0,1\}$$

  - We want to model p(y|x)

$$p(y_i = 1 \mid x_i, \theta) = \sum_j \theta_j x_{ij} = x_i \theta$$

By definition $p(y_i = 1 \mid x_i, \theta) \in \{0,1\}$. We want to transform the probability to remove the range restrictions, as $x_i\theta$ can take any real value.

USING ODDS

- Odds

  p : probability of an event occurring

  1 – p : probability of the event not occurring

  The odds for event i are then defined as

$$odds_i = \frac{p_i}{1 - p_i}$$

  Taking the ***log*** of the odds removes the range restrictions.

$$\log\left(\frac{p_i}{1 - p_i}\right) = \sum_j \theta_j x_{ij} = x_i \theta$$

  This way we map the probabilities from the [0; 1] range to the entire number line (real value).

LOGISTIC REGRESSION MODEL

Linear Regression

$$h_\theta(x) = \theta^t x$$

Logistic Regression

$$g(\theta^t x) = \begin{cases} 1, & \dfrac{1}{1+e^{-\theta x}} \geq 0.5 \\ 0, & 1 - \dfrac{1}{1+e^{-\theta x}} < 0.5 \end{cases}$$

$$p(y_i = 1 \mid x_i, \theta) = \frac{1}{1+e^{-\theta^t x}}$$

$$p(y_i = 0 \mid x_i, \theta) = 1 - \frac{1}{1+e^{-\theta^t x}}$$

$$p(y_i \mid x_i : \theta) = \left(\frac{1}{1+e^{-\theta^t x}}\right)^{y_i} \left(1 - \frac{1}{1+e^{-\theta^t x}}\right)^{1-y_i}$$

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that y = 1 on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$$h_\theta(x) = 0.7$$

Tell patient that 70% chance of tumor being malignant

$h_\theta(X) = p(y = 1 \mid x; \theta)$ "probability that y = 1, given x, parameterized by θ"

$$P(y = 0|x;\theta) + P(y = 1|x;\theta) = 1$$
$$P(y = 0|x;\theta) = 1 - P(y = 1|x;\theta)$$

Logistic Regression

Multi-class classification: One-vs-all

Multiclass classification

Email foldering/tagging: Work, Friends, Family, Hobby

$y=1$ $y=2$ $y=3$ $y=4$

Medical diagrams: Not ill, Cold, Flu

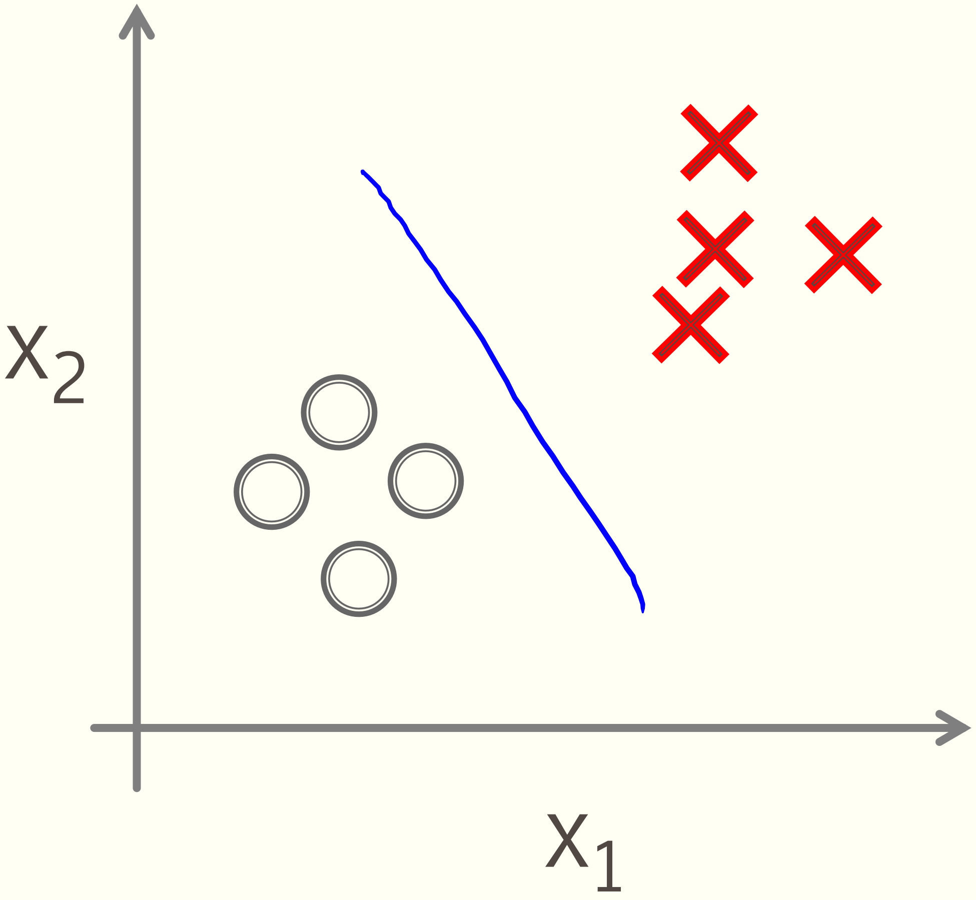
$y=1$ 2 3

Weather: Sunny, Cloudy, Rain, Snow

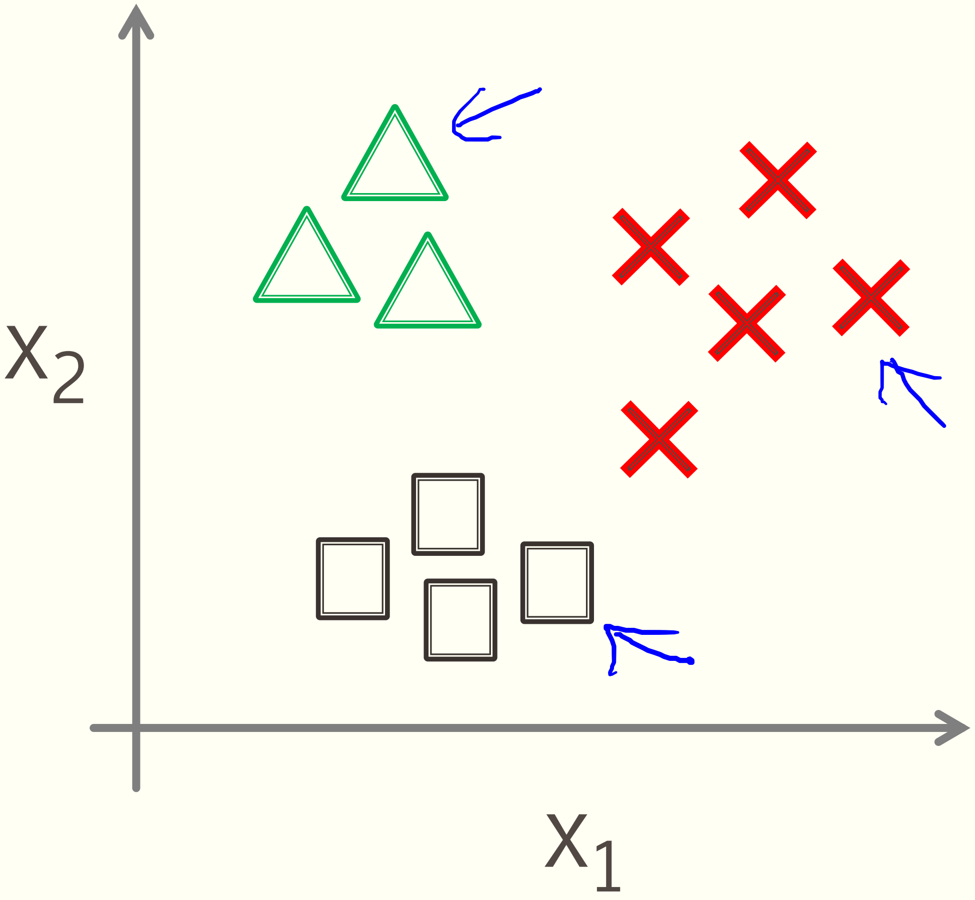
$y=1$ 2 3 4 ←

0 1 2 3

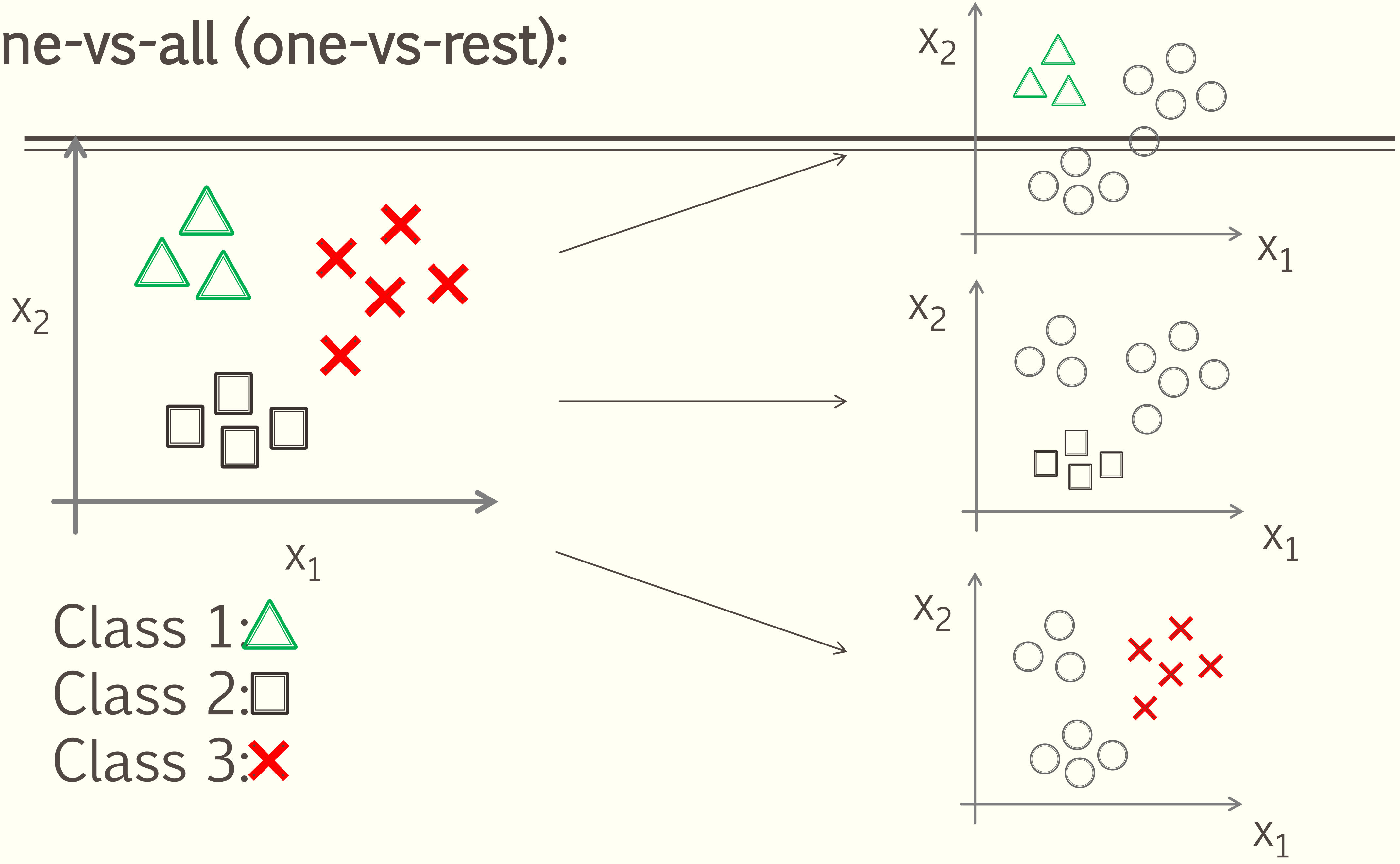
Binary classification:



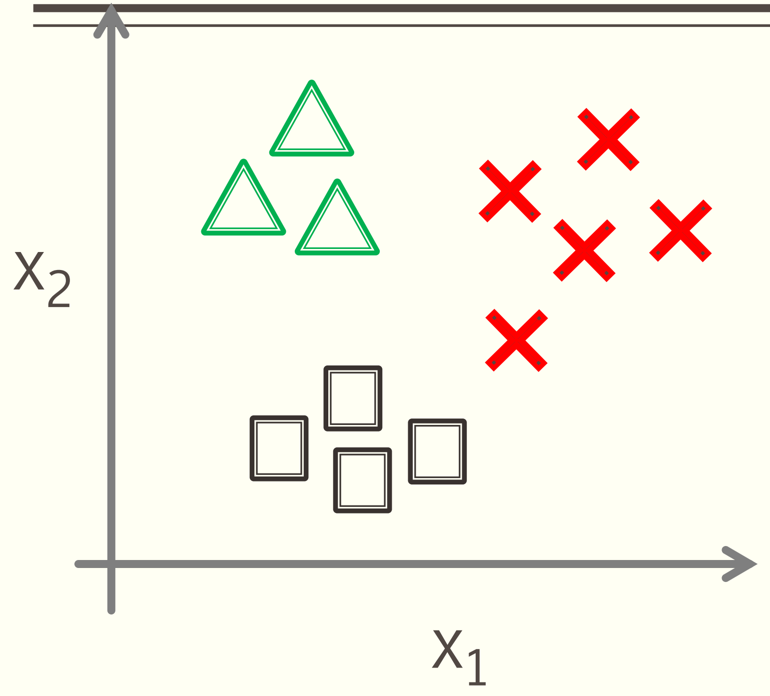
Multi-class classification:

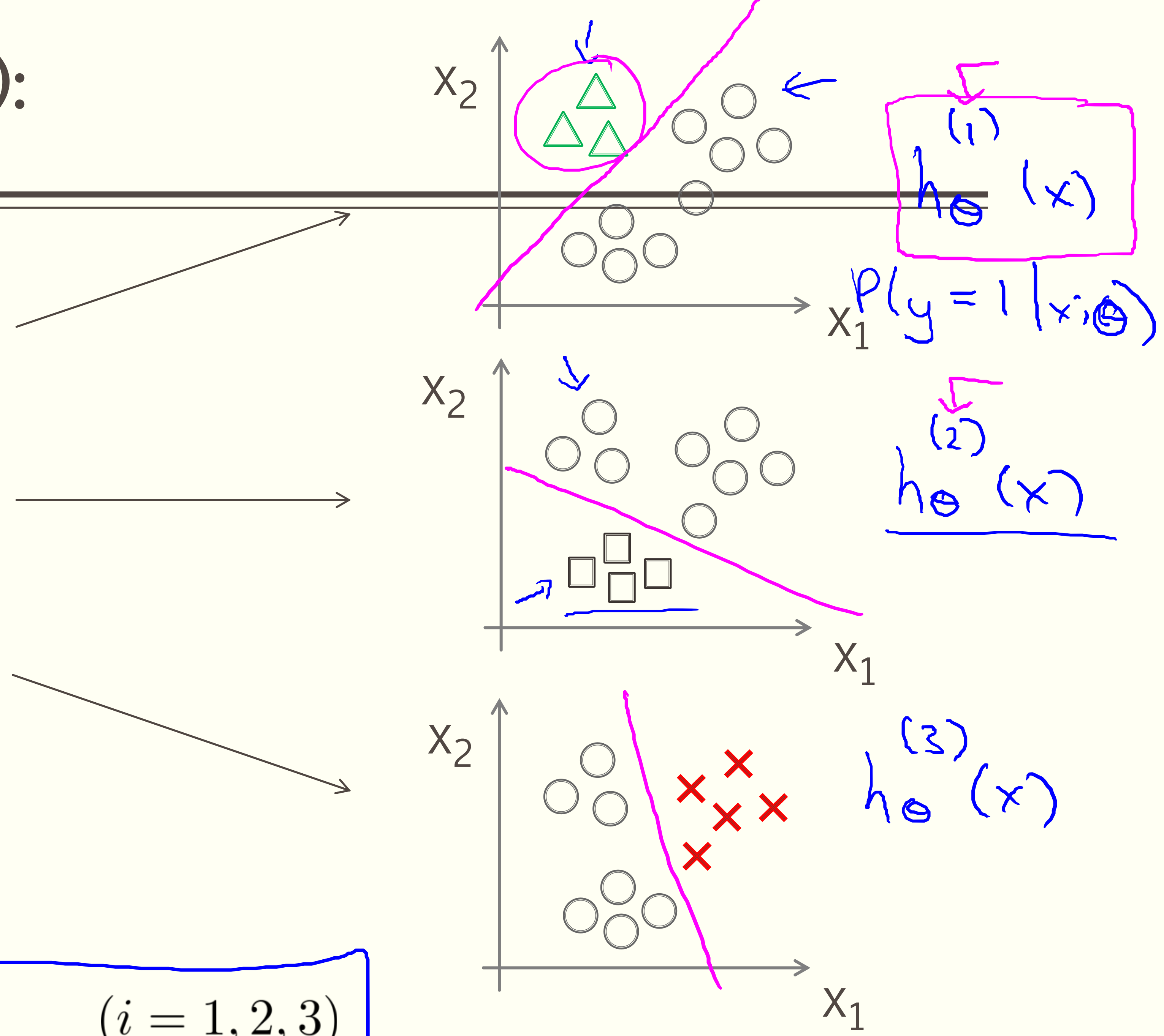


One-vs-all (one-vs-rest):



One-vs-all (one-vs-rest):



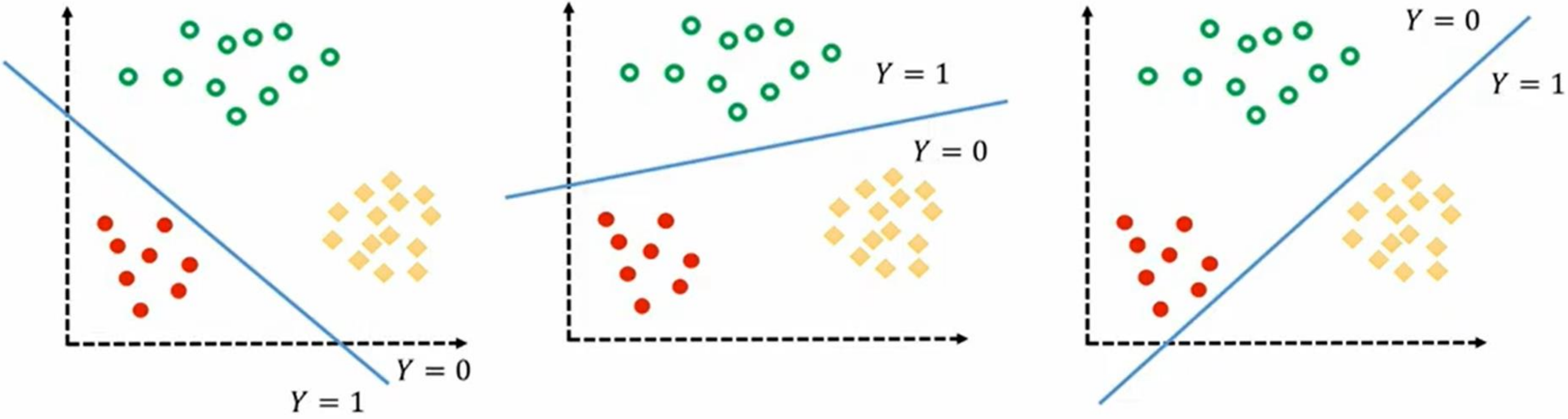


Class 1: △
Class 2: □
Class 3: ×

$$\boxed{h_\theta^{(i)}(x) = P(y = i|x;\theta) \qquad (i = 1, 2, 3)}$$

Multiclass Classification

One-vs-all



$h_\theta^1(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$

$h_\theta^2(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$

$h_\theta^3(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$

Multiclass Classification

One-vs-all